Josef Lutz
Heinrich Schlangenotto
Uwe Scheuermann
Rik De Doncker

Semiconductor Power Devices

Physics, Characteristics, Reliability

Second Edition

Semiconductor Power Devices

Josef Lutz · Heinrich Schlangenotto
Uwe Scheuermann · Rik De Doncker

Semiconductor Power Devices

Physics, Characteristics, Reliability

Second Edition

Josef Lutz
Chair Power Electronics and
Electromagnetic Compatibility, Faculty of
ET/IT
Chemnitz University of Technology
Chemnitz
Germany

Heinrich Schlangenotto
Neu-Isenburg
Germany

Uwe Scheuermann
Semikron Elektronik GmbH & Co. KG
Nuremberg
Germany

Rik De Doncker
Chair Power Generation and Storage
Systems, Faculty of ET/IT
E.ON ERC, RWTH Aachen University
Aachen
Germany

ISBN 978-3-319-70916-1 ISBN 978-3-319-70917-8 (eBook)
https://doi.org/10.1007/978-3-319-70917-8

Library of Congress Control Number: 2017958836

1st edition: © Springer-Verlag Berlin Heidelberg 2011
2nd edition: © Springer International Publishing AG 2018
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface to the Second Edition

The first edition of this book was widely used and accepted by professionals in the field. The progress in power devices, however, makes a second edition necessary.

For this second edition, the basic chapters on semiconductor properties and pn-junctions were revised and extended widely. Effects of doping, current transport, and recombination are now treated much more in detail and depth.

In the chapter on technology, the description of theory of diffusion in silicon is considerably augmented. Aspects on 300-mm technology for Si IGBTs were added. New are the sections on radiation-induced doping and on GaN technology. The chapter on Schottky diodes was revised by an improved treatment of the physics of the metal–semiconductor junction and extended by sections on Merged Pin Schottky diodes. In the chapter on thyristors, the description of the gate-commutated thyristor GCT is added. The chapter on MOS transistors and field-controlled wide-bandgap devices replaces the former chapter on MOSFETs. Despite the progress in wide-bandgap devices, IGBTs are still seen as the main volume components for future power electronics, new aspects on reverse conducting IGBTs were added, and future potential of the IGBT is discussed.

Due to the strong progress in packaging, the former chapter on packaging technology is now replaced by two chapters: "Packaging of Power Devices" and "Reliability and Reliability Testing." Especially the reliability sections are strongly expanded considering new test methods, also in the viewpoint of wide-bandgap devices. A comprehensive section on cosmic ray failures is now placed in this section.

Finally, new research results on transient avalanche oscillations were added as well as some aspects on monolithically integrated GaN devices.

Several researchers in power devices have supported this work with helpful discussions, suggestions, and comments. These are especially Arnost Kopta and Munaf Rahimo from ABB Semiconductors, Markus Behet from EpiGaN, Richard Reiner from Fraunhofer IAF Freiburg, Daniel Hofmann from Fuji Electric, Thomas Laska, Roland Rupp, Hans-Joachim Schulze, and Ralf Siemieniec from Infineon, Dan Kinzer from Navitas, Marion Junghänel from Semikron, Karl Nesemann from SMA, Tomoyuki Shoji from Toyota Nando Kaminski from University of Bremen,

Ulrich Schwarz from Chemnitz University of Technology, Christian Felgemacher from University of Kassel, and Axel Richter from Baden-Wuerttemberg Cooperative State University. Several Master and Ph.D.students at Chemnitz University of Technology have supported part of the work, especially Menia Beier-Möbius, Riteshkumar Bhojani, Haiyang Cao, Susanne Fichtner, Jörg Franke, Christian Herold, Shanmuganathan Palanisamy, Peter Seidel, and Guang Zeng. Stefanie Glöckner has given support with improvements of the English text. Finally, the authors thank the many other researchers and students in power electronics, which supported this second edition with critical comments and discussions.

Chemnitz, Germany — Josef Lutz
Neu-Isenburg, Germany — Heinrich Schlangenotto
Nuremberg, Germany — Uwe Scheuermann
Aachen, Germany — Rik De Doncker
October 2017

Preface to the First Edition

Power electronics is gaining more and more importance in industry and society. It has the potential to substantially increase the efficiency of power systems, a task of great significance. To exploit this potential, not only engineers working in the development of improved and new devices but also application engineers in the field of power electronics need to understand the basic principles of semiconductor power devices. Furthermore, since a semiconductor device can only fulfill its function in a suitable environment, interconnection and packaging technologies with the related material properties have to be considered as well as the problem of cooling, which has to be solved for reliable applications.

This book was written for students and for engineers working in the field of power device design and power electronics applications. The focus was set on modern semiconductor switches such as power MOSFETs and IGBTs together with the essential freewheeling diodes. The practicing engineer may start his/her work with the book with the specific power device. Each chapter presents first the device structure and the generic characteristics and then a more thorough discussion is added with the focus on the physical function principles. The in-depth discussions require the principles of semiconductor physics, the functioning of pn-junctions, and the basics of technology. These topics are treated in depth such that the book will also be of value for the semiconductor device specialist.

Some subjects are treated in particular detail and presented here for the first time in an English textbook on power devices. In device physics, this is especially the emitter recombination which is used in modern power devices to control forward conduction and switching properties. A detailed discussion of its influence is given using parameters characterizing the emitter recombination properties. Furthermore, because of the growing awareness of the importance of packaging techniques for reliable applications, chapters on packaging and reliability are included. During the development of power electronic systems, engineers often are confronted with failures and unexpected effects with the consequence of time-consuming efforts to isolate the root cause of these effects. Therefore, chapters on failure mechanisms and oscillation effects in power circuits are included in this textbook to supply guidance based on long-time experience.

The book has emerged from lectures on "Power devices" held by J. Lutz at Chemnitz University of Technology and from earlier lecture notes on "Power devices" from H. Schlangenotto held at Darmstadt Technical University in 1991–2001. Using these lectures and adding considerable material on new devices, packaging, reliability, and failure mechanisms, Lutz published in German the book *Halbleiter-Leistungsbauelemente – Physik, Eigenschaften, Zuverlässigkeit* in 2006. The English textbook presented here is far more than a translation; it was considerably extended with new material.

The basic chapters on semiconductor properties and pn-junctions and a part of the chapter on pin diodes were revised and enhanced widely by H. Schlangenotto. J. Lutz extended the chapters on thyristors, MOSFETs, IGBTs, and failure mechanisms. U. Scheuermann contributed the chapter on packaging technology, reliability, and system integration. R. De Doncker supplied the introduction on power devices as the key components. All the authors have contributed, however, also to other chapters not written mainly by themselves.

Several researchers in power devices have supported this work with helpful discussions, support in translations, suggestions, and comments. These are especially Arnost Kopta, Stefan Linder, and Munaf Rahimo from ABB Semiconductors, Dieter Polenov from BMW, Thomas Laska, Anton Mauder, Franz-Josef Niedernostheide, Ralf Siemieniec, and Gerald Soelkner from Infineon, Martin Domeij and Anders Hallén from KTH Stockholm, Stephane Lefebvre from SATIE, Michael Reschke from Secos, Reinhard Herzer and Werner Tursky from Semikron, Wolfgang Bartsch from SiCED, Dieter Silber from University of Bremen, Hans Günter Eckel from the University of Rostock. Several diploma and Ph.D. students at Chemnitz University of Technology have supported part of the work, especially Hans-Peter Felsl, Birk Heinze, Roman Baburske, Marco Bohlländer, Tilo Pollera Matthias Baumann, and Thomas Basler. Thomas Plum and Florian Mura from RWTH Aachen have translated the chapter on MOSFETS, and Mary-Joan Blümich has given support with improvements of the English text. Finally, the authors thank many other researchers and students in power electronics, who supported this work with critical comments and discussions.

Chemnitz, Germany — Josef Lutz
Neu-Isenburg, Germany — Heinrich Schlangenotto
Nuremberg, Germany — Uwe Scheuermann
Aachen, Germany — Rik De Doncker
March 2010

Contents

1 Power Semiconductor Devices—Key Components for Efficient Electrical Energy Conversion Systems . . . 1
1.1 Systems, Power Converters and Power Semiconductor Devices . . . 1
1.1.1 Basic Principles of Power Converters . . . 3
1.1.2 Types of Power Converters and Selection of Power Devices . . . 5
1.2 Operating and Selecting Power Semiconductors . . . 8
1.3 Applications of Power Semiconductors . . . 11
1.4 Power Electronics for Carbon Emission Reduction . . . 14
References . . . 18

2 Semiconductor Properties . . . 21
2.1 Introduction . . . 21
2.2 Crystal Structure . . . 24
2.3 Energy Gap and Intrinsic Concentration . . . 26
2.4 Energy Band Structure and Particle Properties of Carriers . . . 31
2.5 The Doped Semiconductor . . . 35
2.6 Current Transport . . . 45
2.6.1 Carrier Mobilities and Field Currents . . . 45
2.6.2 High-Field Drift Velocities . . . 52
2.6.3 Diffusion of Carriers, Current Transport Equations and Einstein Relation . . . 54
2.7 Recombination—Generation and Lifetime of Non-equilibrium Carriers . . . 57
2.7.1 Intrinsic Recombination Mechanisms . . . 59
2.7.2 Recombination at Recombination Centers Including Gold, Platinum and Radiation Defects . . . 61
2.8 Impact Ionization . . . 81

2.9 Basic Equations of Semiconductor Devices 88
2.10 Simple Conclusions 92
2.10.1 Temporal and Spatial Decay of a Minority Carrier Concentration 92
2.10.2 Temporal and Spatial Decay of a Charge Density 93
References 94

3 pn-Junctions 101
3.1 The pn-Junction in Thermal Equilibrium 101
3.1.1 The Abrupt Step Junction 104
3.1.2 Graded Junctions 111
3.2 Current-Voltage-Characteristics of the pn-Junction 114
3.3 Blocking Characteristics and Breakdown of the pn-Junction 122
3.3.1 Blocking Current 122
3.3.2 Avalanche Multiplication and Breakdown Voltage 126
3.3.3 Blocking Capability with Wide-Bandgap Semiconductors 135
3.4 Injection Efficiency of Emitter Regions 137
3.5 Capacitance of pn-Junctions 144
References 147

4 Introduction to Power Device Technology 149
4.1 Crystal Growth 149
4.2 Neutron Transmutation for Adjustment of the Wafer Doping 151
4.3 Epitaxial Growth 154
4.4 Diffusion 156
4.4.1 Diffusion Theory, Impurity Distributions 157
4.4.2 Diffusion Constants and Solubility of Dopants 165
4.4.3 High Concentration Effects, Diffusion Mechanisms 168
4.5 Ion Implantation 170
4.6 Oxidation and Masking 175
4.7 Edge Terminations 177
4.8 Passivation 182
4.9 Recombination Centers 183
4.10 Radiation-Induced Doping 189
4.11 Some Aspects on Technology of GaN Devices 191
References 196

5 pin Diodes 201
5.1 Structure of the pin Diode 201
5.2 I–V Characteristic of the pin Diode 203

5.3 Design and Blocking Voltage of the pin Diode 204
5.4 Forward Conduction Behavior 210
5.4.1 Carrier Distribution 210
5.4.2 Junction Voltages 213
5.4.3 Voltage Drop Across the Middle Region 215
5.4.4 Voltage Drop in the Hall Approximation 216
5.4.5 Emitter-Recombination, Effective Carrier Lifetime and Forward Characteristic 218
5.4.6 Temperature Dependency of the Forward Characteristics 227
5.5 Relation Between Stored Charge and Forward Voltage 228
5.6 Turn-on Behavior of Power Diodes 230
5.7 Reverse-Recovery of Power Diodes 232
5.7.1 Definitions 232
5.7.2 Reverse-Recovery Related Power Losses 239
5.7.3 Reverse Recovery: Charge Dynamic in the Diode . . . 243
5.7.4 Fast Diodes with Optimized Reverse-Recovery Behavior 251
5.7.5 MOS-Controlled Diodes 261
5.8 Outlook 268
References 269

6 Schottky Diodes 271
6.1 Energy Band Diagram of the Metal-Semiconductor Junction 271
6.2 Current-Voltage-Characteristics of the Schottky Junction 273
6.3 Structure of Schottky Diodes 275
6.4 Ohmic Voltage Drop of a Unipolar Device 276
6.4.1 Comparison of Silicon Schottky Diodes and pin Diodes for Rated Voltages of 200 and 100 V 280
6.5 Schottky Diodes Based on SiC 280
6.5.1 SiC Unipolar Diode Characteristics 280
6.5.2 Merged Pin Schottky (MPS) Diodes 285
6.5.3 Switching Behavior and Ruggedness of SiC Schottky and MPS Diodes 289
References 292

7 Bipolar Transistors 295
7.1 Function of the Bipolar Transistor 295
7.2 Structure of the Bipolar Power Transistor 297
7.3 I–V Characteristic of the Power Transistor 297
7.4 Blocking Behavior of the Bipolar Power Transistor 299
7.5 Current Gain of the Bipolar Transistor 301
7.6 Base Widening, Field Redistribution and Second Breakdown 306

7.7 Limits of the Silicon Bipolar Transistor . . . 309
7.8 SiC Bipolar Transistor . . . 309
References . . . 310

8 Thyristors . . . 313
8.1 Structure and Mode of Function . . . 313
8.2 I–V Characteristic of the Thyristor . . . 317
8.3 Blocking Behavior of the Thyristor . . . 318
8.4 The Function of Emitter Shorts . . . 320
8.5 Modes to Trigger a Thyristor . . . 321
8.6 Trigger Front Spreading . . . 323
8.7 Follow-up Triggering and Amplifying Gate . . . 324
8.8 Thyristor Turn-off and Recovery Time . . . 327
8.9 The Triac . . . 329
8.10 The Gate Turn-off Thyristor (GTO) . . . 330
8.11 The Gate Commutated Thyristor (GCT) . . . 335
References . . . 339

9 MOS Transistors and Field Controlled Wide Bandgap Devices . . . 341
9.1 Function Principle of the MOSFET . . . 341
9.2 Structure of Power MOSFETs . . . 343
9.3 Current-Voltage Characteristic of MOS-Transistors . . . 346
9.4 Characteristics of the MOSFET Channel . . . 347
9.5 The Ohmic Region . . . 351
9.6 Compensation Structures in Modern MOSFETs . . . 353
9.7 Temperature Dependency of MOSFET Characteristics . . . 357
9.8 Switching Properties of the MOSFET . . . 359
9.9 Switching Losses of the MOSFET . . . 364
9.10 Safe Operating Area of the MOSFET . . . 365
9.11 The Inverse Diode of the MOSFET . . . 366
9.12 SiC Field Effect Devices . . . 371
9.12.1 SiC JFETs . . . 371
9.12.2 SiC MOSFETs . . . 374
9.12.3 The SiC MOSFET Body Diode . . . 377
9.13 GaN Lateral Power Transistors . . . 378
9.14 GaN Vertical Power Transistors . . . 385
9.15 Outlook . . . 386
References . . . 387

10 IGBTs . . . 391
10.1 Mode of Function . . . 391
10.2 The I–V Characteristic of the IGBT . . . 394
10.3 The Switching Behavior of the IGBT . . . 395
10.4 The Basic Types PT-IGBT and NPT-IGBT . . . 398
10.5 Plasma Distribution in the IGBT . . . 402

10.6 Modern IGBTs with Increased Charge Carrier Density 404
10.6.1 Plasma Enhancement by High n-Emitter Efficiency 404
10.6.2 The "Latch-up Free Cell Geometry" 408
10.6.3 The Effect of the "Hole Barrier" 409
10.6.4 Collector Side Buffer Layers 411
10.7 IGBTs with Bidirectional Blocking Capability 412
10.8 Reverse Conducting IGBTs 414
10.9 The Potential of the IGBT 418
References 422

11 Packaging of Power Devices 427
11.1 The Challenge of Packaging Technology 427
11.2 Package Types 429
11.2.1 Capsules 430
11.2.2 The TO-Family and Its Relatives 433
11.2.3 Modules 437
11.3 Physical Properties of Materials 443
11.4 Thermal Simulation and Thermal Equivalent Circuits 445
11.4.1 Analogy Between Thermal and Electrical Parameters 445
11.4.2 One-Dimensional Equivalent Networks 452
11.4.3 The Three-Dimensional Thermal Network 454
11.4.4 The Transient Thermal Resistance 455
11.5 Parasitic Electrical Elements in Power Modules 458
11.5.1 Parasitic Resistances 459
11.5.2 Parasitic Inductances 462
11.5.3 Parasitic Capacities 466
11.6 Advanced Packaging Technologies 469
11.6.1 Silver Sintering Technology 470
11.6.2 Diffusion Soldering 472
11.6.3 Advanced Technologies for the Chip Topside Contact 475
11.6.4 Improved Substrates 479
11.6.5 Advanced Packaging Concepts 481
References 485

12 Reliability and Reliability Testing 489
12.1 The Demand for Increasing Reliability 489
12.2 High Temperature Reverse Bias Test 492
12.3 High Temperature Gate Stress Test 495
12.4 Temperature Humidity Bias Test 499
12.5 High Temperature and Low Temperature Storage Tests 502
12.6 Temperature Cycling and Temperature Shock Test 503

12.7 Power Cycling Test 505
12.7.1 Power Cycling Test Execution 505
12.7.2 Power Cycling Induced Failure Mechanisms 511
12.7.3 Models for Lifetime Prediction 522
12.7.4 Separation of Failure Modes 526
12.7.5 Mission Profiles and Superposition of Power Cycles 530
12.7.6 Power Cycling Capability of Molded TO Packages 534
12.7.7 Power Cycling of SiC Devices 536
12.8 Cosmic Ray Failures 541
12.8.1 The Salt Mine Experiment 541
12.8.2 Origin of Cosmic Rays 542
12.8.3 Cosmic Ray Failure Patterns 545
12.8.4 Basic Failure Mechanism Model 547
12.8.5 Basic Design Rules 548
12.8.6 Extended Model Considering the nn^+ Junction 553
12.8.7 Further Design Aspects in Extended Models 558
12.8.8 Cosmic Ray Stability of SiC Devices 559
12.9 Statistical Evaluation of Reliability Test Results 563
12.10 Further Reliability Tests 574
References 576

13 Destructive Mechanisms in Power Devices 583
13.1 Thermal Breakdown—Failures by Excess-Temperature 583
13.2 Surge Current 586
13.3 Overvoltage – Voltage Above Blocking Capability 590
13.4 Dynamic Avalanche 596
13.4.1 Dynamic Avalanche in Bipolar Devices 596
13.4.2 Dynamic Avalanche in Fast Diodes 598
13.4.3 Diode Structures with High Dynamic Avalanche Capability 608
13.4.4 Turn-off of Over-Current and Dynamic Avalanche in IGBTs 612
13.5 Exceeding the Maximum Turn-off Current of GTOs 615
13.6 Short-Circuit in IGBTs 616
13.6.1 Short Circuit Types I, II and III 616
13.6.2 Thermal and Electrical Stress in Short Circuit 621
13.6.3 Current Filamentation at Short Circuit 626
13.7 Failure Analysis in IGBT Circuits 630
References 633

14 Power Device Induced Oscillations and Electromagnetic Disturbances 637
14.1 Frequency Range of Electromagnetic Disturbances 637
14.2 LC Oscillations 640
14.2.1 Turn-off Oscillations with IGBTs Connected in Parallel 640
14.2.2 Turn-off Oscillations with Snappy Diodes 642
14.2.3 Turn-off Oscillations with Wide Bandgap Devices 645
14.3 Transit-Time Oscillations 647
14.3.1 Plasma-Extraction Transit-Time (PETT) Oscillations 648
14.3.2 Dynamic Impact-Ionization Transit-Time (IMPATT) Oscillations 655
14.3.3 Transient-Avalanche (TA) Oscillations 659
14.3.4 Summarizing Remarks on Transit-Time Oscillations 663
References 664

15 Integrated Power Electronic Systems 667
15.1 Definition and Basic Features 667
15.2 Monolithically Integrated Systems – Power IC's 670
15.3 GaN Monolithic Integrated Systems 673
15.4 System Integration on Printed Circuit Board 676
15.5 Hybrid Integration 679
References 685

Appendix A: Modeling Parameters of Carrier Mobilities in Si and 4H-SiC 689

Appendix B: Correlates to Recombination Centers 691

Appendix C: Avalanche Multiplication Factors and Effective Ionization Rate 697

Appendix D: Thermal Parameters of Important Materials in Packaging Technology 703

Appendix E: Electric Parameters of Important Materials in Packaging Technology 705

Index 707

Symbols

A	Area (cm^2)
B	Fulop constant: Proportionality factor for $\alpha_{eff} \sim E^{\mathrm{n}}$
$c_{n,p}$	Capture coefficient for electrons/holes (cm^{3}s^{-1})
$c_{An,p}$	Auger capture coefficient for electrons/holes (cm^{3}s^{-1})
C	Capacitance (As/V)
C_j	Junction capacitance (As/V)
D	Diffusion constant (cm^2/s)
D_A	Ambipolar diffusion constant (cm^2/s)
$D_{n,p}$	Diffusion constant of electrons/holes (cm^2/s)
$e_{n,p}$	Emission rate of electrons/holes (s^{-1})
E	Energy (J, eV)
E_C	Lower edge of the conduction band (eV)
E_F	Fermi-Level (eV)
E_g	Bandgap (eV)
E_V	Upper edge of the valence band (eV)
E_{off}	Turn-off energy (J)
E_{on}	Turn-on energy (J)
$\mathcal{E}$	Electric field strength (V/cm)
$\mathcal{E}_c$	Electric field strength at avalanche breakdown (V/cm)
F	Statistic distribution function
$g_{n,p}$	Therm. generation rate of electrons/holes (cm^{-3}s^{-1})
$G_{n,p}$	Net generation rate of electrons/holes (cm^{-3}s^{-1})
G_{av}	Avalanche generation rate (cm^{-3}s^{-1})
$h_{n,p}$	Emitter parameter of n/p emitter (cm^{4}s^{-1})
i	= I(t); current, time dependent (A)
I	Current (A)
I_C	Collector current (A)
I_D	Drain current (A)
I_E	Emitter current (A)
I_F	Diode forward current (A)

I_R	Current in blocking direction (A)
I_{RRM}	Reverse recovery current maximum (A)
j	Current density (A/cm^2)
$j_{n,p}$	Current density of electron/hole current (A/cm^2)
j_s	Saturation current density (A/cm^2)
k	Boltzmann constant ($1.38066 \cdot 10^{-23}$) (J/K)
L	Inductivity (H)
L_{par}	Parasitic inductivity (H)
L_A	Ambipolar diffusion length (cm)
L_D	Debye length (cm)
$L_{n,p}$	Diffusion length of electrons/holes (cm)
n, p	Density of free electrons/holes (cm^{-3})
n_0, p_0	Density in thermodynamic equilibrium (cm^{-3})
n^*, p^*	Density of minority carriers outside therm. equilibrium (cm^{-3})
n_i	Intrinsic carrier density (cm^{-3})
n_L, p_L	Density at the left edge of the flooded zone (cm^{-3})
n_R, p_R	Density at the right edge of the flooded zone (cm^{-3})
n_{av}, p_{av}	Density of electrons/holes generated by avalanche (cm^{-3})
N_A	Acceptor density (cm^{-3})
N_C	Effective density of states of the conduction band (cm^{-3})
N_D	Donator density (cm^{-3})
N_{eff}	Effective doping density $\lvert N_D - N_A \rvert$ (cm^{-3})
N_r	Density of deep centers (cm^{-3})
N_r^+, N_r^-	Density of positively/negatively charged deep centers (cm^{-3})
N_V	Effective density of states of the valence band (cm^{-3})
q	Elementary charge ($1.60218 \cdot 10^{-19}$) (As)
Q	Charge (As)
Q_F	Charge carrying the forward current in a bip. device (As)
Q_{RR}	Measured stored charge of a diode (As)
$r_{n,p}$	Therm. recombination rates of electrons/holes (cm^{-3}s^{-1})
$R_{n,p}$	Net recombination rates of electrons/holes (cm^{-3}s^{-1})
R	Resistor (Ohm)
R_{off}	Gate resistance at turn-off (Ohm)
R_{on}	Gate resistance at turn-on (Ohm)
R_{pr}	Projected range (cm)
Rth	Thermal resistance (K/W)
s	Soft factor of a diode (–)
S	Particles per area (cm^{-2})
t	Time (s)
T	Temperature (°C, K)
v	= V(t); voltage, time dependent (V)
V	Voltage (V)
V_{bat}	Battery voltage/DC link voltage (V)
V_B,V_{BD}	Avalanche breakdown voltage (V)

V_C	Forward voltage of a transistor[1] (V)
V_{drift}	Voltage drop across an n⁻-layer (V)
V_{bi}	Built-in voltage of a pn-junction (V)
V_F	Forward voltage (diode) (V)
V_G	Gate voltage (V)
V_{FRM}	Forward recovery voltage peak of a diode (V)
V_M	Voltage peak (V)
V_R	Voltage in blocking direction (V)
V_s	Threshold voltage diode / thyristor / IGBT (V)
V_T	Threshold voltage channel MOSFET, IGBT (V)
$v_{n,p}$	Velocity of electrons/holes (cm/s)
$v_{d(n,p)}$	Drift velocity of electrons/holes (cm/s)
v_{sat}	Saturation drift velocity at high electric field (cm/s)
w_B	Width of the n^--layer (cm)
$w,\ w_{SC}$	Width of the space charge layer (cm)
x	Coordinate (cm)
x_j	Depth of the pn-junction (cm)
α	Current gain in common-base circuit
α_T	Transport factor
$\alpha_{n,p}$	Ionization rates v of electrons/holes (cm^{-1})
α_{eff}	Effective ionization rate (cm^{-1})
β	Current gain in common-emitter circuit
γ	Emitter efficiency
ε_0	Dielectric constant in vacuum ($8.85418 \cdot 10^{-14}$) (F/cm)
ε_r	Relative dielectric constant (Si: 11.7)
$\mu_{n,p}$	Mobility of free electrons/holes ($cm^2V^{-1}s^{-1}$)
ρ	Space charge (As/cm^3)
σ	Electric conductivity ($Acm^{-1}V^{-1}$)
$\tau_{n,p}$	Lifetime of excess eletrons/holes (s)
$\tau_{n0,p0}$	Low-level lifetime of excess electrons/holes (s)
$\tau_{A,n},\ \tau_{A,p}$	Auger lifetime of electrons/holes (s)
τ_{HL}	Carrier lifetime at high injection level (s)
τ_{eff}	Effective carrier lifetime (s)
τ_g	Generation lifetime (s)
τ_{rel}	Relaxation time (s)
Φ	Ionization integral

[1] **Remark:** In data sheets of manufacturers usually instead of V_C the symbol V_{CE} (collector–emitter–voltage), for V_G the acronym V_{GE} (IGBT) or V_{GS} (MOSFET) is used, for V_T the symbol $V_{GS(th)}$. Similar symbols are used for the current. The shorter symbols have been chosen in this work.

Chapter 1
Power Semiconductor Devices—Key Components for Efficient Electrical Energy Conversion Systems

1.1 Systems, Power Converters and Power Semiconductor Devices

In a competitive market, technical systems rely on automation and process control to improve their productivity. Initially, these productivity gains were focused on attaining higher production volumes or less (human) labor-intensive processes to save costs. Today, attention is paid towards energy efficiency because of a global awareness of climate change and, above all, questions related to increasing energy prices, as well as security of energy and increasing urbanization. Consequently, it is expected that the trend towards more electrical systems will continue and accelerate over the next decades. As a result, the need to efficiently process electrical energy will dramatically increase.

Devices that are capable of converting electrical energy from one form into another, i.e. transforming electrical energy, have been a major breakthrough technology since the beginning of electrical power systems and are considered key enabling technologies. For example, without transformers, large-scale power generation, transmission and distribution of electrical power would not have been possible. Interestingly, very few people today are aware that without this invention, initially called secondary generator [Jon04], we would not have been able to create such an efficient, safe and (locally) environmentally clean power supply system. Of course, as transformers, or generally speaking electro-magnetic devices, can only transform voltage or control reactances, their use in automation systems remained limited. At the beginning of electrification, frequency and phase control could only be realized using electro-mechanical conversion devices (i.e. motors, generators). However, these machines were bulky, required maintenance, had high losses and remained expensive. Furthermore, these electro-mechanical devices had rather low control bandwidth. Therefore, they operated mostly at fixed set points. Today, most automation and process control systems require more

© Springer International Publishing AG 2018
J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8_1

flexible energy conversion means to vary dynamically voltage or to regulate current, frequency, phase angle, etc.

At present, power electronics is the most advanced electrical energy conversion technology that attains both high flexibility and efficiency. As an engineering field, power electronics came into existence about 50 years ago, with the development and the market introduction of the so-called silicon controlled rectifier, known today as the thyristor [Hol01, Owe07]. Clearly, power electronics and power semiconductor devices are closely intertwined fields. Indeed, in its Operations Handbook, the IEEE Power Electronics Society defines the field of power electronics as *"This technology encompasses the effective use of electronic components, the application of circuit theory and design techniques, and the development of analytical tools toward efficient electronic conversion, control, and conditioning of electric power"* [PEL05].

Simply stated, **a power electronics system is an efficient energy conversion means using power semiconductor devices**. A power electronics system can be illustrated with the block diagram shown in Fig. 1.1.

A special class of power electronic systems are electrical drives. A block diagram of an electrical drive is illustrated in Fig. 1.2. Electrical drives are used in propulsion systems, power generation (wind turbines), industrial and commercial drives, for example in heating ventilation and air conditioning systems, and in motion control. In an electrical drive, the control of the electro-mechanical energy converter, the latter being a highly sophisticated load from a control perspective, is

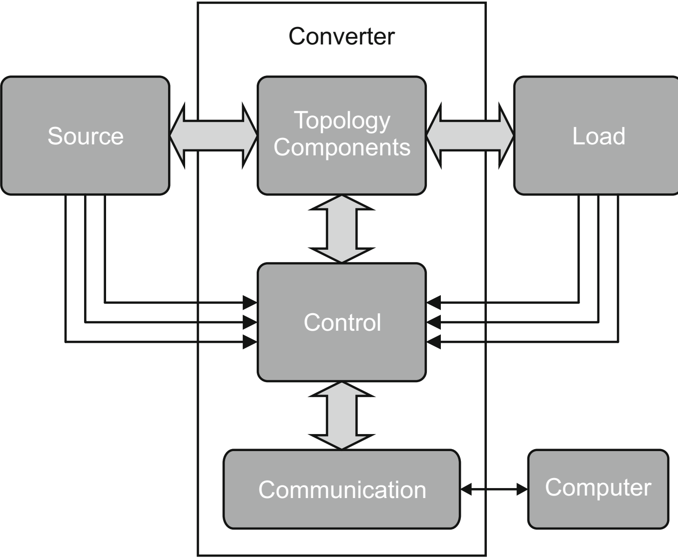


Fig. 1.1 Power electronic systems convert and control electrical energy in an efficient manner between a source and a load. Sensor interfaces to the source and load, as well as information and communication links, are often integrated

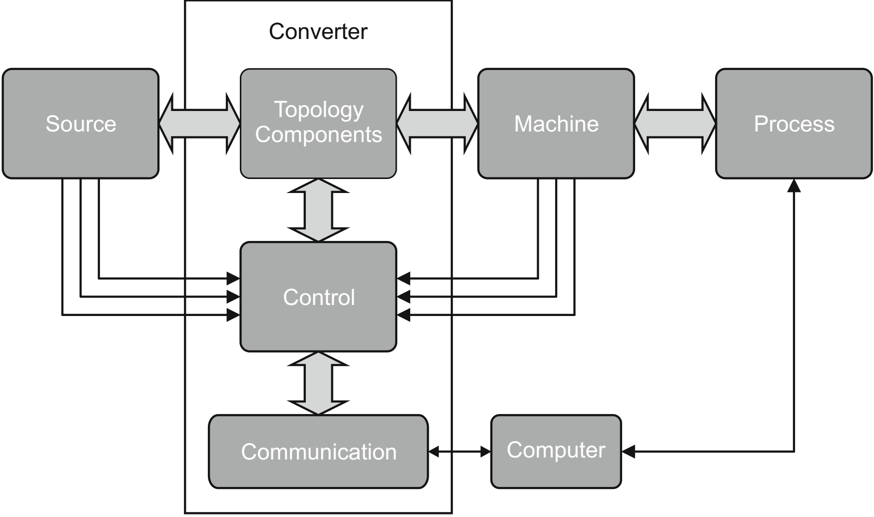


Fig. 1.2 Highly dynamic electrical drives systems comprise power electronic converters and electrical machines or actuators with dedicated control to convert electrical energy into mechanical motion

integrated in the power electronics converter control. Most research institutions that deal with power electronic converter technology also work on electrical drive technology, because this field still represents one of the largest application areas (expressed in installed apparent power) of power electronic converters [Ded06]. In the near future, despite the increased use of power converters in photovoltaic systems and computer power supplies, it is expected that this dominance of drives will remain. Most experts predict that the existing industrial markets for drives will continue to grow and will be complemented by newly developed markets, such as wind turbines, more electric ships and aircrafts and electric mobility, i.e. trains, trams, trolley busses, automobiles, scooters and bikes.

1.1.1 Basic Principles of Power Converters

Looking into the generic power electronic converter block diagram of Fig. 1.1, more details can be revealed when considering the operating principles and the topology of a modern power electronic converter. Basically, to make power electronic converters work, three types of components are needed:

- Active components, i.e. the power semiconductor components, that turn on and off the power flow within the converter. The devices are either in the off-state (forward or reverse blocking) or in the on-state (conducting).

- Passive components, i.e. transformers, inductors and capacitors, which temporarily store energy within the converter system. Based on the operating frequency, voltage, cooling method and level of integration, different magnetic, dielectric and insulation materials are used. For a given power rating of the converter, higher operating (switching) frequencies enable smaller passive components.
- Control unit, i.e. analog and digital electronics, signal converters, processors and sensors to control the energy flow within the converter such that the internal variables (voltage, current) follow computed reference signals that guarantee proper behavior of the converter according to the external commands (that are obtained via a digital communication link). Today, most control units also provide status and system level diagnostics.

As power electronic converters ought to convert electrical energy efficiently (efficiencies above 95%), linear operation of power devices is no option. Rather, the devices are operated in a switching mode. Hence, in the power supply area, to make this distinction, power converters are called "switched-mode power supplies". The basic idea behind all power converters to control and convert the electrical energy flowing through the converter is to break down this continuous flow of energy in small packets of energy, process these packets and deliver the energy in another, but again a continuous, format at the output. Hence, power converters are true power processors! In doing so, all converter topologies must respect fundamental circuit theory principles. Most importantly, the principle that electrical energy can only be exchanged efficiently via a switching network when energy is exchanged between dual components, i.e. energy stored in capacitors or voltage sources should be transferred to inductors or current sources.

As described in guidelines and standards, for example IEEE 519-1992 [IEE92] and IEC 61000 -3-6 [IEC08], and to protect sources and loads, the energy flow at the input and at the output of the converter has to be continuous, substantially free from harmonics and electromagnetic noise. To make the energy flow continuous, filter components are necessary. Note that in many applications these filter components can be part of the source or the load. To minimize cost of filter components, to comply with international standards and to improve efficiency, the control units of inverters, DC-to-DC converters and rectifiers tend to switch the power devices at constant switching frequency, using pulse width modulation (PWM) techniques, sometimes called duty-cycle control. Basic circuit theory and component design proves that higher switching frequencies will lead to smaller passive elements and filter components. Hence, all converter designs strive to increase switching frequencies to minimize overall converter costs. However, as will be discussed in the next sections, higher switching frequencies impact converter efficiency. As a result, a balance has to be found between investment material and production costs and efficiency. Note that efficiency also determines the energy costs of the conversion process over the entire life span of the converter.

1.1.2 *Types of Power Converters and Selection of Power Devices*

Power electronic converters can be categorized in various ways. Today, with power electronics it is possible to convert electrical energy from AC to DC (rectifier), from DC to DC (DC-to-DC converter) and from DC back to AC (inverter).

Although some converters can convert AC directly to AC (matrix- and cyclo-converters), most AC-to-AC conversion is done using a series connection of a rectifier and an inverter. Hence, as shown in Fig. 1.3, most converters possess at least one DC-link, where the energy is temporarily stored between the different conversion stages. Based on the type of DC-link used, the converters can be divided in current source and voltage source converters. Current source converters use an inductor to store the energy magnetically and operate with near constant current in the DC-link. Their dual, i.e. the voltage source converter, uses a capacitor to keep the DC voltage constant.

In case of AC supplies and loads, the converter could take advantage of the fact that the fundamental component of the line current or the load current crosses zero. These converters are called line-commutated or load-commutated converters and are still common in controlled rectifiers and high-power resonant converters as well as synchronous machine drives, using thyristors. A three-phase bridge rectifier is shown in Fig. 1.4. Detailed analysis shows that these converters produce line side harmonics and cause considerable lagging reactive power [Moh02]. Consequently, large filters and reactive power (so-called VAR-) compensation circuits are needed to maintain high power quality. As these filters cause losses and represent a

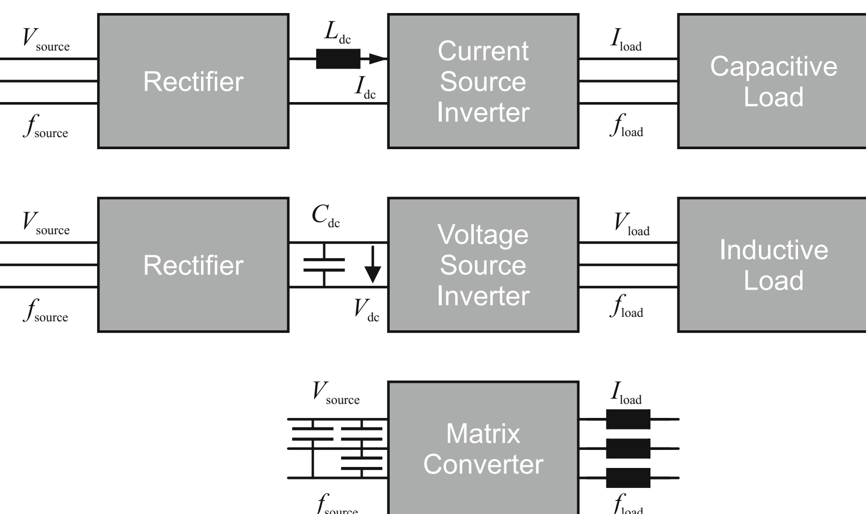


Fig. 1.3 DC-link converters and matrix converters can convert electrical power between (three-phase) AC supplies and loads. Most converters use a combination of rectifier and inverter

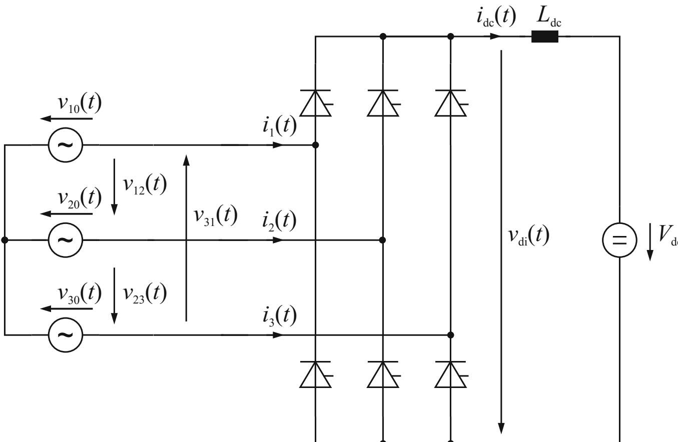


Fig. 1.4 Elementary diagram of line-commutated rectifier circuit, based on thyristors

substantial investment cost, line commutated (rectifier) circuits are slowly phased out in favor of forced commutated circuits that use active turn-off power semiconductor devices, i.e. power transistors (MOSFET, IGBT) or turn-off thyristors (GTO, IGCT). Active rectifier circuits (actually inverters operating in rectifying mode) can eliminate the need for VAR compensators and reduce or eliminate harmonic filter components.

However, not only the type of converter (rectifier, inverter or DC-to-DC converter), but also the type of topology selected (voltage source or current source) has a profound impact on the characteristics and the type of semiconductor devices that are required. A three-phase current source inverter and a three-phase voltage source inverter are illustrated in Fig. 1.5 to point out the operating differences of the devices.

In current source converters, the devices need to have forward and reverse blocking capability. These devices are called symmetrical voltage blocking devices. Although symmetrical blocking turn-off devices do exist, in practice, the reverse blocking capability is often realized by connecting or integrating a diode in series with the active turn-off semiconductor switch (transistor or turn-off thyristor). Hence, in this case, higher conduction losses must be tolerated as compared to asymmetric blocking devices. As will be shown in this book, the physics of power semiconductor switches leads to the fact that the design of symmetrical blocking turn-off devices (with integrated reverse blocking pn-junction) somehow relates to thyristor-based structures (see Chaps. 8 and 10.7). As these devices are more suitable for high power applications (voltages above 2.5 kV), some high-power converter manufacturers still use symmetrical (GCT) devices in high-power (above 10 MVA) current source converters [Zar01]. The main advantage of such converters is the fact that a current source converter is fault-tolerant against internal and external short circuits.

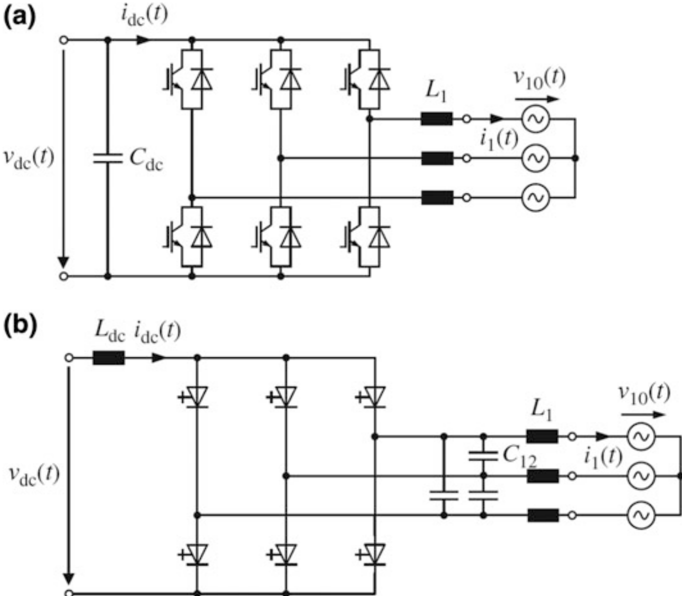


Fig. 1.5 Elementary diagrams of **a** voltage source inverter (VSI) and **b** current source inverter (CSI) circuits

Voltage source converters require a reverse conducting device because they inevitably drive inductive loads at their AC terminals. Hence, to avoid voltage spikes, when a device turns off current, a freewheeling path is needed. This reverse conduction or freewheeling capability of semiconductor switches can be realized by connecting or integrating a diode anti-parallel to the turn-off device. As this additional junction is not in series with the main turn-off device, no additional voltage drop occurs in the current path of the converter. Hence, with the present state of device technology, voltage source converters tend to be more efficient than current source converters, especially at partial load conditions [Wun03]. Indeed, at partial load the current source converter still has a high circulating current in the DC-link of the converter, while the voltage source operates at reduced current, even when the DC-link capacitor carries full voltage.

In practice, due to the lower losses in the DC-link capacitor as compared to the DC-link inductor, the size of a voltage source converter can become considerably smaller than that of a current source converter. In addition, most loads and sources behave inductively (at the switching frequency). Hence, voltage source converters may not require additional impedances or filters, while a current source converter requires capacitors at its output terminals. Taking all these engineering considerations into account, one can understand the growing importance of voltage source converters as compared to current source converters. The device manufacturers have responded to this growing market by optimizing far better asymmetric transistors and thyristors with respect to conduction and switching losses, which led to

considerable efficiency improvements and less cooling costs. Furthermore, most voltage source converter topologies use the two-level phase leg configuration that was shown in Fig. 1.5a. This phase leg topology has become so universal that device manufacturers offer complete phase legs integrated in single modules as elementary building blocks, called *power electronic building blocks* (PEBBs), thereby reducing manufacturing cost and improving reliability (see Chap. 14). As power electronics is becoming a mature technology, one can state that in the near future most new converter designs (with ratings from few mVA to several GVA) will be voltage source type converters.

1.2 Operating and Selecting Power Semiconductors

When designing a power converter, many details need to be considered to achieve the design goals. Typical design specifications are low cost, high efficiency or high power density (low weight, small size). Ultimately, thermal considerations, i.e. device losses, cooling and maximum operating temperature, determine the physical limits of a converter design. When devices are operated within their (electrical) safe operation area (SOA), conduction and switching losses dominate device losses. The underlying physics of these losses are described and analyzed in this book. However, it should be noted that the converter designer can substantially minimize these losses by making proper design decisions. In general, the outcome of the design depends greatly on the selection of:

- device type (unipolar, bipolar, transistor, thyristor) and rating (voltage and current margins, frequency range)
- switching frequency
- converter layout (minimizing parasitic stray inductances, capacitances and skin-effects)
- topology (two-level, multi-level, hard-switching or soft-switching)
- gate control (switching slew rate)
- control (switching functions, minimizing filters, EMI)

Furthermore, as losses cannot be avoided, the design of the cooling system (liquid cooling, air cooling) has a strong impact on the selection of the type of packaging. Several types of packaging technologies are currently available on the market: discrete, module and press-type packages. Whereas, the discrete and module packages can be electrically insulated, which allows all devices of a converter to be mounted on one heat sink, the press-type packages can be cooled on both sides. Typically, discrete devices are used in switched mode power supplies (up to 10 kW). Higher power levels, up to 1 MW, require parallel connection of multiple semiconductor chips and make use of the module type package, while double-sided cooled packages (single wafer disk type designs or multiple-chip press-packs) are used at the highest power levels up to several Gigawatts. Details of system architecture are discussed in Chap. 11.

As already stated, the device switching power (product of maximum blocking voltage and repetitive turn-off current) and its maximum switching frequency are important criteria for a first selection of a power semiconductor in many applications. Next to this theoretical application limit, the practical application range of silicon devices depends also on cooling limits and economical factors.

At present several device structures have been developed, each offering specific advantages. The structures of today's most important power semiconductor devices are shown in Fig. 1.6. However, it is worth mentioning that in modern applications classical bipolar transistors have been superseded by IGBTs (Insulated Gate Bipolar Transistors), being basically a MOS-controlled bipolar device. Details on each of the devices in Fig. 1.6 will be given in Chaps. 5–10.

As production of silicon devices has made great progress over the past 50 years, the application range of silicon devices has expanded and became better understood. Fig. 1.7 illustrates the practical application range of each type of silicon device in classical power converters (rectifiers and hard-switching power converters).

Note that for these applications the operation ranges are within a hyperboloid. In other words, the product of switching power (product of max. voltage and current) and switching frequency that can be attained per device in practical conversion

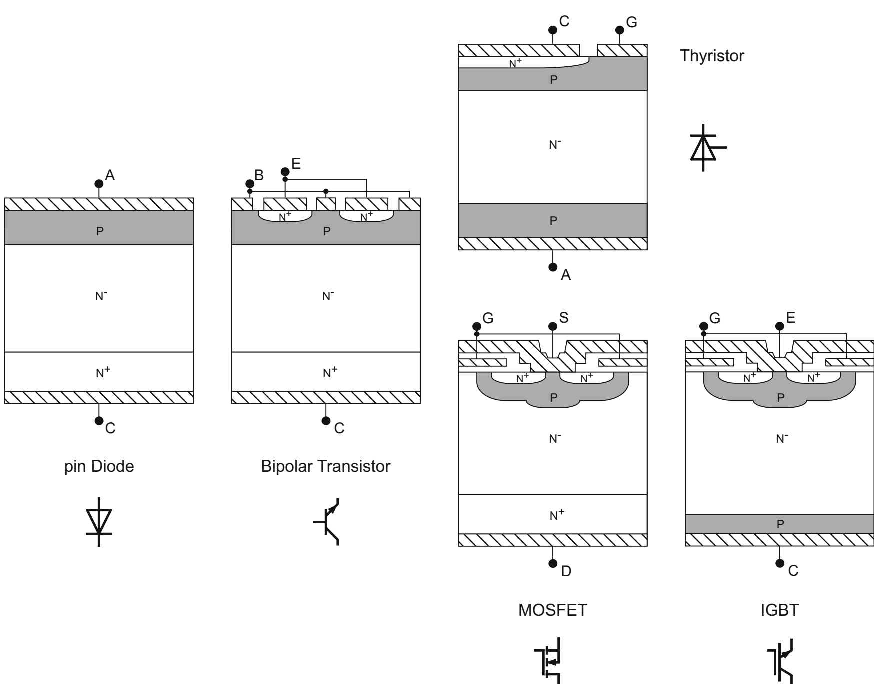


Fig. 1.6 Basic structures of common power semiconductor devices

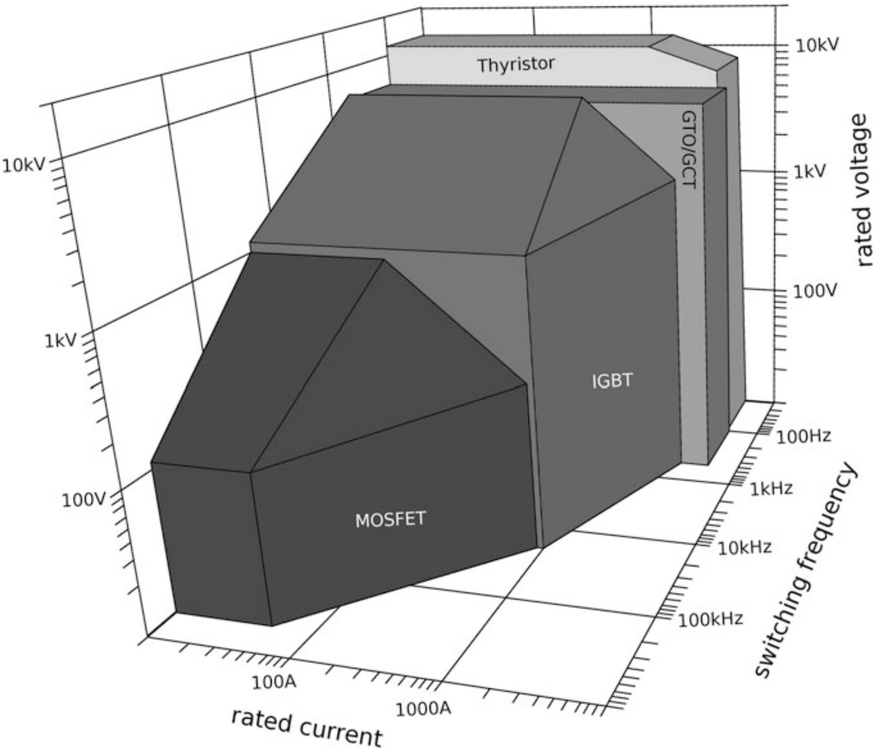


Fig. 1.7 Operating range of silicon power semiconductor devices

systems using silicon devices, assuming classical hard-switching converter configurations and similar type of cooling, appears to be fairly constant:

$$P_{\mathrm{sw-hard}} \cdot f_{\mathrm{sw}} = V_{\mathrm{max-hard}} \cdot I_{\mathrm{max-hard}} \cdot f_{\mathrm{sw}} \approx 10^{9}\ \mathrm{VA/s}$$

This frequency-power product is a good performance indicator for how well the designer was able to maximize utilization of the power semiconductors and to improve the power density of the converter. Indeed, as pointed out earlier, increasing switching frequency also reduces size of transformers, machines and filter components (at constant apparent power). Actually, if passive components of the same type (electromagnetic or electrostatic) are being considered, for example inductors, transformers and machines, they also experience a similar frequency-power product barrier as they use the same materials (copper, silicon-steel and insulation materials), operating at same maximum temperatures.

To reduce switching losses, soft-switching converters or wide bandgap materials, such as Silicon Carbide (SiC) devices, can break this technology barrier. For example, soft-switching resonant converters are being applied successfully in switched-mode power supplies and DC-to-DC converters. In soft-switching

converters, not only the switching losses of the turn-off devices are being reduced, but also the reverse recovery losses of the power diodes are mostly eliminated. As will be shown in this book, reverse recovery effects in power diodes not only increase switching losses, but also are a root cause for high HF noise (EMI) in converters. To limit these EMI effects, designers are forced to slow down switching transients, which leads to higher switching losses in hard-switching converters. As soft-switching converters utilize resonant snubber techniques, these losses do not occur and the switching frequency can be increased or, alternatively, the output power of the converter may be augmented. Soft-switching (resonant or transition-resonant) converters typically improve the frequency-power product by a factor of up to five:

$$P_{sw-soft} \cdot f_{sw} \approx 5 \times 10^9 \text{ VA/s}$$

As was stated, yet another approach to increase power density of converters is the use of SiC diodes. SiC diodes have near zero reverse recovery current. Hence, the silicon turn-off devices can be operated with higher turn-on and turn-off slew rates. These hybrid silicon-SiC designs are currently under investigation as SiC diodes are becoming available at higher power levels. Combining this hybrid concept with high-frequency soft-switching principles, i.e. using the parasitic elements of the devices (capacitances) and packages (inductances) as resonant components, the highest power densities can be attained. These concepts already find their implementation in ultra compact power supplies. Also, high-power DC-to-DC converters start to make use of these principles. One can estimate that the frequency-power product of the silicon switches in these hybrid converters can become as high as 10^{10} VA/s.

1.3 Applications of Power Semiconductors

One can conclude that the field of power electronics and of power semiconductors is still evolving at a rapid pace. Soon, all electric power will pass, not only through copper, dielectric or magnetic materials but also through semiconductors, often several times, because most applications require energy conversions or because increased efficiency is required in these energy conversion processes.

As was mentioned above, converters are being used over a wide power range, with ratings from milli-Watts or mVA (technically speaking, it is more correct to use apparent power) up to Gigawatts. Depending on the required voltage and current ratings of the power semiconductors, different types of power semiconductors are being used. At the low power end (1 VA up to 1 kVA), switched-mode power supplies for battery chargers, mostly for portable communication devices and power tools, as well as for electronic systems (audio, video and controllers) and personal computer systems form a major global market. Pushed by legislation, these power supplies have steadily augmented efficiency by improving control and

developing better power devices and passive components. Modern power supplies also have reduced standby losses. The trend is towards higher switching frequencies because less material is needed for filter components. Hence, most power supplies in this power range are using power MOSFET devices to convert electrical energy.

Another major market for power electronic systems are electronic ballasts in lighting systems. New energy efficient light sources (fluorescent, gas discharge lamps, LED, OLED) require control and conversion of the electrical power to operate. The main challenge is to develop power electronic circuits that are cheap and that can be mass-produced. Moreover, the overall life-cycle assessment (to assess impact on environment) of light sources seems to favor more efficient lighting systems [Ste02]. New legislation in the EU will phase out incandescent light bulbs.

Drive applications span a power range from few 10 VA up to 100 MVA. In automotive applications, many small drives (100 VA up to 1 kVA) are fed from the on-board power source, nominally 12 or 24 V. Hence, MOSFET devices are most common in these applications. On the other hand, grid connected drives have to cope with the different grid standard voltage levels. For example, single-phase systems for households in North America and power systems in the aircraft industry, have 115 V (rms) phase-voltage at 60 Hz or 400 Hz, respectively. Higher power single-phase systems offer 230 V line-to-line. In Europe, single-phase systems are 230 V, while three-phase line-to-line voltages equal 400 V. Canada and the US also have 460 V three-phase power systems. Typically, the highest low-voltage power systems have 660 V line voltage (IEC 60038 defines low-voltage systems up to 1000 V). To cope with all standards and to lower production costs, device manufacturers have settled on few voltage levels that cover most grid connected applications (rectifiers and inverters). Consequently, power devices with a breakdown voltage of 600, 1200 and 1700 V have been developed. As transistor type devices offer short circuit protection at low cost, IGBTs are predominantly being used in drives fed from power grids. Medium voltage drives (grid voltage from 1000 V up to 36 kV) use, depending on drive rating, transistor (IGBTs) and turn-off thyristor (GTO or GCT) type devices. Above 3 kV, i.e. at higher voltage and power ratings (above 5 MW), three-level converters [Nab81] based on GCTs seem to dominate the market. However, at very high power levels above 15 MW, load commutated inverters (LCIs) using thyristors are still produced by some manufacturers, for example in rolling mills and compressor drives [Wu08].

Drives in traction applications such as locomotives, trains and trams also face many different voltage standards. In Europe, several DC (600, 1500 and 3000 V) and AC (16.7 and 50 Hz) systems are used. Older converter designs used thyristors to control torque of various types of machines (DC, synchronous and asynchronous machines). Typically, one converter would drive multiple motors (multi-axle design). More and more, IGBT based converters are being used and single-axle designs are preferred. Hence, the required rating of the converters in traction systems has gone down, which favors designs based on transistor type devices. Most importantly, the load-cycle capacity of the converter is essential for the required reliability, especially in traction applications. In this area, research is on-going to

improve device package and cooling system reliability to reduce converter life-cycle costs (more details can be found in Chap. 14).

Yet another modern drive application at the lower power spectrum (10 W) is the electronic toothbrush. This household appliance is a true power electronics marvel. A switched-mode power supply transforms the power of the AC line (115 V or 230 V phase voltage) to medium frequency (50 kHz) AC power to allow a contactless energy transfer (via a split transformer core) to the hand-held battery fed toothbrush. A rectifier converts the medium frequency to DC. A step-down converter regulates the charging current to the battery and the electronics. An electronic commutated brushless PM machine drives the mechanical gears that move the brush in a rocking motion. Note that the complexity of this toothbrush approaches that of an electric vehicle. At these power levels, control and power devices are highly integrated to make mass production possible at reasonable cost. However, often these low power applications are precursors of what can be achieved with high-level integration at higher power levels in the future.

Power electronics is used in generator systems whenever constant speed operation of a turbine or an engine cannot be guaranteed. A typical application is maximum power point tracking of generators driven by combustion engines (10 to 1000 kW range). More recently, power generation with wind turbines is inverter driven. Power levels of wind turbines have grown from 50 kW in 1985 to 5.0 MW in 2004 [Ack05]. Wind turbine manufacturers expect off-shore wind turbines to reach 10 MW per unit in the future. These large units will be "full converter" units in contrast to the doubly-fed generators systems, that are currently mostly used in on-shore applications. Doubly-fed generators (also called rotating transformers) use AC-to-AC converters that are rated typically lower than 60% of the turbine power. This solution tends to be economically advantageous when using low-voltage (400 V or 690 V) generators, up to 5.0 MW. Note that worldwide approximately 120 GVA of inverter apparent peak power has been installed in the last decade to satisfy the demand for wind power [Wea09].

Another high-power application is transport of electrical energy over long distances using high-voltage DC (HVDC) transmission. Classical HVDC systems use three-phase bridge type rectifiers based on thyristors. Some variants use direct light triggered thyristors, although the requirement of diagnostic status feedback (via a glass-fiber, due to the high-voltage basic insulation level requirements) often favors separate light-triggered thyristors or thyristors triggered using a classical gate driver (both methods use energy stored in the snubber capacitor to trigger the thyristor via a glass fiber). The first HVDC systems date from 1977 and are still in use. However, increasing power demand over long distances (mostly hydro-power), for example in the so-called BRIC countries (Brazil, Russia, India and China) has given HVDC a new boost. HVDC technology is now operating with ±500 kV, delivering 3 GW of power, while new systems will operate at ±800 kV, transmitting 6 GW [Ast05]. These transmission systems are current source type converters and are designed to deliver power from point-to-point. Voltage source type transmission systems are being implemented in those areas, where more decentralized power generation takes place. These systems (called HVDC Light or HVDC Plus) currently use

press-pack IGBTs or IGBT modules. The functional advantages of voltage source systems, i.e. independent active and reactive power control, PWM voltage control, lower harmonics and smaller filter requirements, have enabled voltage source converter technology to compete economically against classical HVDC at power levels up to 1 GW [Asp97]. Currently, off-shore wind power plants are under construction using voltage source systems to transmit power via undersea cables.

Electrolysers for electrowinning and electroplating are yet another high power application in power electronics. Contrary to HVDC, very high DC currents at modest voltage (200 V to 500 V) have to be controlled [Wie00]. Units delivering more than 100 kA have been constructed based on thyristor rectifiers. In the future, electrolysers may play a growing role when energy from renewable power sources is converted and stored in hydrogen [Bir06].

A growing market for power electronics are converters for photovoltaic (PV) systems, especially grid connected PV systems. High efficiency, also at partial load, drives the design of PV converters. Units from 150 W (module converters), 5 kW (string converters) up to 1 MW (central converters) are being produced [Qin02]. Most designs use IGBT devices. Depending on geographical latitude, most road maps of PV cell manufacturers foresee PV at parity with electrical energy cost by 2015 (southern Europe) and 2020 (central Europe). Large-scale PV systems as well as solar thermal systems are envisaged in the near future around the Equator. To transport the electrical energy, HVDC transmission systems will be needed that span entire continents. These super-grids are under study and can be realized with today's state-of-the-art power electronics [Zha08].

The more the energy demand of the world will rely on renewable power sources, the more electrical storage capacity will be needed. High power battery storage systems are being demonstrated for over a decade in Japan using high-temperature sodium-sulfur batteries [Bit05]. Lithium-ion battery technology will further increase power density and energy density [Sau08]. Furthermore, if electric vehicles, all driven by power electronic converters, are used on a massive scale, it is anticipated that these vehicles can provide sufficient storage capacity to substantially load-level renewable power sources.

1.4 Power Electronics for Carbon Emission Reduction

Power electronics is significant for society's future. Power semiconductor components are drivers and enablers to the technological advancements that reduce carbon emissions. In Japan, it is discussed that, in the future, our way of life and practices will be an "intelligent electrified society". Power components will be responsible for the increased efficiency of the individual applications. They will serve as the key components in a variety of fields.

In 2010 Mutsuhiro Mori from Hitachi published a study called "Power Semiconductor Devices Creating Comfortable Low Carbon Society" [Mor10] with the estimation that the market volume for power semiconductors will grow 10-fold

by 2050. It contains long time technical trends. These trends, extended by further European studies like [Pop12] and other material, are briefly addressed in the following sections.

Energy Efficiency

At the opening session of APEC 2013, B.J. Baliga stated that IGBT-based power electronics has saved “75 Trillion pounds” CO_2-emission in the last 20 years. This is about 33.4 Billion tons CO_2, which is equivalent to the emissions of 390 large power plants of 1 GW each over a period of 20 years, assuming the 2013 emission factor (1 kWh = 0.596 kg CO_2) and an average load factor of 84%. Hence, power semiconductor devices are enablers for higher efficiency, leading to the fact that 390 central power plants have become expendable. However, this notion is not reflected in society. The “greenest” electricity is the one that does not need be produced.

Around 50% of the total electricity consumption occurs in electric motors in the industry and other applications. Between 40 to 50% of the applications show efficiency gains by using power electronic variable speed motor drives. Today, 80–85% of all drives are already power electronic controlled. Combining all these factors, the total electric energy savings potential with further variable speed drives is about 5–6% of the total electrical power consumption in Europe, by application of existing power devices [Pop12]. The next step is to equip motor drives with power devices that show significantly reduced losses.

About 21% of electricity is consumed for lighting [EPE07]. Replacing traditional fluorescent sources by high-efficiency ones using electronic ballast reduces the energy consumption by 61% [Pop12]. Further savings are related to new technologies based on solid-state lighting (LED) requiring effective digital controlled power converters. All this makes application of power devices necessary. A large ecological potential and strong growth area for power electronics exists in energy efficiency.

Electromobility

Mobility and traffic must be sustainable, i.e. without excessive CO_2, particle and noise emissions. Here, power electronic inverters are essential sub-systems for all transport systems of the future, whether it is e-vehicles, hybrids, rail vehicles or vehicles powered by fuel cells. At the moment, the view on future transport is too narrowly limited to electric cars. Individual mobility is an important factor in our standard of living. It is compatible with ecology if we ensure that different modes of transport—railway transport, public traffic, individual traffic complement each other.

In this scenario, the future driver will be mobile but possibly car-less. He or she uses a car when need arises. Hence, car sharing organizations are only the beginning of an alternative use of vehicles. In addition, a good railway system is a crucial factor as well. Future traffic requires a high amount of power devices. In the forecast given in [Mor10], individual mobility is seen as the largest growth volume for power devices (Fig.1.8).

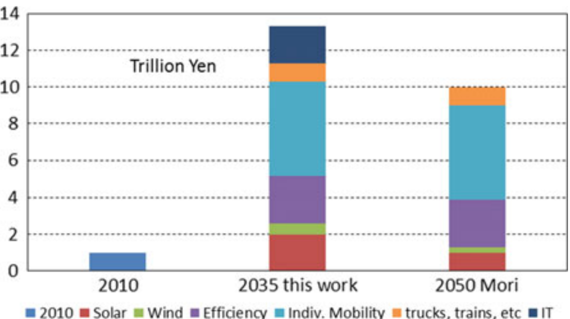


Fig. 1.8 Expected market volume of power semiconductors. Forecast 2050 by M. Mori [Mor10], forecast with higher engagement for sustainability 2035 this work: [Lut17]

Information Technology

In 2006, already 10% of the world's electricity was consumed by information technology (IT) [EPE07], while forecasts point to a strong increase, thus regarding 2017 we better assume 15%. The Kitakyushu Research Group for Sustainability estimates: until 2025 the data traffic will increase by a factor of 200, the required electricity consumption by a factor of five. From a sustainability point-of-view, this would be a disaster.

Electric energy demanded by data centers and servers in Europe was 56 TWh in 2007 and was predicted to have an increment to 104 TWh in 2020. In a typical data center, less than half of this power is delivered to the computing units, which includes microprocessors, memory and disk drives. The rest of the power is lost in power conversion, distribution and cooling [Pop12]. Conventional AC-distributed architectures suffer from low efficiency due to many conversion steps. Using high-voltage DC distribution at 400 V and high-efficiency DC-DC converters, the overall efficiency can be increased from 50 to 70%. There are further strong activities necessary to reduce energy consumptions of microprocessors etc.

Wireless communication is highly electricity consuming, mainly due to the poor efficiency of base stations. A typical base station (2007) was equipped for an input power of 2 kW, with an average power consumption of 1.3 kW. The radiation output power was 20 W, a comparatively high power for usual short transmission distances. A typical telecom radio base station with an output power of 120 W has a power consumption of more than 10 kW. This translates into a system efficiency of 1.2% [Pop12], which is quite low.

The efficiency of the power transmitter is barely 6% [Pop12]. Innovative solutions like the use of multilevel converters and a linear regulator allow the use of a switching frequency equal to the bandwidth of the envelope signal [Vas10]. For RF devices, progress with the GaN HEMT has shown that it is possible to realize highly efficient switch-mode amplifiers at microwave frequencies [Pop12].

A wide extension of IT applications is intended in every day life and the industrial production with the slogan "Internet of Things (IoT)" and "Industry 4.0", even the term "next industrial revolution" is used. Unfortunately, energy efficiency is rarely considered in the IoT and Industry 4.0 communicated outlines and road-maps. Without energy and energy efficiency considerations these "revolutions" will

conflict with ecological constraints. Moreover, the next industrial revolution must consider sustainability: From linear economy (natural resource—production—consumption—waste) to a global circular economy [Vol15].

Renewable Energy and Smart Electricity Distribution

Wind energy and solar energy need power electronic inverters. The sales of wind and solar inverter manufacturers are already a significant part of the power device companies' business, with a strong growth rate. According to the Global Wind Energy Council, a cumulated capacity of 487 GW of wind turbines was installed in 2016 [Gwe16]. Taking into account a 50-50% mix of doubly-fed and full-converter wind generators, De Doncker estimated at the IEEE PEDG 2017 meeting that this represents approximately 750 GVA of three-phase power electronic inverters. Furthermore, taking data of PowerWeb [Pow17], which shows that 285 GW of PV systems were installed globally in 2016, he estimates that approximately 315 GVA of grid-connected inverters have been installed (assuming 10% VAR compensation capacity). As a consequence, inverter cost per kVA has dropped significantly over the past decades from 500/kVA down to less than 25/kVA, which is nowadays about the same cost as a 50 Hz standard transformer. [Ded14]. According to Bloomberg, PV module cost in 2016 has dropped to below 0.22 $ per Watt [Blo17], leading to more investments in PV than in wind power. Clearly, with power electronics and PV, both being produced from silicon dioxide (i.e. sand), a sustainable, low-cost energy supply can be built that consumes less copper and steel. Hence, from a technical and economical point of view, all components and sub-systems are now available to allow an electricity supply from 100% renewable resources. Since important renewable sources are fluctuating, storage units and intelligent control are necessary. Power electronic actuators are key elements. They are required to control current flow, to adjust generation, storage and consumption, to deliver and compensate reactive power. A very effective solution for grid stabilization is the implementation of HVDC lines with the Modular Multilevel Converter in full-bridge topology based on high voltage IGBTs [Dor16]. In addition, research is on-going to explore the use of DC distribution systems at medium and at low-voltage to make distribution grids more flexible in routing energy better between decentralized power generators and "prosumers" [Ded14]. Interconnected DC distribution grids have reduced infrastructure and storage costs and operate more efficiently than classical, radial AC grids [Sti16].

Power Device Market Volume Forecast

An analysis of technical trends and forecast for market volume was made in [Mor10], the expected volume is displayed in Fig. 1.8. The study is based on the G8 Summit 2008 results, which aim at a 50% CO_2-reduction until 2050, continuing usage of nuclear energy, introducing carbon capture storage, and targets at only 33% renewable electricity. Meanwhile, the Fukushima disaster occurred. With regard to the upcoming climate disaster, which makes quality of life of future generations questionable, faster and stronger action is required. Therefore, we need to achieve these aims earlier than stipulated. In Fig. 1.8 [Lut17], the volume for

solar and wind is doubled and effort for IT is added. The forecast of [Mor10] is given and "2035 this work" from [Lut17] added. Political decisions and framework conditions are of strong influence. However, every forecast for a modern and sustainable society gives a strong increase of the volume of power devices.

While this first chapter discussed power conversion systems from the application view, we will, after discussing in depth the physics and technology of power electronic devices and components, return to the system design at the end of this book from a bottom-up perspective.

One can conclude that with power electronics, vast amounts of energy can be saved (due to efficient control of processes). In addition, power electronics is a key enabling technology to make the electrical energy supply more robust and flexible, so that a more sustainable energy supply can be realized. By definition, at the heart of power electronics are power semiconductor devices that enable this efficient energy conversion. Consequently, a deep understanding of power semiconductors is a must for any electrical engineer who wishes to contribute towards a more sustainable world.

References

[Ack05] Ackermann, T.: Wind Power in Power Systems. Wiley, Chichester (2005)

[Asp97] Asplund, G., Eriksson, K., Svensson, K.: DC transmission based on voltage source converters. In: CIGRE SC14 Colloquium, South Africa (see also library.abb.com) (1997)

[Ast05] Astrom, U., Westman, B., Lescale, V., Asplund, G.: Power transmission with HVDC at voltages above 600 kV. In: IEEE Power Engineering Society Inaugural Conference and Exposition in Africa, pp. 44–50 (2005)

[Bir06] Birnbaum, U., Hake, J.F., Linssen, J., Walbeck, M.: The hydrogen economy: technology, logistics and economics. Energy Mater. Mater. Sci. Eng. Energy Syst. **1**, 152–157 (2006)

[Bit05] Bito, A.: Overview of the sodium-sulfur battery for the IEEE stationary battery committee. Power Eng. Soc. General Meeting **2**, 1232–1235 (2005)

[Blo17] Bloomberg New Energy Finance. https://about.bnef.com/new-energy-outlook/visited. Oct 2017

[Ded06] De Doncker, R.W.: Modern electrical drives: design and future trends. In: CES/IEEE 5th International Power Electronics and Motion Control Conference (IPEMC 2006), 1, pp. 1–8 (2006)

[Ded14] De Doncker, R.W.: Power electronic technologies for flexible DC distribution grids. In: 2014 International Power Electronics Conference (IPEC-Hiroshima 2014—ECCE ASIA), Hiroshima, 2014, pp. 736–743 (2014)

[Dor16] Dorn, J., et al.: Full-bridge VSC: an essential enabler of the transition to an energy system dominated by renewable sources. In: IEEE Power and Energy Society General Meeting (PESGM) (2016)

[EPE07] EPE/ECPE.: In: Position Paper on Energy Efficiency—The Role of Power Electronics, March (2007)

[Gwe16] Global Wind Statistics 2016.: Report of the Global Wind Energy Council. www.gwec.net

[Hol01] Holonyak, N.: The silicon p-n-p-n switch and controlled rectifier (thyristor). IEEE Trans. Power Electron. **16**, 8–16 (2001)

[IEE92] IEEE 519-1992.: IEEE Recommended Practices and Requirements for Harmonic Control in Electrical Power Systems, Institute of Electrical and Electronics Engineers (1992)

[IEC08] IEC 61000-3-6.: Electromagnetic compatibility (EMC)—Part 3–6: limits—assessment of emission limits for the connection of distorting installations to MV, HV and EHV power systems. International Electrotechnical Commission (2008)

[Jon04] Jonnes, J.: Empires of Light: Edison, Tesla, Westinghouse, and the Race to Electrify the World. Random House, New York NY (2004)

[Lut17] Lutz, J.: Semiconductor power devices as key technology for a future sustainable society. In: Proceedings 7. ETG Fachtagung Bauelemente der Leistungselektronik und ihre Anwendungen, Bad Nauheim, ETG Fb. 152 (2017)

[Moh02] Mohan, N., Undeland, T.M., Robbins, W.P.: Power Electronics: Converters, Applications, and Design, 3rd edn. Wiley, New York NY (2002)

[Mor10] Mori, M.: Power semiconductor devices creating comfortable low carbon society. Hitachi, CiteWeb id: 20081073821 (2010)

[Nab81] Nabae, A., Takahashi, I., Akagi, H.: A new neutral-point-clamped PWM inverter. IEEE Trans. Indus. Appl. **1**, 518–523 (1981)

[Owe07] Owen, E.L.: SCR is 50 years. IEEE Indus. Appl. Mag. **13**, 6–10 (2007)

[PEL05] PELS Operations Handbook.: IEEE PELS Webpages. http://ewh.ieee.org/soc/pels/pdf/PELSOperationsHandbook.pdf (2005)

[Pop12] Popovic-Gerber, J., et al.: Power electronics enabling efficient energy usage: energy savings potential and technological challenges. IEEE Trans. Power Electron. **27**(5), 2338–2353 (2012)

[Pow17] Powerweb—Forecast International’s Energy Portal. http://www.fi-powerweb.com

[Qin02] Qin, Y.C., Mohan, N., West, R., Bonn, R.H.: Status and needs of power electronics for photovoltaic inverters. Sandia National Laboratories Report, SAND2002, pp. 1535 (2002)

[Sau08] Sauer, D.U.: Storage systems for reliable future power supply networks. In: Droege, P. (ed.) Urban Energy Transition—from Fossil Fuels to Renewable Power. Elsevier, pp. 239–266 (2008)

[Ste02] Steigerwald, D.A., Bhat, J.C., Collins, D., Fletcher, R.M., Holcomb, M.O., Ludowise, M.J., Martin, P.S., Rudaz, S.L.: Illumination with solid state lighting technology. IEEE J. Sel. Top. Quantum Electron. **8**, 310–320 (2002)

[Sti16] Stieneker, M., De Doncker, R.W.: Medium-voltage DC distribution grids in urban areas. In: 2016 IEEE 7th International Symposium on Power Electronics for Distributed Generation Systems (PEDG), Vancouver, BC, pp. 1–7 (2016)

[Vas10] Vasic, M., Garcia, O., Oliver, J.A., Alou, P., Diaz, D., Cobos, J.A.: Multilevel power supply for high-efficiency RF amplifiers. IEEE Trans. Power Electron. **25**(4), 1078–1089 (2010)

[Vol15] Volkert, C.A.: Wirtschaft gegen Umwelt: Grundsatzkritik an der Wegwerfproduktion. Tagungsband Offene Akademie (2015)

[Wea09] World Wind Energy Association.: World Wind Energy Report 2008. http://www.wwindea.org. (2009)

[Wie00] Wiechmann, E.P., Burgos, R.P., Holtz, J.: Sequential connection and phase control of a high-current rectifier optimized for copper electrowinning applications. IEEE Trans. Indus. Electron. **47**, 734–743 (2000)

[Wu08] Wu, B., Pontt, J., Rodríguez, J., Bernet, S., Kouro, S.: Current-source converter and cycloconverter topologies for industrial medium-voltage drives. IEEE Trans. Indus. Electron. **55**, 2786–2797 (2008)

[Wun03] Wundrack, B., Braun, M.: Losses and performance of a 100 kVA dc current link inverter. In: Proceedings of European Power Electronics Association Conference EPE 2003, Toulouse, topic 3b, pp. 1–10 (2003)

[Zar01] Zargari, N.R., Rizzo, S.C., Xiao, Y., Iwamoto, H., Satoh, K., Donlon, J.F.: A new current-source converter using a symmetric gate-commutated thyristor (SGCT). IEEE Trans. Indus. Appl. **37**, 896–903 (2001)

[Zha08] Zhang, X.P., Yao, L.: A vision of electricity network congestion management with FACTS and HVDC. In: IEEE Conference on Electric Utility Deregulation and Restructuring and Power Technologies, 3rd conference, pp. 116–121 (2008)

Chapter 2
Semiconductor Properties

2.1 Introduction

Research on semiconductors has a long history [Lar54, Smi59]. Phenomenologically they are defined as substances whose electrical resistivity covers a wide range, about 10^{-4} to 10^{9} Ωcm, between that of metals and insulators and which at high temperatures decreases with increasing temperature. Other characteristics are light sensitivity, rectifying effects and, most typical, an extreme dependency of the properties even on minute impurities. After reaching a basic understanding of their physical nature in the 1930s and 1940s, semiconductors are defined now often by the energy band model and impurity levels leading to the observed phenomena: Semiconductors are solids whose conduction band is separated from the valence band by an energy gap E_g and at sufficiently low temperatures is completely empty, whereas all states of the valence band are occupied. Most important for application in devices, however, is that the conductivity can be controlled over a wide temperature range by impurities, and that there are two types of impurities, donors which release electrons causing n-type conductivity and acceptors which provide positive carriers, the holes, leading to p-type conductivity. This allows the fabrication of pn-junctions.

Semiconductors for power devices must have a sufficiently large energy gap or 'band gap' to insure that the intrinsic carrier concentration, present also without doping, stays below the doping concentration of the weakest doped region up to a sufficient operation temperature, for example 450 K. This is the precondition that the doping structure remains effective. A sufficient band gap is also advantageous because the critical field strength, up to which the material can withstand electric fields without breakdown, increases with band gap. The critical field is a key parameter for the design and function of power devices. Unnecessary large energy gaps, on the other hand, can prove disadvantageous because the ionization energy of the impurities becomes larger with increasing band gap and hence they release only an unfavorably low number of carriers at room temperature. Also built-in and threshold voltages get larger with wider band gap.

© Springer International Publishing AG 2018

J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8_2

Other properties of the semiconductor are also important. An essential demand is first that the mobilities of free electrons and holes are sufficiently high. Furthermore most power devices based on high carrier injection require a semiconductor with an indirect bandgap such as silicon (see Sect. 2.4). Only these semiconductors allow a sufficient control of the excess carrier lifetime, which determines dynamical and stationary characteristics. Of course also the properties of impurities needed for the pn-structure and adjusting the lifetime are absolutely essential. As a significant chemical semiconductor property we mention the stable native oxide of silicon, which is a prerequisite for almost the whole modern semiconductor technology. A detailed description of the relevant semiconductor properties follows in the present Chap. 2, but their impact on device characteristics and technology will come up at many points in the book.

Semiconductors for power devices are always mono-crystalline, of single crystal form, because, first, only single crystals guarantee a homogeneous space charge under blocking bias and as few as possible energy levels in the band gap. High blocking voltages and low leakage currents are possible only under these conditions. Second, carrier mobilities in single crystal semiconductors are much higher than in polycrystalline. This allows correspondingly higher current densities and the use of smaller devices.

The advantages obtained passing from poly- to mono-crystalline material can be seen looking at the first commercial semiconductor power rectifiers made of poly-crystalline selenium. These rectifiers were produced from about 1940 until 1970 [Spe58, Pog64]. With current densities of at best 1 A/cm^2 and blocking voltages up to 40 V per cell, packages with relatively large volume were necessary to handle the given currents und voltages. The properties were to a good part due to the poly-crystalline state, although partly also to the semiconductor Se itself. Silicon diodes replacing Se rectifiers (and Ge diodes) from the end of the 1950s allow current densities and blocking voltages more than two orders of magnitude higher.

The era of modern semiconductors started at the end of the 1940s with the advent of hyper-pure single crystal germanium together with a breakthrough in understanding pn-junctions and the invention of the transistor. Ge was replaced by silicon in the 1950s, and nowadays it has no importance for power devices. Its main drawback is the relative small band gap of only 0.67 eV (at 300 K) which results in a low allowed operating temperature ($\approx$70 °C). Si is by far the most commonly used semiconductor also for power devices, and one can read more about it in this book. The band gap of 1.1 eV and other properties of Si are very suitable for most applications. Si belongs to the group IV of the periodic system, following carbon, C, and preceding Ge. A compound semiconductor consisting of the group III element gallium and the group V element arsenic is gallium-arsenide, GaAs. This semiconductor has gained significance for microwave devices, since it allows high switching frequencies. It has a bandgap of 1.4 eV and a very high electron mobility (five times that of Si). In the field of power devices, the high electron mobility renders it suitable for high voltage Schottky diodes. Schottky diodes of GaAs for 300 V and higher are available on the market.

A compound semiconductor, which after much research and development efforts in the last decades has gained now high significance for power devices, is silicon carbide, SiC [Fri06]. It has a large band gap ranging from 2.3 to 3.3 eV depending on the crystal modification (polytype, see the next section). Consequently, the maximum operating temperature up to which the intrinsic carrier concentration is small enough, and the critical field strength at which breakdown occurs are much higher than for Si. Whereas the higher possible temperature could not be utilized much till now because metallurgical properties of contacts and packaging limits the operating temperature, extensive use was made of the benefits of the high critical field, which are a much higher allowed doping concentration and smaller width of the base region required for a given blocking voltage. This means that fast Schottky diodes and other unipolar devices for high voltages can be fabricated, and pn-devices with extremely high blocking voltage (possibly up to 20 kV) and low power losses are possible. The main advantage of SiC devices is that they allow higher operation frequencies. A difference to Si technology is that diffusion of impurities cannot be used as a doping method because the diffusion constants in SiC are too small. Although the technology of SiC is less versatile and developed and more expensive than that of Si, the potential of SiC devices in power electronics is high. Not only Schottky diodes and MOSFETs for currents up to about 50 A and voltages up to 1700 V are on the market but also "full SiC" power modules for hundreds of amperes and voltages up to 1200 V and more. A phase of intensive testing is going on with the hope of widespread applications. To take advantage of the very small switching times, gate circuits and electronic environment have to be fundamentally readapted to keep inductive voltage peaks and high-frequency oscillations below a dangerous level. We will deal with SiC devices at the proper points of the book.

In recent years, intensive investigations have been performed on microwave and power devices on the base of GaN [Ued05]. This is a semiconductor which with regard to the bandgap of 3.4 eV and a correspondingly high critical field is similar to 4H-SiC. Although the technology has profited greatly from the developments in the LED area, the fabrication of large single crystals and wafers of GaN offers still problems. GaN-devices demonstrated till now are lateral devices using GaN films grown epitaxially on substrates of silicon, 4H-SiC or saphire (Al_2O_3). 8-inch GaN-on-Si wafers can meanwhile be produced. A basic element of GaN devices is a highly conducting electron layer ('two-dimensional electron gas', 2DEG) induced by the high electric polarization of GaN and AlGaN which latter is deposited as a thin layer on the GaN. Owing to the high critical field the dimensions can be chosen small, so a small on-resistance can be obtained per area. Field effect transistors with a blocking voltage up to 1.8 kV and a current of more than 100 A have been demonstrated [Ike08]. The main interests are at mid-range voltages (<1 kV) for high switching frequency applications. Special devices of the 600 V class are on the market [Mor14, Ish15, Hon15]. Compared with SiC the developments are still in an earlier stage.

A further subject of investigations in view of a possible future material for power devices has been the diamond modification of carbon. Diamond has an energy gap

of 5.5 eV and in pure form is an insulator. Its critical field is extremely high (10–20 MV/cm). If it contains impurities like boron or phosphorus it behaves like a semiconductor whose carriers have higher mobilities than in silicon. A handicap of diamond is the high ionization energy of the dopants, especially of donors. Hence the impurities are ionized only to a small extent at room temperature. Concepts to partially overcome this drawback have been proposed and samples of high-voltage Schottky diodes have been demonstrated [Twi04, Ras08]. The employment of diamond devices in power electronics is expected at best in the longer term.

2.2 Crystal Structure

The atoms in a crystal are arranged in a three-dimensional lattice at lattice points around which they can vibrate. The lattice is to a great extent determined by the type of bonding between the atoms. In all the named semiconductors (apart from Se) the bonds in all directions are (widely) covalent, i.e. formed by an electron pair with opposite spins. This is the rule for group IV elements. Because they have four valence electrons in the outer shell, an atom can undergo covalent bonds with four nearest neighbors which, like the central atom, contribute an electron to the respective bond. In this way, a closed shell of eight shared electrons is gained for each atom establishing a very stable configuration. Similar considerations hold for III–V compounds such as GaAs whose constituents *on average* have four valence electrons per atom. Symmetrically arranged the four nearest neighbors are located at the corners of a tetrahedron the midpoint of which is occupied by their common bonding partner. For the element semiconductors, this tetrahedral bonding can be realized only by one lattice structure, the diamond lattice. This is hence the lattice of Ge, Si and the diamond modification of carbon.

Figure 2.1a shows the cubic unit cell of the diamond lattice. It can be described as superposition of two face-centered cubic (fcc) sublattices, which by definition have lattice points at the corners and midpoints of the side faces of the cube. The two sub-lattices are shifted by a quarter of the space diagonal. In the lower left part of the figure, the covalent bonding of an atom on the space diagonal with its four nearest neighbors is marked. This part of the figure is redrawn in Fig. 2.1b to show the tetrahedral bonding geometry more clearly.

GaAs has the same lattice structure, except that now one of the face-centered cubic sub-lattices is occupied by Ga and the other by As atoms. If in Fig. 2.1b an As atom is in the center, the four neighbors at the corners are Ga and vice versa. This so-called zincblende lattice is the structure of most III–V semiconductors [Mad64].

An exception is GaN which crystallizes preferably in the wurtzite structure, a hexagonal lattice called after a second polytype of ZnS appearing besides zinc-blende. The also appearing zincblende modification of GaN is less stable and not used in devices. In the wurtzite lattice the tetrahedral bonding between the neighbors of both types is realized as well, only the orientation of neighboring tetrahedra

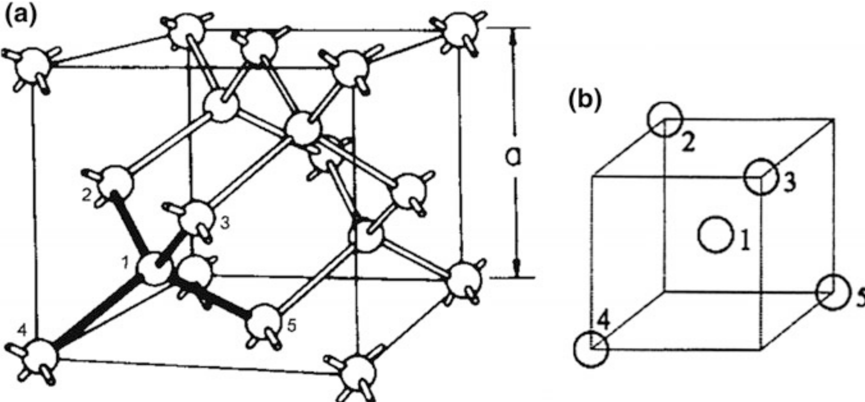


Fig. 2.1 **a** Cubic unit cell of the diamond lattice. The tetrahedral bonding between an atom 1 on the space diagonal and four nearest neighbors 2–5 is marked. **b** Central atom with four neighbor atoms at the corners of a tetrahedron. The lattice constant for silicon is a = 5.43Å, the distance between the center of Si atoms $d = 3^{1/2}/4\ a$ = 2.35 Å (adapted from [Sze81] and [Hag93])

towards each other is different from the zincblende structure. Actually, the tetrahedral bonding system of XY-compounds is compatible with many other lattices. This is shown in fact by SiC which crystallizes in the zincblende structure and likewise in many other polytypes, most of which are hexagonal. In all cases, the arrangement of nearest neighbors is identical, each silicon atom being surrounded by four carbon atoms at the corners of a tetrahedron with the Si in the center, and vice versa. Only the arrangement of more distant atoms varies from polytype to polytype [Mue93]. The atomic distance between nearest neighbors is always 0.189 nm = 1.89 Å, nearly the mean value between the atomic distances in diamond with 1.542 Å and Si with 2.35 Å. For power devices, the hexagonal polytype called 4H-SiC is preferred. The prefix indicates that the structure repeats itself after stacking of four Si-C double planes. In this nomenclature the wurtzite polytype of GaN has a 2H-structure, since the lattice repeats itself after two double planes of Ga-N.

The used crystal structures are compiled in the following list:

Ge, Si, diamond-C	Diamond lattice, cubic
GaAs	Zincblende lattice, cubic
GaN	Wurtzite lattice, 2H-hexagonal
4H-SiC	4H-hexagonal

The crystal orientation influences to some extent processing parameters and physical properties. In silicon, the thermal oxidation rate, the epitaxial deposition rate, the density of surface states and the elastic constants depend on orientation, whereas mobilities and diffusion constants due to the cubic lattice are (isotropic) scalars. To describe crystal planes and directions one uses the Miller indices. These

are defined via the reciprocals of the intercepts of the considered plane with the crystallographic axes, expressing their ratios by a triple of smallest integers. For example, the six faces of a cube intersect two of the axes in the point ∞, hence two of the Miller indices are 0, and the planes are denoted as (1 0 0), (0 1 0), (0 −1 0) and so on. The family of these planes is denoted as {100} planes. The crystal planes which have equal intercepts on the axes are the {111} planes. In silicon technology crystal wafers (slices) with a {100} or a {111} surface are used.

2.3 Energy Gap and Intrinsic Concentration

The energy gap results because all the valence electrons of the semiconductor atoms are required for the complete covalent bonds and because a certain amount of energy is necessary for breaking electrons out of the bonds so that they can move freely and hence conduct current. This is expressed by the band diagram shown in Fig. 2.2, where the energy of the electrons is plotted versus a space coordinate x.

The valence band contains the states for the electrons in the bonds, and the conduction band represents the states of the electrons which are free for conduction. Between the top of the valence band, E_V, and the bottom of the conduction band, E_C, no energy levels are present, $E_g = E_C - E_V$ is the energy gap. Here we refer to an "intrinsic" semiconductor whose properties are not influenced by impurities but are due to the semiconductor itself. At least the energy E_g is necessary to break an electron out of the bonds and create a conduction electron. At zero absolute temperature, this energy is not available for the valence electrons (external activation excluded), and since no carriers are present, the semiconductor behaves like an insulator. At $T > 0$ a certain number of valence electrons is elevated into the conduction band. By this process, however, not only conduction electrons are created, but also an equal number of empty electron places remain in the valence band and these voids, called "holes", can conduct current likewise. As a missing valence electron the hole has the opposite, i.e. a positive charge. Because in an electric field the empty place is filled repeatedly by a neighboring valence electron, the hole travels in opposite direction to the electrons. This is similar to a bubble in water moving against gravity. It follows, that one can ascribe also an energy to the

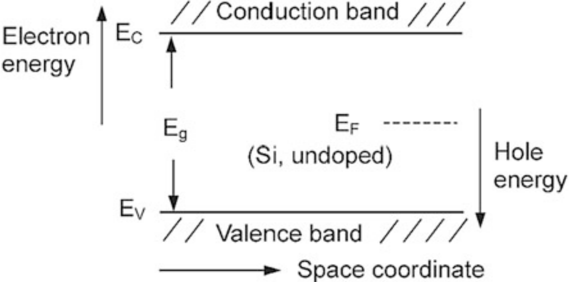


Fig. 2.2 Energy band model

holes which is directed downward in Fig. 2.2 (see the calculation below). So far, the picture of a (classical) vacancy at a point x in the otherwise full band of valence electrons is successful. With this understanding the hole is an auxiliary quantity for a simpler description of the motion of all involved valence electrons which are the actual carriers. However there are other decisive experiments, particularly measurements of the Hall effect, which contradict this classical picture and show the holes as positive charge carriers on their own. This is confirmed by quantum theory according to which the holes are independent stable (quasi-) particles on the same grounds as conduction electrons. A detailed discussion of these features will be given in the next paragraph.

We calculate now the intrinsic concentration of carriers in a pure semiconductor and introduce simultaneously some more general relationships for electrons and holes which are applicable also when impurities are present. Since the thermal generation of carriers is counteracted by recombination annihilating them, this leads to a thermal equilibrium described by statistical physics. The occupation probability of states with energy E, defined as the number of electrons n_{E} per number of states with that energy, N_{E}, is given by the Fermi distribution

$$\frac{n_E}{N_E} = \frac{1}{1 + e^{\frac{E-E_F}{kT}}} \tag{2.1}$$

In this equation k denotes the Boltzmann constant, T the absolute temperature and $\mathrm{E_F}$ the Fermi level, which in statistical thermodynamics is known under the name "chemical potential". Fermi energy $\mathrm{E_F}$ is a constant of the system determined so that the sum over n_{E} returns the total electron concentration. Equation (2.1) is a consequence of the quantum mechanical principle, that a state cannot be occupied by more than one electron. As shown by the equation, the states with energy smaller than $\mathrm{E_F}$ are mostly occupied by an electron, while the states with $E > E_F$ are mostly empty. The number of occupied divided by the number of unoccupied states of energy E, called the occupation degree, is

$$\frac{n_E}{N_E - n_E} = e^{-(E-E_F)/kT} \tag{2.2}$$

In the intrinsic case and often also in doped semiconductors (see next chapter), the occupation probability is small, $n_E << N_E$. In this so-called non-degenerate case, (2.2) turns into the classical Boltzmann or Maxwell-Boltzmann distribution:

$$\frac{n_E}{N_E} = e^{-(E-E_F)/kT} \tag{2.3}$$

Since primarily only states near the bottom of the conduction band are occupied and primarily only states near the top of the valence band are empty, we assume at first that the states in the bands are concentrated at the edges and their number in integrated form is given by effective densities of states (number per volume) N_C, N_V.

Taking (2.3) at the energy E_C one obtains for the concentration of conduction electrons in thermal equilibrium:

$$n = N_C \cdot e^{-(E_C - E_F)/kT} \tag{2.4}$$

Using (2.2) at $E = E_V$ and considering that the density of unoccupied states in the valence band is identical with the hole concentration p, whereas the density of occupied states is $N_V - p \approx N_V$, the hole concentration in thermal equilibrium is obtained as follows:

$$p = N_V \cdot e^{-(E_F - E_V)/kT} \tag{2.5}$$

This shows that the holes behave statistically as particles, with energy scale inverted compared with electrons. Multiplication of (2.4) and (2.5) using $E_C - E_V = E_g$ yields:

$$n \cdot p = n_i^2 = N_C N_V \cdot e^{-E_g/kT} \tag{2.6}$$

where $n_i = n = p$ is the intrinsic concentration. Equation (2.6) is the mass law equation of the reaction $(0) \leftrightarrow n + p$, describing generation and, inversely, the recombination of an electron-hole pair.

Since the condition $n = p = n_i$ has not been used in the derivation, the intrinsic conduction represents only a special case of (2.4) (2.5) and (2.6), actually they are applicable also to doped semiconductors where $n \neq p$. Doped semiconductors will be treated in detail in one of the next sections. Only the Boltzmann distribution (2.3) is assumed for thermal equilibrium, meaning that the doping is not too high (case of non-degeneracy). As shown by (2.6), the np product is a constant independent of the Fermi level, but dependent on the bandgap and temperature.

The Fermi level for the intrinsic case is obtained from (2.4) and (2.5) setting $n = p$:

$$E_i = \frac{Ev + Ec}{2} - \frac{kT}{2} \cdot \ln \frac{Nc}{Nv} \tag{2.7}$$

Because of similar values of the densities of states N_C, N_V, the Fermi level in intrinsic semiconductors is located close to the middle of the bandgap.

In spite of the simplifying assumption on the distribution of the states in the bands, (2.4), (2.5) and (2.6) are applicable to the actual situation. To take into account that with increasing T more states above respectively, below the band edges are occupied, this results only in a temperature dependency of the effective densities of states N_C and N_V. The density of states N_E increases with distance ΔE from the edges as $\sqrt{\Delta E}$. Multiplying this with the Boltzmann factor of Eq. (2.3) and integrating, one obtains again (2.4) and (2.5), where N_C, N_V now are proportional to $T^{3/2}$ [Sze02]. Considering that also the band parameters themselves vary a little with T, one obtains for Si [Gre90]:

$$N_C = 2.86 \times 10^{19} \left(\frac{T}{300} \right)^{1.58} /\text{cm}^3$$
$$N_V = 3.10 \times 10^{19} \left(\frac{T}{300} \right)^{1.85} /\text{cm}^3 \qquad (2.8)$$

These numbers are large compared with the doping concentrations in most cases, as will be seen later. Compared with the number of Si atoms per cm^{-3}, 5.0×10^{22}, they are small.

The bandgap is approximately a constant. However, seriously considered, it decreases slightly with temperature. This can be expressed for Si and other semiconductors in the form [Var67]:

$$E_g(T) = E_g(0) - \frac{\alpha \cdot T^2}{(T + \beta)} \qquad (2.9)$$

The bandgap parameters of this equation together with the effective densities of states are compiled for Si, GaAs, 4H-SiC and GaN in Table 2.1 [Thr75, Gre90, Lev01, Mon74].

The intrinsic carrier densities calculated with these data for Si, GaAs, 4H-SiC are shown in Fig. 2.3 as functions of temperature. Also germanium is included. From Si to SiC the intrinsic concentration decreases extremely for a given temperature due to the exponential dependence on energy gap. Considering a given value of n_i, the absolute temperature at which this value is adopted is proportional to E_g, if the temperature-dependence of the pre-exponential factor in (2.6) is neglected.

The temperature at which n_i reaches a value comparable with the impurity concentration of the lowest doped region in a device, defines a limit, above which the pn-structure begins to be leveled out and loses its normal function. If n_i is dominating, its exponential increase with temperature and the corresponding decrease of resistance, together with thermal feedback, can lead to current constriction and destruction of the device. In a Si device with 1000 V blocking voltage, a doping in the range of 10^{14} cm^{-3} is necessary for the base region to sustain the voltage. To meet the condition $n_i < 10^{14}$ cm^{-3}, the temperature must remain below about 190 °C, as is shown in Fig. 2.3. With 4H-SiC, temperatures of more than

Table 2.1 Bandgap parameters and effective density-of-states of some semiconductors

	Si	GaAs	4H-SiC	GaN
$E_g(0)$ (eV)	1.170	1.519	3.263	3.47
α*10^4 (eV/K)	4.73	5.405	6.5	7.7
β (K)	636	204	1300	600
$E_g(300)$ (eV)	1.124	1.422	3.23	3.39
$N_C(300)$ (cm^{-3})	2.86×10^{19}	4.7×10^{17}	1.69×10^{19}	2.2×10^{18}
$N_V(300)$ (cm^{-3})	3.10×10^{19}	7.0×10^{18}	2.49×10^{19}	4.6×10^{19}

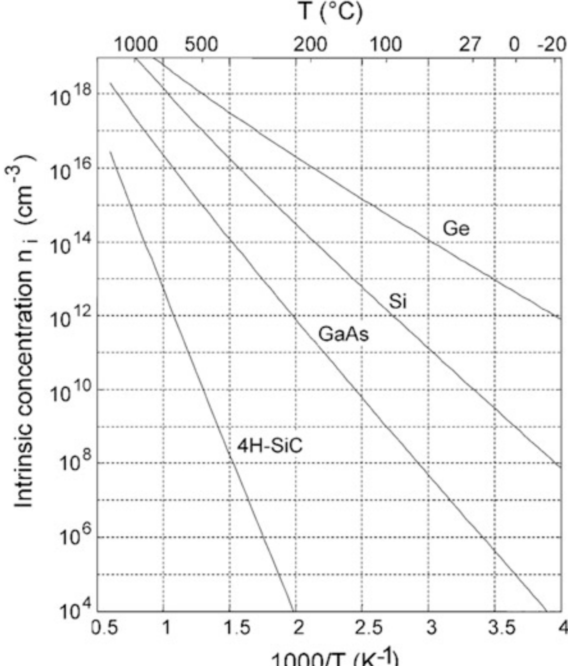


Fig. 2.3 Intrinsic carrier density for Ge, Si, GaAs and 4H-SiC as a function of temperature

800 °C would be allowed for 1000 V devices from this requirement. In practice, interconnect and packaging materials, however, set a much lower temperature limit.

Equation (2.6) shows why semiconductors with a wide bandgap, if nearly intrinsic, behave like insulators. In 4H-SiC, the intrinsic concentration as given by above data is even at 400 K only 0.3 cm^{-3}, corresponding to a resistivity of $\approx 2 \times 10^{16}$ Ωcm. Although the practically attainable resistivity is several orders of magnitude smaller, SiC can be used for insulating layers with high thermal conductance.

For later use in the case of doped semiconductors we note an alternative form of (2.4) and (2.5) which is obtained using the intrinsic concentration and intrinsic Fermi level to eliminate the effective densities of states and energies of band edges. Dividing (2.4) by the specialized form $n_i = N_C \exp(-(E_C - E_i)/kT)$ it converts to:

$$n = n_i \cdot e^{\frac{E_F - E_i}{k \cdot T}} \tag{2.10}$$

The hole concentration in thermal equilibrium can be written as follows:

$$p = n_i \cdot e^{\frac{E_i - E_F}{k \cdot T}} \tag{2.11}$$

2.4 Energy Band Structure and Particle Properties of Carriers

Besides the E(x) band diagram of Fig. 2.2, there is the more detailed energy band representation in k-space, $E(\vec{k})$, which allows further insight into fundamental semiconductor properties. Here the electron energy is plotted versus the wave vector $\vec{k}$ of a wave packet, which solves the quantum mechanical Schrödinger equation for an electron in the lattice-periodic potential of the crystal. Of main interest are the valence $E(\vec{k})$ band with highest maximum and the conduction $E(\vec{k})$ band with lowest minimum. In Fig. 2.4 these bands are shown for Si and GaAs for specified directions in $\vec{k}$-space.

The energy difference between the absolute minimum of the conduction band and the maximum of the valence band is the band gap E_g shown in Fig. 2.2. The maximum of the valence band is nearly always at $\vec{k} = 0$. In GaAs, also the minimum of the conduction band is located at this position. A semiconductor of this kind is called a direct semiconductor. In Si, the conduction band has a minimum far away from $\vec{k} = 0$. Semiconductors of this type are called indirect. Since in Si a minimum lies in each of the {100} directions, there are six minima in a unit cell.

Whether a semiconductor has a direct or indirect band gap is decisive for the probability of transitions between the bands. This determines the suitability for

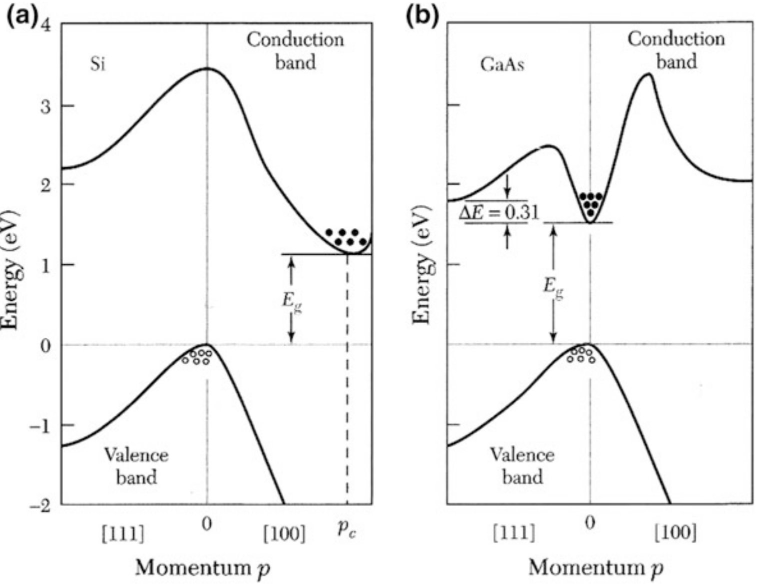


Fig. 2.4 Energy band E(k) for Si (indirect semiconductor) and GaAs (direct semiconductor). Figure from Sze, [Sze02], reproduced with permission of John Wiley & Sons Inc.

optical devices, but has some significance also for power devices. The influence on transition probability follows because the crystal momentum $\vec{p} = \hbar\vec{k}$ ($h = \hbar/2\pi, = h/2\pi, h$ is Planck's constant) has the property of a momentum of the electron regarding its reaction on external forces. Internal forces arising from the periodic potential in the crystal are taken into account in another way (see below). An external force $\vec{F}$, owing to an electric or magnetic field, causes an acceleration which, as in Newton's second law of mechanics, is given by $d\vec{p}/dt = \vec{F}$. In transitions between bands the change in crystal momentum $\vec{p}$ has to be counted in the law of conservation of momentum. The recombination of an electron at the bottom of the conduction band with a hole at $\vec{k} = 0$ occurs under emission of a photon, which receives nearly the whole energy released, but has negligible momentum. In indirect semiconductors the recombination can occur only if the crystal momentum of the conduction electron can be transferred to a quantum of lattice vibrations, a phonon. Hence, compared with a direct semiconductor such as GaAs, where the recombination occurs without participation of a phonon, the radiative band-to-band recombination in indirect semiconductors has a much lower probability. Therefore, only direct semiconductors are used for LEDs and lasers. The lifetime associated with the radiative band-to-band recombination represents an upper limit of the lifetime of minority carriers. For GaAs, a radiative minority carrier lifetime $\tau = 1/(B \cdot N)$ (B is the recombination constant) of 6 µs is estimated for a doping concentration of $N = 1 \times 10^{15}$ cm^{-3}; for 10^{17} cm^{-3} this means $\tau = 60$ ns [Atk85]. This is sufficient to allow satisfactory operation of pin diodes for a medium voltage range, but it is detrimental for bipolar transistors and thyristors with commonly used doping structures. In Si the radiative recombination constant B is four orders of magnitude smaller [Sco74]. Hence very high lifetime values are possible in Si. The recombination radiation in Si, having a wavelength $\lambda \cong hc/E_g = 1.1$ µm, is often used as a tool to investigate and test the internal operation of devices. Besides Si also Ge and all polytypes of SiC are indirect semiconductors. Some of the III–V compound semiconductors are of the direct type like GaAs and GaN, while others are of the indirect type like GaP.

The $E(\vec{k})$ bands are also basic for the behavior of electrons and holes as charge carriers. We will outline this here shortly, referring for a more complete discussion to the books of Moll [Mol64] and Spenke [Spe58]. If an external force is applied, the electrons near the minimum of the conduction band are accelerated. However, the increase in kinetic energy is only relatively small, because the acceleration is stopped after a short relaxation time owing to the non-ideality of the crystal, i.e. by scattering by phonons and impurities. Thus the wave vector and kinetic energy are reduced nearly to their initial values (on the statistical average) and remain in the band not far away from the minimum. Hence the kinetic energy, $E_{n,kin} = E - E_C$, can be expressed using Taylor's expansion as

$$E_{n,kin} = \frac{1}{2}\frac{d^2E}{dk^2} \cdot (\vec{k} - \vec{k}_m)^2 \quad (2.12)$$

where $\vec{k}_m$ denotes the $\vec{k}$-vector at the band minimum. We assume for simplicity that the $E(\vec{k})$ function depends only on the absolute value k of $\vec{k}$, and not on the orientation. Defining a "particle momentum" as $\vec{p}_n = \hbar\,(\vec{k} - k_m)$, (2.12) turns into

$$E_{n,kin} = \frac{1}{2\hbar^2}\frac{d^2E}{dk^2}\vec{p}_n^{\,2} = \frac{\vec{p}_n^{\,2}}{2m_n} \quad (2.13)$$

where we have defined, furthermore,

$$m_n \equiv \hbar^2 / \frac{d^2E}{dk^2} \quad (2.14)$$

m_n has the dimension of a mass and is called the effective mass of the electrons. It has the order of magnitude of electron mass in vacuum, but is not equal to it. The velocity of the electron is $\vec{v}_n = \vec{p}_n / m_n$. Because the momentum vector $\vec{p}_m$ of the conduction band minimum in $\vec{p}_n = \vec{p} - \vec{p}_m$ is constant, the relation between force and acceleration can be written also as $\vec{F} = d\vec{p}_n / dt$. Hence, quantum mechanics leads to the result that conduction electrons in the lattice-periodic potential obey relationships of mechanics; they react on external forces like mass points with a positive effective mass m_n and a negative charge −q. The interaction with the internal field of the periodic potential is taken into account by the effective mass. The effective mass is not a scalar but a tensor, but this is not important for an essential understanding in fact of the model.

Holes behave quite analogously. This is obtained representing a hole by the full valence band minus an electron. The full band does not conduct current because the parts from + $\vec{k}$ and $-\vec{k}$ compensate each other. The contribution of an electron to the current density is $-q\vec{v}_n$, and that of the missing electron, the hole, $+q\vec{v}_n$. Since the velocity of the hole is identical to that of the missing electron, the equation of motion yields $\dot{\vec{v}} = \vec{F}_n / m_{n,v} = \dot{\vec{v}}_p = \vec{F}_p / m_p$. Here $\vec{F}_n$ denotes the force on the electron, $\vec{F}_p$ the force on the hole, which due to the opposite charge has the opposite sign for electric and magnetic fields, $\vec{F}_p = -\vec{F}_n$. $m_{n,v}$ is the effective mass of a valence electron. Hence we have to define $m_p = -m_{n,v}$ to obtain the classical relation between force and acceleration. Because at the top of the valence band $d^2E/dk^2 < 0$, the effective mass of the valence electron is negative, and that of the hole positive:

$$m_p = -\hbar^2 / \frac{dE}{dk}(E = E_V) \quad (2.15)$$

For $E_{kin,p} \equiv E_V - E$ one obtains similarly as for the electrons: $E_{kin,p} = \vec{p}^2/2m_p = m_p\vec{v}_p^2/2$. Hence we have obtained that also holes obey the relationships of mechanics, they behave like mass points with a positive charge q and positive effective mass. Quantum mechanics leads to a particle picture, called quasi-particle model, which is justified on the same basis for holes as for conduction electrons. It has a wide range of applicability and is used throughout device physics.

Quantitatively, the situation is more complicated because the effective mass of electrons in Si and other semiconductors is a tensor and depends strongly on orientation. However, the effective mass entering the mobility and conductivity, the "conductivity effective mass" $m_{n,c}$, is an average over the equivalent minima in k-space, and this averaging yields a scalar in cubic crystals [Smi59]. The conductivity effective mass in Si is $m_{n,c}$ = 0.27 m_0 (m_0 free electron mass) [Gre90]. Also for holes in Si a scalar value is used for the conductivity effective mass, it amounts to $m_{p,c}$ = 0.4 m_0 at 300 K. Both masses depend only very weakly on temperature.

Regarding the quasi-particle model for holes, the question is whether it implicates essentially different results compared with the previous classical picture of a void or bubble in the sea of bonding electrons. The answer is yes. Most evidently this is shown by the Hall-effect, measurements of which are the main experimental tool to investigate the basic processes underlying conductivity. In these experiments, a magnetic field $\vec{B}$ is applied perpendicular to a semiconductor strip in which a longitudinal current is flowing, the lateral voltage V_H at the strip is measured, see Fig. 2.5.

Since the carriers are forced to flow along the strip, the Lorentz force $Q\,\vec{v} \times \vec{B}$ is balanced by the formation of a lateral electric field $\vec{E}_H = -\vec{v} \times \vec{B}$ where $\vec{v}$ is the vector of (drift) velocity of the carriers and Q their charge. If x and z are the coordinates in the directions of $\vec{v}$ and $\vec{B}$, respectively, then the "Hall field" $\vec{E}_H$ has the direction of the positive y coordinate. With the scalar components in this coordinate system denoted by normal letters (so that $\vec{v}$ = (v, 0, 0), $\vec{B}$ = (0, 0, B), $\vec{E}_H$ = (0, E_H, 0)), one obtains

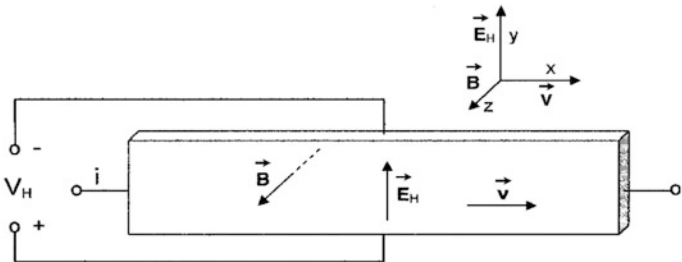


Fig. 2.5 Basic arrangement for Hall measurements

$$E_H = v \cdot B = \frac{j}{Q \cdot C} B = R_H \cdot j \cdot B \quad \text{with} \quad R_H = \frac{1}{Q \cdot C} \tag{2.16}$$

Here the velocity was substituted by the current density according to $j = Q \cdot C \cdot v$ where C denotes the carrier concentration. R_H is called “Hall constant”. If in a p-type semiconductor the hole current is in principle achieved by a motion of valence electrons as in the classical bubble model, the Hall constant would have the same negative sign as for an n-type specimen since then also $Q = -q$. Furthermore, because the concentration of valence electrons, C, is much higher than that of the empty states and also much higher than the concentration of electrons in an n-type semiconductor of equal conductivity, the absolute value of R_H should be very small after the classical bubble model. Actually, however, the Hall constant measured for specimens doped with acceptors is positive in contrast to that of n-type samples and in magnitude it is comparable for both types of doping. Hence the holes manifest themselves as independent entities which in a magnetic field experience a Lorentz force like positively charged mass points. Hence p-type conductivity is a conductivity type of its own, equivalent to n-type conduction. Before the quantum theory of solids was developed, the positive Hall constant found in many specimens was a very irritating phenomenon.

2.5 The Doped Semiconductor

If in a silicon crystal some atoms in the lattice are replaced by atoms of an element of group V in the periodic table, e.g. phosphorus, each impurity atom has one electron more in the outer shell than necessary for the four covalent bonds. Therefore, one electron is only loosely bonded and needs only a small amount of energy—available probably as thermal energy—to be removed from the impurity atom and freed for conduction. These elements donating their fifth electron to the conduction band are called donors. The result is an extrinsic, n-type semiconductor with “n” pointing to the negative charge of the carriers. On the other hand, if the silicon atoms are replaced at some lattice points by atoms of an element of group III, e.g. boron, each impurity atom has one electron less than necessary for the four covalent bonds. Since the bonds between the impurity and the neighboring silicon atoms are nearly as tight as between the silicon atoms themselves, there is only little energy necessary for moving an electron out of a Si–Si bond in the neighborhood to the impurity to complete its bonds with the four silicon neighbors. Accepting an electron of the valence band and thus generating a mobile hole, these impurities leading to p-type conductivity are called acceptors. In the energy band picture, the donors have energy levels close to the conduction band in the bandgap, the acceptors close to the valence band edge. The energy levels of important dopants in Si and 4H-SiC are shown in Fig. 2.6.

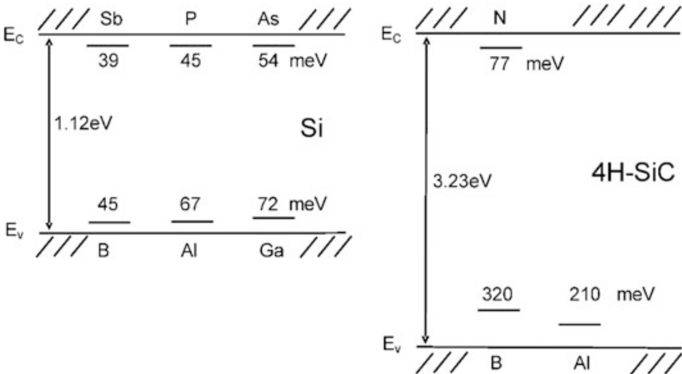


Fig. 2.6 Energy levels of doping impurities in Si and 4H-SiC. The differences from the respective band are given, i.e. the ionization or activation energies for elevating an electron or hole from the neutral impurity atom into the conduction or valence band, respectively

In a quantitative model, impurities of group 3 and 5 are considered as hydrogen-like systems, the ground state energy of which represents the ionization energies $\Delta E_D = E_C - E_D$, $\Delta E_A = E_A - E_V$ [Smi59, Koh57]: In a neutral donor the electron not used for the lattice bonds is bound to the ion D^+ by the electrical field which is idealized to the Coulombic form $\mathsf{E} = q/4\pi\varepsilon r^2$ as in a homogeneous medium. Using for it the macroscopic dielectric constant $\varepsilon = \varepsilon_r \cdot \varepsilon_0$ of the semiconductor and for the electron its effective mass m_n in the host crystal, the formulas of the hydrogen atom result, modified only by the numeric values of ε and m_n. Since the orbit radii obtained in this way are significantly larger than the distance between the semiconductor atoms, the assumptions are justified as a rough approximation. As ionization energy one obtains

$$\Delta E = \frac{m_n q^4}{8\varepsilon^2 h^2}$$

where h is Planck's constant. For donors in Si ($\varepsilon = 11.7 \cdot \varepsilon_0$, $m_{eff} = 0.27 \cdot m_0$, m_0 free-electron mass) this yields: $\Delta E_D = 26.8\text{meV}$. Similarly, a neutral acceptor is understood as the ion A^- (bound in the lattice by complete covalent bonds) surrounded by a hole in the idealized Coulomb field. With the hole effective mass of 0.4 m_0, the ionization energy of acceptors in silicon is obtained as $\Delta E_A = 40.0\,\text{meV}$. As is seen from Fig. 2.6, the experimental ionization energies are of the order of these model values, but somewhat higher. In SiC, the differences are larger. Apart from neglected effects of the immediate neighborhood of the impurity, the electron or hole assigned in the model to the orbit of the ion may be involved to some extent in the bonding of the ion in the lattice. Deep levels in the bandgap as shown by recombination centers (see Sect. 2.7.2) are beyond the range of this model.

Because of the small ionization energies and the many states in the conduction and valence bands available for electrons and holes, most donors and acceptors in Si

are ionized at room temperature. This follows from (2.1) or (2.2) if the Fermi level lies below the level of the donor or above the acceptor level in either case of doping. Considering the degree of ionization in detail we use the notation in the case of donor doping denoting the total impurity concentration with N_D and the concentrations of neutral and ionized impurities with N_D^0, N_D^+, respectively. Then the number of occupied donor levels in (2.2) is $n_E = N_D^0$ and the number of unoccupied levels $N_E - n_E = N_D - N_D^0 = N_D^+$, hence the ratio of both, the occupation degree, is

$$\frac{N_D^0}{N_D^+} = g \cdot e^{-\frac{E_D - E_F}{kT}} \tag{2.17}$$

where we have added a "degeneracy factor" g. This factor, which for donors is $g = 2$, is necessary because the neutral donors D^0 exist in two states depending on the spin orientation of the trapped electron. Nevertheless only one electron can be bound because for a second the Coulomb field is no longer present. Since the state D^- does not appear, the Fermi energy does not include the degeneracy of the D^0 state and hence one has to take into account it separately [Spe58, Sho59]. The Fermi energy can be eliminated dividing Eq. (2.17) by (2.4), which yields:

$$\frac{N_D^0/N_D^+}{n/N_C} = g \cdot e^{\frac{E_c - E_D}{kT}} \tag{2.18}$$

As long as the intrinsic concentration and hence the hole density $p = n_i^2/n$ is small, the neutrality condition is $N_D^+ = n$ so that $N_D^0 = N_D - n$. Inserting this and solving for n, one obtains after simple conversion

$$n = \frac{N_D}{\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{N_D}{N_C} \cdot g \cdot e^{\frac{\Delta E_D}{kT}}}} \tag{2.19}$$

For small ΔE_D and $N_D << N_C$, this yields $n \approx N_D$ as expected. For acceptor doping the hole concentration is obtained exchanging N_D, N_C and ΔE_D by the respective acceptor quantities N_A, N_V and ΔE_A (see the more general Eq. (2.23) at $N_D = 0$). However, the degeneracy factor of acceptor levels in Si, Ge and SiC is $g = 4$ since there are two degenerate valence bands at $k = 0$, from which the levels are split off, and this results in a further degeneracy in addition to spin degeneracy [Bla62] [1]. For both n- and p-type conductivity, the degeneracy factor reduces the ionization ratio n/N_D or p/N_A, respectively.

[1] Actually the occupation of excited states causes an additional degeneracy and enhancement of g [Bla62, p. 140 ff]. In Si this effect seems to be relative small owing to the high energy difference between the excited states and the ground state. Since a calculation under real conditions is not available, the effect is not taken into account.

As is confirmed by inserting the ionization energies of Fig. 2.6 and the effective density of states from (2.8), dopants in Si are nearly completely ionized at room temperature up to doping concentrations of 10^{17} cm^{-3}. Considerable de-ionization is obtained at higher concentrations. In this doping range one has to take into account, however, that the ionization energy decreases with increasing concentration starting from the values given in Fig. 2.6 for small densities. For a Ga impurity concentration of 5×10^{17} cm^{-3}, which is about the maximum doping of the p-base in a thyristor, an activation energy of 65 meV was determined [Wfs60]. With this ΔE_A an ionization ratio of 66% of the total Ga concentration is obtained at 300 K. As function of temperature, the carrier concentration is shown in Fig. 2.7 for this example and additionally for phosphorus with a concentration of 10^{14} cm^{-3}. Whereas for the concentration 5×10^{17} cm^{-3} the deionization is noticeable up to 400 K, for a doping concentration of 10^{14} cm^{-3} the ionization remains complete down to a temperature of 80 K. Above $T = 430$ K the intrinsic concentration becomes comparable with $N_D = 10^{14}$ cm^{-3} and causes an increase of the carrier concentration.

To include this effect, the neutrality condition inserted in (2.18) has to take into account the minority carrier concentration. Writing $n = N_D^+ + n_i^2/n$ where in this range $N_D^+ = N_D$, one obtains

$$n = \frac{N_D}{2} + \sqrt{\left(\frac{N_D}{2}\right)^2 + n_i(T)^2} \quad (2.20)$$

In SiC the energy levels of dopants lie deeper in the bandgap (see Fig. 2.6), and hence a larger part of the doping atoms remains neutral, particularly for acceptors. For Al, the preferred acceptor dopant in SiC devices, the acceptor version of Eq. (2.19) with $N_V = 2.5 \times 10^{19}$ cm^{-3} [Gol01] yields for a doping concentration $N_A = 1 \times 10^{16}$ cm^{-3} an ionization ratio of only 35% at 300 K. This can strongly influence the device characteristics. Boron in SiC shows even a smaller degree of

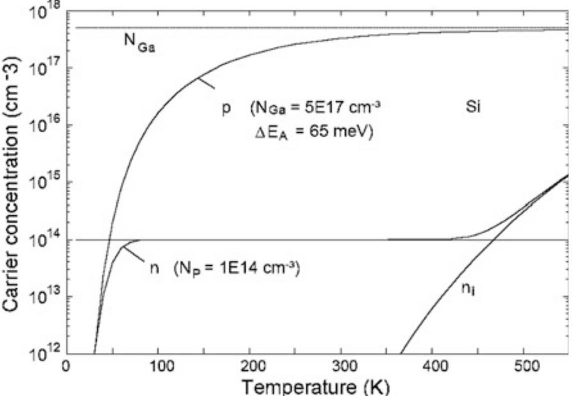


Fig. 2.7 Carrier concentration as a function of temperature for a Ga doping of 5×10^{17} cm^{-3} and a phosphorus doping of 10^{14} cm^{-3}

ionization. Incomplete ionization is an obstacle for fabrication of layers with good emitter efficiency with SiC.

In GaN, silicon doping is the typical choice for intended n-doping. The activation energy of silicon in GaN is 5–9 meV, this allows effective doping [Qua08]. Mg is used as acceptor, its ionization energy is in the range 140–210 meV [Lev01], in [Qua08] a value of 173 meV is reported. Also native defects act as dopants in GaN, a vacancy on a nitrogen place V_N acts as donor, a vacancy on a Ga-site V_{Ga} is a shallow acceptor.

In diamond, the energy levels are even deeper in the bandgap [Gil79]. Boron is a deep acceptor level with activation energy of 370 meV [Gil79]. Phosphorus exhibits donor states with activation energies 0.8–1.16 eV [Oka90]. Therefore, the ionization ratio at room temperature will be low, and it is difficult to reach a significant carrier density for current transport.

Often it is important to know the Fermi level for a given doping density. For complete ionization and temperatures where $n_i << N_D$, N_A, the Fermi-level is related in a simple manner to the doping concentration. If we restrict again to the nondegenerate case, one obtains inserting $n = N_D$ in (2.4):

$$E_C - E_F = k \cdot T \cdot \ln\left(\frac{N_C}{N_D}\right) \tag{2.21}$$

For silicon this and the analogous equation for acceptor doping are plotted in Fig. 2.8 for 300 and 400 K. The Fermi energy is a linear function of the logarithm of the doping density.

According to Eq. (2.19) the ratio n/N_D decreases with increasing doping density N_D and, assuming ΔE_D as constant, becomes small at high N_D. Really this is not observed, on the contrary at doping concentrations of 10^{19} cm^{-3} and higher the impurities are completely ionized. This is explained by the mentioned decrease of the ionization energy and other high doping effects. The decrease of ΔE_D, ΔE_A is caused by screening of the impurity charge by free carriers, e.g. D^+ by electrons, as well as by the formation of a tail of states at the neighbored band, because the

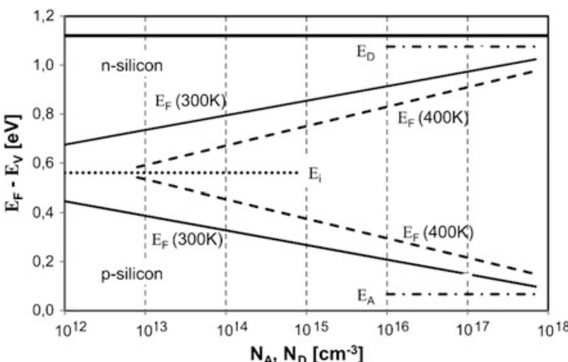


Fig. 2.8 Fermi level in n- and p-type Si versus doping concentration at T = 300 K

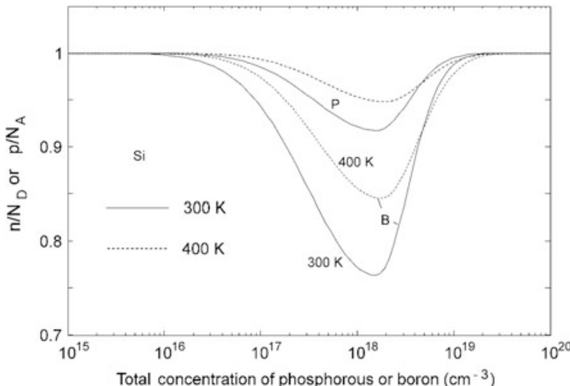


Fig. 2.9 Ionization ratio of phosphorus and boron in Si at 300 and 400 K as a function of the total concentration of the respective impurity. After Kuzmicz [Kuz86]

periodic lattice potential is disturbed by the Coulomb field of the statistically distributed impurities. Additionally, the levels of the impurity atoms spread and form an impurity band because the wave functions of the bound impurity states overlap at high concentrations and hence the levels split up. These effects begin more than an order of magnitude below the concentration $N_{C,V}/2$, where degeneracy sets in. Since they are not yet completely mastered theoretically from first principles, practical approaches use semi-theoretical or empirical modelling concepts with adjusted parameters. A calculation of Kuzmicz [Kuz86], who considers the mentioned impurity band and the decrease of the mean ionization energy, yielded the dependence of the ionization ratio on impurity concentration as shown in Fig. 2.9.

The figure refers to phosphorus and boron in Si, the most commonly used dopants. The curves were recalculated using analytical expressions developed in [Kuz86] on the base of numerical calculations. The deionization has its maximum near 2×10^{18} cm^{-3} and amounts to 9% for phosphorus and 23% for boron at 300 K. Above 10^{19} cm^{-3}, the ionization is again complete. The deionization in the intermediate range becomes noticeable in the temperature dependence of the resistivity.

What if both donors and acceptors are present simultaneously? This is always the case in devices doped by impurity diffusion and particularly in the vicinity of diffused pn-junctions, which are defined by $N_A = N_D$. To see the influence of this compensation we consider the case that the acceptor doping predominates and the difference $N_A - N_D$ is large against the intrinsic concentration n_i. Due to the deep lying Fermi level, all the donors are then ionized, $N_D^+ = N_D$. The hole concentration in a neutral region is hence given by $p = N_A^- - N_D$. Usually it is assumed that also the acceptors are completely ionized, so the hole density is equal to the net impurity concentration, $p = N_A - N_D$. Taking the deionization of the acceptors into account, one obtains for the ionization ratio of the *net* acceptor concentration $N_{A,net} = N_A - N_D$ with $r \equiv N_A^- / N_A$:

$$\frac{p}{N_{A,net}} = \frac{rN_A - N_D}{N_{A,net}} = r - (1-r)\frac{N_D}{N_{A,net}}$$

Related to the reduced net acceptor doping the deionization of the acceptors becomes a higher weight, hence $p/N_{A,net}$ decreases with compensating doping N_D. But also the acceptor ionization ratio r itself depends on the concentration N_D. Expressing N_A^-/N_A^0 by Eq. (2.2) and multiplying with (2.5) one obtains

$$\frac{N_A^-}{N_A^0}p = \frac{N_V}{g}\exp\left(-\frac{E_A - E_V}{kT}\right) \equiv C \tag{2.22}$$

Again a degeneracy factor g is included to allow for the enhanced occupation probability of the neutral impurity state. The concentration C is the mass law constant of the reaction $N_A^- + p \Leftrightarrow N_A^0$. Since now p is reduced by the compensating dopant, the occupation degree $\gamma \equiv N_A^-/N_A^0$ is enhanced and thus also the ionization ratio $r = \gamma/(1+\gamma)$. Inserting the neutrality condition $N_A^- = p + N_D$ and the equation $N_A^0 = N_A - N_A^- = N_A - p - N_D$ into (2.22) one obtains for the hole density

$$\frac{(p+N_D)p}{N_{A,net} - p} = C \tag{2.23}$$

This has the solution

$$p = \frac{2N_{A,net}}{1 + \frac{N_D}{C} + \sqrt{\left(1+\frac{N_D}{C}\right)^2 + 4\frac{N_{A,net}}{C}}} \tag{2.24}$$

Equation (2.24) is a generalization of the acceptor version of (2.19), into which it turns with $N_D = 0$. The equation shows explicitly how the fraction of the ionized net acceptor doping depends, apart from $N_{A,net} = N_A - N_D$ itself, on the compensating dopant concentration N_D. Because C is of the order the magnitude 10^{18} to 10^{19} cm^{-3}, the effect is only significant for $N_D \gtrsim 5 \times 10^{16}$cm^{-3}, hence for a realistic description one has to consider high-doping effects. For not too high concentrations these can be expressed approximately by an effective decrease of the ionization energy $\Delta E_A = E_A - E_V$ with doping densities. Since this leads to an increase of the concentration C according to (2.22), the decrease of p with N_D as given by (2.24) is strongly weakened. The ionization energy of boron was reported to depend on doping densities as $\Delta E_A = 46 - 3 \times 10^{-5} \cdot (N_A^+ + N_D)^{1/3}$ meV [Li78]. In Fig. 2.10 the dependence of $p = N_{A,net}^-$ on N_D obtained with this ΔE_A is plotted for two temperatures. The net acceptor density $N_A - N_D$ is taken to be constant and equal to 1×10^{16} cm^{-3}. At 300 K the ionization ratio decreases to 80% up to $N_D = 2 \times 10^{18}$ cm^{-3}. Hence the deionization in regions of strong compensation is considerable already at low net doping densities. For absent compensation, $N_D = 0$, the acceptor ionization ratio obtained from (2.24) with the stated $\Delta E_A(N_A)$ is up to

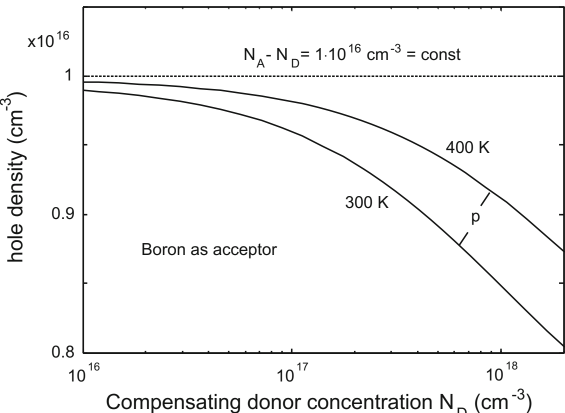


Fig. 2.10 Hole concentration for a constant net acceptor doping $N_A - N_D$ versus compensating donor density N_D

nearly $N_A = 2 \times 10^{18}$ cm^{-3} in rough accordance with the curves for boron in Fig. 2.9, beyond this doping the approach fails. In literature [Bla62, p. 132 ff] doping with more than one impurity species is treated extensively using equations like (2.24).

Till now we have assumed that the semiconductor is neutral and in thermal equilibrium. In devices, however, there are regions with space charge where carriers are depleted and, on the other hand, regions with injected carriers adding to those supplied by dopants. Departing from above calculations, the ionization in space charge regions is nearly always complete because the depletion of carriers ($p << N_A^-$) shifts the reaction $A \Leftrightarrow A^- + \oplus$ to higher N_A^-. Under conditions of high injection, however, the deionization is enhanced, and especially in this case it can affect device operation. These effects can be calculated quantitatively from Eqs. (2.18) and (2.22).

In spite of the limits shown by this discussion, the doping atoms in silicon are usually to a fairly high degree ionized at operating temperatures $T \gtrsim 250\,K$. In fact *most analytical calculations and programs for device simulation are just based on the assumption of complete ionization throughout the device and during the whole switching time considered.* The above discussion shows, where a more exact treatment may be worth the enhanced effort. For devices in SiC and GaN this is often the case.

The high doping effects mentioned are responsible for another phenomenon which influences the characteristics of power devices: the reduction of bandgap and the associated increase of the intrinsic concentration n_i at high doping densities. From measurements on the p base of bipolar transistors Slotboom and De Graaff [Slo76] obtained the following empirical relationship for the bandgap narrowing, ΔE_g, in dependence of doping concentration N:

$$\begin{aligned} \Delta E_g &= 9 \times 10^{-3}\text{eV} \cdot \left(\ln \frac{N}{10^{17}\text{cm}^{-3}} + \sqrt{\left(\ln \frac{N}{10^{17}\text{cm}^{-3}}\right)^2 + 0.5} \right) \\ &\approx 18\text{meV} \cdot \ln \frac{N}{10^{17}\text{cm}^{-3}} \quad \text{for N} > 5 \times 10^{17}\text{cm}^{-3} \end{aligned} \tag{2.25}$$

Calculations considering band tails and impurity band were found to be in rough accordance with (2.25) [Slo77]. The dependency observed for regions with acceptor doping is used also for highly doped n regions. Another effect which has been shown by Lanyon and Tuft [Lan79] to result in an effective reduction of bandgap is the electrostatic field energy required when an electron-hole pair is created within the surrounding of majority carriers. This stored energy included in the band gap becomes smaller with increasing (majority) carrier density due to screening. Based on this mechanism, the following theoretical formula has been derived in the range where Maxwell-Boltzmann statistics apply [Lan79]

$$\Delta E_g = \frac{3q^2}{16\pi\varepsilon} \left(\frac{q^2 n}{\varepsilon kT} \right)^{1/2} = 22.6 \left(\frac{n}{10^{18}} \frac{300}{T} \right)^{1/2} \text{meV} \tag{2.26}$$

ΔE_g depends here on the carrier concentration n (in case of n doping) rather than the doping concentration, which can be of significance especially in space charge regions. The numerical expression on the right hand side of (2.26) is obtained for Si using the dielectric constant $\varepsilon = 11.7 \cdot \varepsilon_0$. Although band tails and impurity band are not considered, (2.26) is also claimed to describe largely the measurements which were partly reinterpreted [Lan79]. Differing from the results of Slotboom and De Graaff, (2.25), the bandgap reduction according to (2.26) depends on temperature.

A plot of the two relationships is shown in Fig. 2.11. Around 10^{18} cm^{-3}, the equation of Lanyon and Tuft yields a considerably smaller bandgap narrowing than that of Slotboom and De Graaff, above 2×10^{19} cm^{-3} it is the other way round. Both evaluations are not free of assumptions; the experimental result (2.25) is supported theoretically in [Slo77].

The primarily determined quantity in [Slo76] is the intrinsic carrier concentration n_i in dependence on doping concentration. For conversion to bandgap narrowing the effective density of states N_V or N_C of the relevant band is assumed to be unchanged. Hence as obtained by inserting the reduced bandgap $E_g = E_{g0} - \Delta E_g$ into (2.6), the following relation is used for conversion between n_i and ΔE_g:

$$n \cdot p = n_i^2 = N_C N_V e^{\frac{-E_g(N)}{kT}} = n_{i0}^2 e^{\frac{\Delta E_g}{kT}}, \tag{2.27}$$

where n_{i0} denotes the intrinsic concentration at low doping (or carrier) concentrations corresponding to the bandgap E_{g0}. We will use these results in later Sects. 3.4 and 5.4.5 for the injection efficiency, i.e. the influence of recombination in highly doped emitter regions on device characteristics.

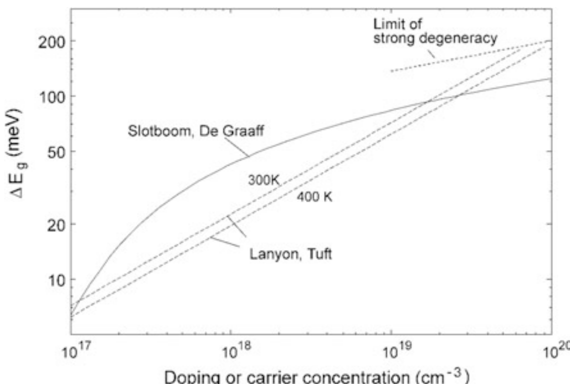


Fig. 2.11 Bandgap narrowing as a function of doping or carrier density according to [Slo76], Eq. (2.25) (solid line), and [Lan79], Eq. (2.26) (dashed lines), respectively. The dotted lines represents a limiting expression for strong degeneracy [Lan79]

At the end of this paragraph we mention some frequently used notations. The terms *majority carriers* and *minority carriers* are used to identify the carriers of the type supplied by the (dominating) doping atoms and the carriers of the second kind, which are in minority and related with the first by the mass law equation $n \cdot p = n_i^2$ in thermal equilibrium. In this case, the density of minority carriers is in silicon many orders of magnitude smaller than the majority concentration. For a donor density $N_D = 1 \times 10^{14}$ cm^{-3} one obtains at room temperature with $n_i \approx 10^{10}$ cm^{-3} a majority carrier concentration $n = 1 \times 10^{14}$ cm^{-3} and a minority concentration $p = n_i^2/n = 1 \times 10^6$ cm^{-3}. At higher temperatures, the minority carrier concentration can be calculated inserting (2.20) into $p = n_i^2/n$, see also Fig. 2.7. Whereas in thermal equilibrium the minority carriers are insignificant for the properties of bulk material, they play an essential part for the functioning of pn-devices based on injection and extraction of minority carriers. This will be investigated later.

In devices with layers of different doping levels, a superscript index '−' or '+' is added to the symbols for the conductivity type to indicate the doping level. So n^-, p^- signify n and p layers with a much lower doping than neighboring regions; likewise n^+, p^+ mean layers with a much higher doping level. Typically the doping ranges indicated in this manner are as follows:

$$
\begin{array}{ll}
\mathrm{n^-, p^-} & 10^{12}-10^{14}\mathrm{cm}^{-3} \\
\mathrm{n, p} & 10^{15}-10^{18}\mathrm{cm}^{-3} \\
\mathrm{n^+, p^+} & 10^{19}-10^{21}\mathrm{cm}^{-3}
\end{array}
$$

Normally a power device contains in the interior a weakly doped n^- region which essentially determines the characteristics. Heavily doped n^+ and p^+ layers border on the metalized surfaces.

2.6 Current Transport

2.6.1 Carrier Mobilities and Field Currents

As discussed in Sects. 2.3 and 2.4, electrons and holes in the semiconductor behave essentially like free particles which, however, are scattered by vibrating lattice atoms, impurity ions and other scattering centers. Like molecules in a gas they have a kinetic thermal energy which on statistical average is

$$E_{kin} = \frac{m}{2} v_{th}^2 = \frac{3}{2} kT \tag{2.28}$$

where v_{th} is the mean thermal velocity and m the effective mass of the respective carrier. Already at room temperature the mean thermal velocity is very high: With the electron effective mass $m = m_n = 0.27 \cdot m_0$, one obtains from (2.28) $v_{th} = \sqrt{(3kT/m_n)} = 2.2 \times 10^7$ cm/s = 220 µm/ns at 300 K. The mean free path between two scattering events is of the order 10 nm in lightly doped Si and still smaller if many impurities are present. Since the thermal motion is statistically distributed over all directions, the current resulting from the large number of carriers is zero between terminals on equal potential. The thermal velocity used in (2.28) is defined as the root of the mean value of the squared velocity, a rough measure for the mean *absolute* velocity. Contrary to it, the normal *linear* mean value of the thermal velocities taking account also of their directions is zero for both electrons and holes, if no field is present.

When an electric field E is applied, each carrier experiences a force $\pm q \cdot \mathsf{E}$ and is accelerated between two collisions. Hence the thermal velocity is superimposed by an additional velocity which for holes has the direction of the electric field and for electrons the opposite. Averaging linearly over time and the carriers of each type, now *non-zero* mean velocities v_n, v_p of the electrons and holes result. These velocities caused by the field are called *drift* velocities. For low fields, defined by the condition that the drift velocities are small compared with the thermal velocity v_{th}, the mean free time τ_c between two collisions which depend on the total velocity is independent of the field. Therefore, in this range, the drift velocities are proportional to the field strength

$$v_{n,p} = \mp \mu_{n,p} \cdot \mathsf{E} \tag{2.29}$$

The proportionality factors μ_n for electrons and μ_p for holes are called *mobilities.* Due to the minus sign used in the case of electrons both mobilities are positive constants. Inserting (2.29) into the condition $v << v_{th}$, and omitting now the indices n or p, one obtains as condition which the field must satisfy for constant mobilities

$$\mathsf{E} << \frac{v_{th}}{\mu} \tag{2.30}$$

The significance of the mobilities follows from their connection with the macroscopic electric current densities. These are obtained from (2.29) as

$$\begin{aligned} j_n &= -q \cdot n \cdot v_n = q \cdot \mu_n \cdot n \cdot \mathsf{E} \\ j_p &= q \cdot p \cdot v_p = q \cdot \mu_p \cdot p \cdot \mathsf{E} \end{aligned} \tag{2.31}$$

with the concentrations n and p of electrons and holes. The sum of both is the total current density j:

$$\begin{aligned} j &= j_n + j_p = q \cdot (\mu_n \cdot n + \mu_p \cdot p) \cdot \mathsf{E} \\ &= \sigma \cdot \mathsf{E} = \mathsf{E}/\rho \end{aligned} \tag{2.32}$$

where

$$\sigma \equiv q \cdot (n \cdot \mu_n + p \cdot \mu_p) = 1/\rho \tag{2.33}$$

is the electrical conductivity and ρ the resistivity. According to these equations, the mobilities are material parameters which determine the ohmic voltage drop $V = \mathsf{E} \cdot \Delta x = \rho \cdot j \cdot \Delta x$ for a given current density and hence the power loss density $V \cdot j$ and heat generation. Hence, the mobilities determine the maximum allowed current density of devices. Together with other characteristics they decide on the suitability of a semiconductor for power devices.

An overview of mobilities in semiconductors is given in Table 2.2. It shows at first that the mobility of holes is in all cases significantly smaller than that of electrons. Hence, n and p regions in devices are not equivalent. Especially in unipolar devices which conduct current only by majority carriers, the weakly doped region required for the blocking voltage is chosen preferentially of n type doping. As mentioned in Sect. 2.1, the very high electron mobility of GaAs offers the possibility to make Schottky diodes with a low on-state voltage drop and simultaneously a relative large thickness necessary for a high blocking voltage. For the

Table 2.2 Mobilities of various semiconductors at room temperature for light doping. For 4H-SiC the mobilities parallel to the hexagonal axis are given, for GaN (2H-type) the mobilities perpendicular to the hexagonal axis in bulk material and additionally the electron mobility in the mentioned 2-dimensional electron gas (2DEG). The table contains also the saturation drift velocity of electrons, a property discussed at the end of this chapter

	μ_n [cm²/(Vs)]	μ_p [cm²/(Vs)]	$v_{sat(n)}$ [cm/s]
Ge	3900	1900	6×10^6
Si	1420	470	1.05×10^7
GaAs	8000	400	1×10^7
4H-SiC	1000	115	2×10^7
GaN	990	150	2.5×10^7
GaN 2DEG	up to 2000		
Diamond	2200	1800	2.7×10^7

hexagonal semiconductors SiC and GaN the mobilities are anisotropic, i.e. they are different in directions parallel ($\mu_{||}$) and perpendicular ($\mu\perp$) to the hexagonal axis. The values in the table are the mobilities parallel to the hexagonal axis, which in vertical devices is the direction of main current flow. In 4H-Si the anisotropy is small, $\mu_{n\|} \approx 1.2\ \mu_{n\perp}$ [Scr94] in contrast to the polytype 6H-SiC which electron mobility $\mu_{n||}$ is about a factor 5 smaller than $\mu_{n\perp}$ and also than $\mu_{n\|}$ in 4H-SiC. This is the main reason why 4H-SiC is preferred now against 6H-SiC, which polytype attracted much research efforts in the 1990s. To judge different semiconductors regarding conduction losses in devices, one has to use the mobilities together with other relevant properties such as the necessary thickness and doping density allowed for a given blocking voltage. Although the mobilities of 4H-SiC are lower than those of Si, in combination with the much smaller thickness and higher allowed doping of the base region a much smaller on-state resistance of unipolar devices can be achieved. This holds also in comparison with GaAs. GaN is with respect to the mobilities comparable to SiC, however, in a two-dimensional electron gas (2DEG) at a heterojunction to AlGaN electron mobilites up to 2000 $cm^2V^{-1}s^{-1}$ are measured, see Chap. 4. High mobilities are measured in diamond, but because of large ionization energies of the dopants the carrier concentration is small.

The mobilities are constant with regard to the field (if small enough), but depend on doping concentration and temperature. An accurate knowledge of their dependence on doping concentration is very important particularly because the mobility determines the relation between resistivity and doping density which is used daily to conclude from the simply measurable resistivity to doping density N. Assuming complete ionization one has for an n type wafer

$$\rho(N) = \frac{1}{q \cdot \mu_n \cdot n} = \frac{1}{q \cdot \mu_n(N)} \cdot \frac{1}{N} \tag{2.34}$$

In an undoped and weakly doped semiconductor, the mobilities are determined by scattering on phonons, i.e. vibrating lattice atoms. Above a doping concentration of 10^{15} cm^{-3}, the mobilities are noticeably and at higher concentrations strongly reduced by collisions with doping ions. At still higher concentrations the scattering by ions is limited by the carriers themselves by screening of the impurity charge. The experimental dependence of mobilities in Si on the donor or acceptor ion concentration, respectively (= majority carrier concentration) is depicted in Fig. 2.12 which also shows fitting curves to the measurements. The figure is valid at room temperature. The experimental curves represent measurements of Thurber and coworkers on phosphorus doped silicon for μ_n [Thu80a] and boron doped silicon for μ_p [Thu80b] using their analytical representation of measurements. For $n > 8 \times 10^{18}$ cm^{-3} an empirical dependence for the electron mobility of Masetti et al. [Mas83] is shown, who determined the mobilities at very high doping levels. The dashed lines are fits to the experiments by the often used formula of Caughey and Thomas [Cau67]:

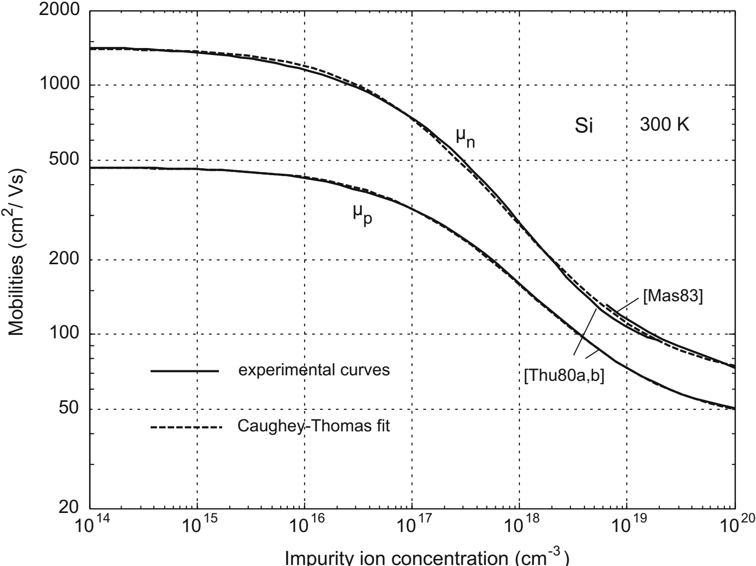


Fig. 2.12 Mobilities in Si as function of impurity ion density

$$\mu = \mu_\infty + \frac{\mu_0 - \mu_\infty}{1 + \left(N/N_{ref}\right)^\gamma} \tag{2.35}$$

The limiting values μ_0 and μ_∞ for low and high concentrations and the concentration N_{ref} at which the mobility adopts the mean value between them, are matched to the experiments. N denotes the impurity ion concentration. Parameter values used for the fit are given in Appendix A, where also their variation with temperature is given. As is seen, this simple approach is well suited to describe the experimental dependence up to impurity densities of 1×10^{20} cm^{-3}. Around 2×10^{18} cm^{-3} the carrier density deviates noticeably from the total impurity density as discussed in the previous paragraph. To obtain the mobility for a given total impurity concentration in this range, first the ionized impurity density has to be determined using Fig. 2.8. Although Eq. (2.35) can be used also with N defined as total impurity density, the plot versus ion density is preferred, because the ionized impurity concentration is the quantity determining the mobilities, while scattering by impurities in the neutral state may be neglected. Hence the plot is applicable with good accuracy also for doping with As, Ga and Al, which are ionized to a different extent. Often complete ionization is assumed when discussing mobilities in the range of operation temperatures of devices. We will also do this in what follows.

Also for silicon carbide the Caughey-Thomas formula (2.35) can be used to describe the dependency of the mobilities on the doping density. The parameters for 4H-SiC are given in Appendix A as well.

So far we have dealt with the mobilities of majority carriers in thermal equilibrium. During on-state of power devices high concentrations of electrons and holes are injected into the weakly doped base region. Hence electrons are scattered

by holes and holes by electrons (electron-hole scattering), and this leads to a similar decrease of the mobilities as scattering by ions. The voltage drop over the base region is strongly dependent on this effect. The mobility sum $\mu_n + \mu_p$ at high injection levels where the electron and hole concentrations are equal to each other ($n \approx p \gg N_{dop}$), has been measured as function of the carrier concentration by Dannhäuser and Krause [Dan72, Kra72]. Figure 2.13 shows their experimental dependency. Theoretically, electron-hole scattering is often described by a classical formula of Fletcher [Fle57] using for combination with the lattice scattering mobility an expression of Debye and Conwell [Deb54]. The mobility sum as function of n obtained from this theory at 300 K is shown in Fig. 2.13 also. Considering that the effective masses of electrons and holes entering the formulae were not matched, the agreement with the measurements is good. A more general mobility model using quantum mechanical scattering theory has been developed by Klaassen [Kla92]. The result of this model for $\mu_n + \mu_p$ is plotted in Fig. 2.13 additionally. The curve runs close to the former ones, except above 3×10^{17} cm^{-3} where the result of Klaassen becomes considerably higher. The difference against the measurements is explained in [Kla92] with heating of the samples by the current pulses used in the measurements of [Dan72, Kra72]. However, also simplifying assumptions of the theory may be a cause for the discrepancy.

Since the mobility theory is very complicated, a simplified description can be useful for a survey. For this purpose it is of interest how the mobilities at high injection levels as function of the carrier concentration compare really with the dependency on the doping density in thermal equilibrium. To show this, the quantity $\mu_n + \mu_p$ obtained from Eq. (2.35) with N replaced by the carrier concentration n = p is plotted in Fig. 2.13 additionally. Again the parameter sets for μ_n and

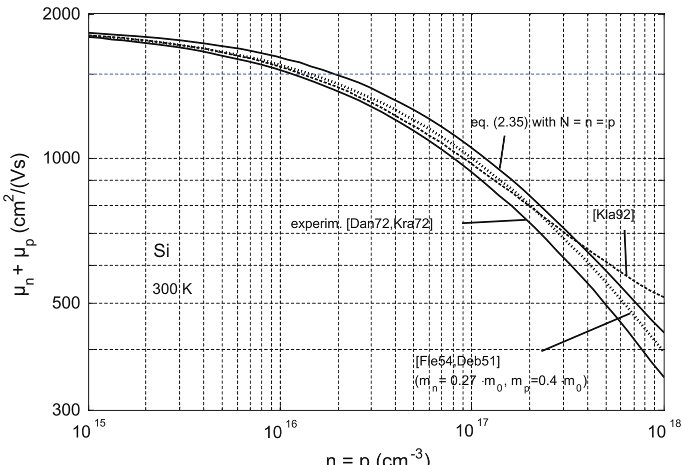


Fig. 2.13 Mobility sum at high injection levels as a function of carrier concentration

μ_p as given in Appendix A are used. As is seen, the obtained mobility sum as function of the carrier concentration differs not much from the other curves. If the difference against the experimental dependency proves true in view of a possible heating and the spread of the measurements (about $\pm$ 10% around the fitting curve), electron-hole scattering is somewhat more effective than ion scattering.

In a rough approximation, however, the measurements suggest that impurity ions in thermal equilibrium and carriers at high injection levels are nearly equally effective in reducing the mobility sum. This will hold hence also for the single mobilities. Since scattering at the two types of scatterers, ions and carriers with opposite charge, may be assumed to be independent of one another, the (approximate) equivalence follows also for a mixture of both. Hence defining N in (2.35) as the total concentration of scatterers, $N = N_D + p$ in the $\mu_n(N)$ expression and $N = N_A + n$ in $\mu_p(N)$, the equation describes the effect of impurity *and* electron-hole scattering. The electron mobility in an n-region and the hole mobility in a p region each with arbitrary minority carrier injection level is obtained hence as:

$$\mu_n = f_n(N_D + p) \tag{2.36a}$$

$$\mu_p = f_p(N_A + n) \tag{2.36b}$$

where $f_n(N)$ and $f_p(N)$ are the function (2.35) with the μ_n and μ_p parameter set, respectively. Scattering of electrons by electrons and of holes by holes has not been considered, since their influence is of second order.[2] Like carrier scattering is included nevertheless to a significant part, because it is contained in the used empirical relationship for thermal equilibrium and the reproduced dependency at high injection level. The range of approximate validity of (2.36a, b) reaches to minority carrier concentrations of about 5×10^{17} cm^{-3}, as pointed out below.

The model can be extended to the presence of a compensating doping. The mobilities are influenced then in addition to attractive also by repulsive impurity ions. Although of considerable practical interest, experimental results on mobilities in compensated semiconductors are hardly available. Theoretically, repulsive Coulomb fields have (approximately) the same scattering cross section and hence impact on the mobilities as attractive [Mol64, Smi59, Kla92], provided the concentrations are smaller than about 5×10^{17} cm^{-3} at normal operating temperatures. This restriction occurs because the carrier velocities and the minimum distances of the scattered particles from the Coulomb centers must not be too small on average. Hence, with this limitation of the concentration of repulsive ions, the mobilities in compensated regions in thermal equilibrium are given by (2.35) with the

[2] Like-carrier scattering does not influence the mobilities directly, because the total momentum of two colliding particles and hence the current they transport are not changed by the impact. Nevertheless it reduces the mobilities by randomizing the velocity distribution of carriers, which makes scattering by ions and phonons more effective [Deb54, Luo71]. The influence of like-carrier scattering on the effectiveness of *electron-hole* scattering is unclear till now.

concentration N defined as the sum of donor and acceptor concentrations, $N = N_D + N_A$. Generally, if also injected carriers are present, the mobilities are given by

$$\mu_n = f_n(N_D + N_A + p) \tag{2.37a}$$

$$\mu_p = f_p(N_A + N_D + n) \tag{2.37b}$$

where $f_n(N)$ and $f_p(N)$ are again the function (2.35) with the μ_n respectively μ_p parameter set. The mobilities are determined by the sum of the doping concentrations (while the majority carrier concentration equals the positive *difference*). (2.37a, b) is applicable also to the mobility of minority carriers. For example the hole mobility in an n-region with $N_A = 0$ follows from (2.37a) in thermal equilibrium as $\mu_p = f_p(N_D + n) = f_p(2N_D)$. Because of the additional scattering by majority carriers the minority carrier mobility is smaller than the mobility of majority carriers for equal doping concentration.

At higher concentrations, however, the scattering cross section of repulsive ions is considerably smaller than for attractive Coulomb fields, as is obtained from a rigorous quantum mechanical scattering theory [Kla92]. Also electron-hole scattering becomes less effective at high carrier concentrations. These effects can be taken into account in (2.37a, b) supplying the concentrations of repulsive impurity ions and of scattering carriers with weighting factors which are smaller than 1. One consequence is that the minority carrier mobilities at doping densities higher than about 2×10^{18} cm^{-3} are found to be *higher* than the respective majority mobilities, although electron-hole scattering adds to repulsive impurity scattering. Below about 2×10^{17} cm^{-3} on the other hand, the weighting factor of the carrier concentrations becomes somewhat higher than 1, reaching about 1.4 [Kla92]. This can be a cause of the deviation of the dependency obtained from (2.35) from the experimental curve in Fig. 2.13. As mentioned, another high concentration effect is screening of the Coulomb centers by oppositely charged carriers. In contrast to the empirical fitting Eq. (2.35), the extensions (2.36a, b), (2.37a, b) do not include screening effects produced by injection and compensation at high concentrations. Injection enhances and compensation lowers screening and mobilities as compared with (2.37a, b). The model of Klaassen [Kla92] includes screening in a rough manner. Like carrier scattering is neglected in the theoretical part of this model.

The *temperature* dependence of mobilities is determined by the same effects as the dependence on doping density. At *low* doping, scattering on thermal lattice vibrations predominates and yields a decrease of mobilities with temperature approximately proportional to T^{-2}. Impurity scattering is shown theoretically to yield a mobility μ_i *increasing* with temperature nearly as $T^{3/2}$, if screening is neglected. This comes from the increase of the thermal velocity causing a decrease of scattering by the Coulomb centers. The increase of thermal velocity, however, implicates also a decrease of screening which at very high doping density inverts the temperature dependence of μ_i to a slow *decrease* with T. These effects of the

partial mobilities μ_1 and μ_i find themselves also in the total mobility μ which is formed according to the approximate Matthiesen rule

$$\frac{1}{\mu} = \frac{1}{\mu_l} + \frac{1}{\mu_i} \tag{2.38}$$

Considering the temperature range 250 – 450 K, the result is that at low doping the (total) mobilities in Si show a strong decrease with increasing T, an approximate independency on T in the doping range 10^{18} cm^{-3} to 10^{19} cm^{-3} and at very high doping a minor decrease. The experimental temperature dependence can be well described – together with doping dependence – using temperature-dependent parameters μ_0, $\mu\infty$, N_{ref} and γ in formula (2.35) as given in Appendix A.

In Fig. 2.14 the temperature dependency of μ_n and μ_p in silicon are shown for two doping densities, one which is typical for the n-base region of power devices (left hand side) and the other for the p-base in thyristors, IGBTs and bipolar transistors (right hand side). The temperature dependence at 3×10^{17} cm^{-3} is considerably weaker than at 1×10^{14} cm^{-3}. That the resistivity $\rho = 1/(q\mu_p N_A)$ of the p base of mentioned devices increases relatively weakly with T, is very desirable for the *lateral* resistance of the p base (see Chaps. 8 and 10). Particularly the on-state voltage drop including the on-resistance of MOSFETs depends essentially on the mobilities and their temperature dependence. Equations (2.36a, b), (2.37a, b) together with the parameter set of Appendix A provide also a model for the temperature dependence of the mobilities at high injections and in the case of compensation.

2.6.2 *High-Field Drift Velocities*

At high electric fields, where condition (2.30) is not satisfied, the drift velocity is no longer proportional to the field, but increases weaker. At very high field strength it

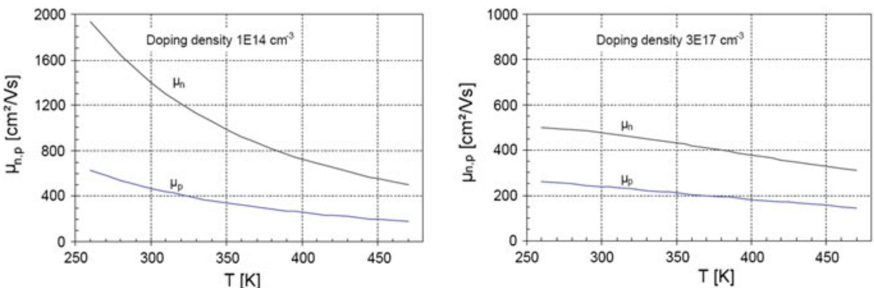


Fig. 2.14 Temperature dependence of mobilities in Si at two doping densities (see text)

approaches a limiting value, the saturation drift velocity v_{sat}. The dependency of the drift velocities of electrons and holes on the electric field is often expressed in the form [Cau67, Tho80]:

$$v_{n,p} = \frac{\mu_{n,p}^{(0)} \cdot \mathsf{E}}{\left(1 + \left(\frac{\mu_{n,p}^{(0)} \cdot \mathsf{E}}{v_{sat(n,p)}}\right)^{\beta}\right)^{\frac{1}{\beta}}} \tag{2.39}$$

where $\mu_{n,p}^{(0)}$ are the low-field mobilities discussed above.[3] The additional marking is introduced, because Eq. (2.29) is used also for the non-linear range of fields where the mobilities $\mu_{n,p} \equiv v_{n,p}/\mathsf{E}$ are field-dependent. For small E, (2.39) turns into (2.29) with $\mu_{n,p} = \mu_{n,p}^{(0)}$. The exponent β was proposed by Caughy and Thomas [Cau67] to be $\beta_n = 2$ for electrons and $\beta_p = 1$ for holes. Jacobini and coworkers [Jac77] have determined the β's renewed and as functions of temperature. They obtained $\beta_n = 2.57 \times 10^{-2} \times T^{0.66}$, $\beta_p = 0.46 \times T^{0.17}$, which yields a value near 1 at 300 K in *both* cases. Besides $\mu_{n,p}$, and $\beta_{n,p}$, also the saturation velocities are functions of temperature. According to Jacobini et al. [Jac77] they decrease with T as $v_{sat(n)} = 1.53 \times 10^9/T^{0.87}$, $v_{sat(p)} = 1.62 \times 10^8/T^{0.52}$ in cm/s.

Using these results in (2.39), one obtains at 300 K the field dependences shown in Fig. 2.15. Below 10^3 V/cm, fields which occur in devices during the conducting state, one obtains the linear dependency described by constant mobilities. Soon above this value, however, the increase of v_n and somewhat later that of v_p becomes sublinear, as is expected also from condition (2.30). At 3×10^4 V/cm the drift velocities approach the respective saturation velocity $v_{sat(n)}$ or $v_{sat(p)}$. These range near 1×10^7 cm/s and thus already reach the order of magnitude of the mean thermal velocity. Fields in the range up to typically 2×10^5 V/cm appear in space charge regions during blocking of devices. Hence for a wide range of blocking voltages, carriers in the space charge region move with saturation velocity.

Relationship (2.39) has been verified experimentally in pure silicon [Jac77], for doped silicon experimental data of the high-field drift velocity are not available. Using the doping dependent mobility $\mu_{n,p}^{(0)}(N)$ in (2.39), however, a physically reasonable dependence on E and doping density N is obtained which satisfies general scaling requirements [Tho80]. The saturation velocities $v_{sat(n,p)}$ are assumed to be independent of N. An alternative approach for $v_{n,p}$ as functions of E and N has been given by Scharfetter and Gummel [Scf69], see Appendix A1.

[3] In [Jac77] a second expression for the saturation velocity of electrons has been given, it reads:

$$v_{sat(n)} = \frac{2.4 \cdot 10^7 \text{cm/s}}{1 + 0.8 \cdot \exp(T/600)}.$$

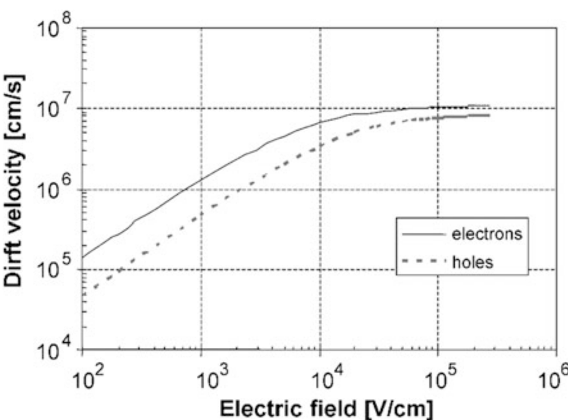


Fig. 2.15 Drift velocity of electrons and holes as function of the electric field. Temperature 300 K

2.6.3 Diffusion of Carriers, Current Transport Equations and Einstein Relation

Unlike the situation in metals, the current in semiconductor devices is caused often not only by an electric field but also by diffusion of carriers. Generally, if some mobile particles have a spatially variable concentration C, they diffuse from a region of high to a region of low concentration. The particle current density J thus arising is proportional to the negative concentration gradient (Fick's first law):

$$J = -D \cdot \nabla C \tag{2.40}$$

where the proportionality factor D is the diffusion constant. This holds also for electrons and holes, whose concentrations can vary because of a variation of the doping concentration or as result of injection of carriers. Due to the charge of the carriers the particle currents are connected with electrical currents. Multiplying with ± q and assuming the concentration gradient to appear in x direction, the electrical diffusion current densities are

$$j_{n,diff} = -q \cdot D_n \cdot \frac{dn}{dx} \tag{2.41}$$

$$j_{p,diff} = -q \cdot D_p \cdot \frac{dp}{dx} \tag{2.42}$$

Together with the field currents, Eq. (2.31), the total current densities are given by the transport equations:

$$j_n = q \cdot \left(\mu_n \cdot n \cdot \mathsf{E} + D_n \cdot \frac{dn}{dx} \right) \tag{2.43a}$$

$$j_p = q \cdot \left(\mu_p \cdot p \cdot \mathsf{E} - D_p \cdot \frac{dp}{dx} \right) \tag{2.43b}$$

The diffusion constants D_n, D_p depend on the same scattering mechanisms as the mobilities. In fact, they are related to the mobilities by the following simple relationship

$$D_{n,p} = \frac{kT}{q} \cdot \mu_{n,p} \tag{2.44}$$

which is called *Einstein relation*. This can be derived from the case of thermal equilibrium [Sho59]: As follows from (2.43b) with the thermal equilibrium condition $j_p = 0$, a concentration gradient dp/dx is connected with a field E and hence with an electrical potential $V(x) = -\int \mathsf{E} dx$. Now the holes obey the Boltzmann distribution

$$p(x) \propto \exp\left(-\frac{qV(x)}{kT} \right) \tag{2.45}$$

as can be concluded from (2.3). The energy states are located here at different points in space. On the other hand, one obtains from (2.43b) with $j_p = 0$

$$\begin{aligned} \frac{d \ln p}{dx} &= \frac{\mu_p}{D_p} \mathsf{E} = -\frac{\mu_p}{D_p} \cdot \frac{dV}{dx} \\ p &\propto \exp\left(-\frac{\mu_p}{D_p} V(x)\right) \end{aligned} \tag{2.46}$$

This is compatible with the Boltzmann distribution (2.45) only if $D_p/\mu_p = kT/q$. Similarly one can proceed for electrons. Hence the Einstein relation (2.45) is an immediate consequence of the Boltzmann distribution. The factor kT/q in (2.45) has the dimension of a voltage and is called thermal voltage. Its value at 300 K is 25.85 mV. Hence the diffusion constants measured in cm^2/s amount to only about 1/40th of the respective mobility in cm^2/Vs.

The Einstein relation is used very extensively not only for thermal equilibrium assumed for its derivation, but also for non-equilibrium up to high current densities and high injection levels. In a series of papers (see [Mna87a, Mna87b, Kan93, Mna98]) a model has been developed and applied claiming that the opposite drift velocities of electrons and holes in the on-state at high injection levels leads to complete breakdown of the Einstein relation due to a mutual drag. This model contradicts partly the conventional mobility theory and results in considerably

different device characteristics.[4] Hence the question arises which of the two approaches is appropriate for power device simulation.
The answer follows from the fact that virtually in all cases of interest in this respect the drift velocity of carriers is small against their thermal velocities:

$$v_{n,p} \ll v_{n,ptherm}. \tag{2.47}$$

If for example the electron current density j_n in the base region of a pin-diode is as high as $5 \times 10^3 A/\mathrm{cm}^2$ and the electron concentration at its minimum $5 \times 10^{17}\mathrm{cm}^{-3}$, the drift velocity of electrons, $v_n = j_n/qn$, is still less than 1% of their mean thermal velocity at 300 K. In the whole range of operating temperatures and even down to 100 K the condition (2.47) is well fulfilled up to high current densities. At high fields, where (2.47) is not satisfied, the diffusion currents are usually negligible and the diffusion constants hence not relevant. Since the Maxwell distribution of carrier velocities holds still approximately, the microscopic scattering processes determining mobilities and diffusion constants are the same as in thermal equilibrium. Hence the relation between these constants should not be changed. Similarly as the doping density the concentrations of injected carriers determines the number of active and passive scatterers without influencing the relation between $\mu_{n,p}$ and $D_{n,p}$.

To prove this analogously to the thermal equilibrium case one has to take into account, that the mutual compensation of field and diffusion currents used in the above derivation does no more happen approximately at high current densities. One can assume however, regarding the case of holes, a field strength $E_{eq,p} \equiv D_p/\mu_p d \ln p/dx$ (equilibrium field for holes), at which the hole diffusion current is compensated by the field current, the ohmic part of the field which is proportional to the hole current being disregarded. Depending on the injected carrier distribution the field $E_{\mathrm{eq,p}}$ varies with space coordinate x. Since $j_p = 0$ for $E = E_{\mathrm{eq,p}}$, the Eq. (2.46) follows as above with the potential $V(x)$ resulting from this field. Because due to the condition (2.47) also the Boltzmann distribution (2.45) is valid approximately, the Einstein relation between μ_p and D_p results for the field $E_{eq,p}(x)$. Since (2.47) however is also the condition for field-independent mobilities (see Eqs. 2.30, 2.29), the relation (2.44) follows for the total field including the

[4] For example, the ambipolar diffusion constant

$$D = \frac{2D_n D_p}{D_n + D_p},$$

relevant for the carrier distribution at high injection levels (see Sect. 5.4.1), is proposed in mentioned papers to be independent of the carrier concentration $n = p$. While D_n decreases with increasing n, the hole diffusion constant D_p *rises*. Obviously this disagrees with (2.45) since μ_n *and* μ_p decrease with n. Also the mobility ratio μ_n/μ_p is obtained to be constant in cited papers in contrast to the conventional theory. This results in strongly different high current characteristics of asymmetrical pn^-n^+ diodes.

ohmic part. The analogous argumentation applies to μ_n and D_n. Hence *under operating conditions of power devices including high current densities and high injection levels, the Einstein relation will be applicable with good accuracy.*

2.7 Recombination—Generation and Lifetime of Non-equilibrium Carriers

In thermal equilibrium, charge carriers are continuously generated thermally and they disappear with the same rate by recombination. In devices during operation, however, carrier densities in the active region are not in thermal equilibrium, they are higher or lower than the densities according to the equilibrium Eqs. (2.4), (2.5) and (2.6). A non-equilibrium state tends to restore itself to equilibrium. The time within which the system strives to achieve this, if injection/extraction and external generation are turned off, is determined by the *lifetime* τ of the non-equilibrium carriers. This is an adjustable quantity which is decisive for the dynamic as well as static behavior of power devices.

The definition of the lifetime uses the net recombination rates R_n and R_p of electrons and holes which are defined as the difference between the thermal re-combination rates $r_{n,p}$ and thermal generation rates $g_{n,p}$:

$$R_n \equiv r_n - g_n, R_p \equiv r_p - g_p.$$

These thermodynamic quantities, which are zero in thermal equilibrium, describe the decrease of n and p with time due to the net thermal recombination. Hence one has

$$R_n \equiv r_n - g_n = -\left(\frac{\partial n}{\partial t}\right)_{rg} \quad R_p \equiv r_p - g_p = -\left(\frac{\partial p}{\partial t}\right)_{rg} \tag{2.48}$$

where the index ‘rg’ marks the part of the time derivatives owing solely to recombination and generation. An in/outflow of carriers into/out of the considered volume element as well as external generation, for example by light, is excluded in (2.48). R_n, R_p depend on doping and carrier densities and increase with increasing deviation of the respective carrier density from the equilibrium density n_0 or p_0. The relationship between R_n, R_p and the excess concentration $\Delta n \equiv n - n_0$, respectively $\Delta p \equiv p - p_0$ is usually not very far from a linear dependency. Therefore it is useful to define lifetimes τ_n, τ_p by the equations:

$$R_n \equiv \frac{\Delta n}{\tau_n}, \qquad R_p \equiv \frac{\Delta p}{\tau_p} \tag{2.49}$$

Compared with R_n, R_p, the lifetimes depend not very strongly on the respective excess concentration, and in some significant cases they are actually constant. This holds for the lifetime of injected minority carriers, the *minority carrier lifetime*, if the injection level is low (density of minority carriers is small compared to the majority concentration).
For the decay of a homogeneous excess concentration, e.g. of *Δp(t)*, Eqs. (2.48) and (2.49) yield:

$$\frac{d\Delta p}{dt} = -\frac{\Delta p}{\tau_p}, \qquad \Delta p = p(0) \cdot e^{-t/\tau_p} \tag{2.50}$$

where in the latter equation τ_p was assumed to be constant. From this the meaning of τ as a *lifetime* of the excess carriers becomes obvious. In the *stationary* case the disappearance of carriers due to recombination (for $R_n > 0$) is compensated by a net inflow of carriers or by external generation. In this case the net recombination rates of electrons and of holes are equal:

$$R_n = R_p = R \tag{2.51}$$

since the number of electrons leaving the conduction band must equal the number of electrons entering the valence band, the charge on possible intermediate levels being constant in the stationary case. If energy levels in the band gap are not involved, (2.51) is valid also for time dependent processes.

We have talked till now about recombination in the *volume* of the semiconductor. Before continuing with this, it is the point here to note that also *recombination at the metallic contacts*, especially at the anode contact, is very significant for modern power devices. The recombination at the anode metal layer contacting a p-region results in an electron particle current to the contact given by

$$J_n = (n - n_0)s \tag{2.52}$$

where s is the surface recombination velocity at the contact. For the recombination process the same hole current is needed, which flows in blocking direction and during on-state reduces the total hole current. It is used hence to control the emitter efficiency. s is typically of the order of magnitude 10^5 cm/s and approximately constant. The high surface recombination velocity arises from a high density of interface states, distributed continuously over the bandgap. The recombination at a contact is controlled in devices by the integral doping concentration in front of it. Since this is treated later in the book together with the impact on switching properties (see Sect. 5.7.4.3), this note suffices here.

Returning to the *volume recombination* three physical *mechanisms* are to be mentioned: (i) recombination at recombination centers formed by ‘deep impurities’ or ‘traps’ which have energy levels deep in the band gap, (ii) band-to-band Auger recombination and (iii) radiative band-to-band recombination. The latter two mechanisms occur in the semiconductor lattice itself and depend only on carrier

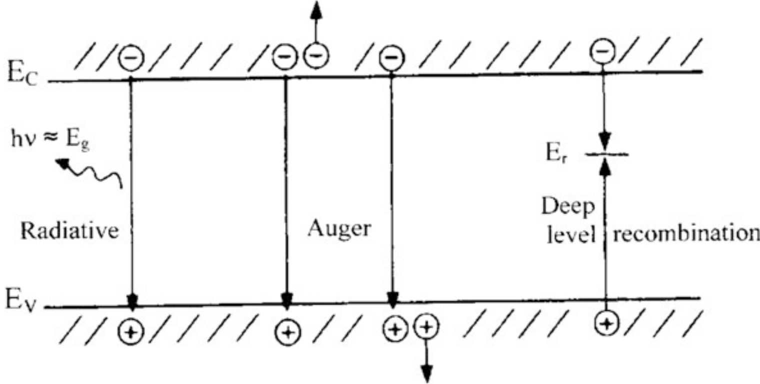


Fig. 2.16 Mechanisms of recombination. **a** radiative band-to-band recombination. **b** Auger recombination with energy transfer to a third carrier and **c** recombination via a deep energy level

concentrations, not directly on the density of normal and deep impurities. The three mechanisms are illustrated in Fig. 2.16. The total recombination rate is made up additive of the single parts, hence according to (2.48) the inverse total *lifetime* is obtained by adding the reciprocal single lifetimes τ_i :

$$\frac{1}{\tau_{tot}} = \sum \frac{1}{\tau_i} \tag{2.53}$$

This applies also to the superposition of single lifetimes caused by different independent kinds of traps if present. In the following sections, the intrinsic mechanisms are described first.

2.7.1 Intrinsic Recombination Mechanisms

(a) Radiative band-to-band recombination: As shown in Sect. 2.4, this direct recombination of an electron and a hole under transfer of the released energy to a light quant has a high probability only in direct semiconductors. According to simple statistics the net recombination rate is

$$R = B \cdot (n \cdot p - n_i^2) \tag{2.54}$$

with the radiative recombination probability B.

The radiative lifetime, for example the hole lifetime in an n-type semiconductor, is obtained with (2.49) as $\tau_{prad} = \Delta p/R = 1/(B \cdot n)$ since $np - n_i^2 = n_0 \Delta p + p_0 \Delta n + \Delta n \Delta p = n \Delta p$ assuming $p_0 << n_0$: Hence, the radiative minority carrier lifetime is inversely proportional to the majority carrier density. In most

cases $n{\cdot}p >> n_i^2$, so that $R \approx B{\cdot}n{\cdot}p$. In GaAs the radiative lifetime is estimated to be 6 μs at a doping density N = 1 × 10^{15} cm^{-3} or 60 ns at 1 × 10^{17} cm^{-3} [Atk85]. This small lifetime limits the applicability of GaAs for bipolar devices. In silicon, the recombination constant at 300 K is B ≈ 1 × 10^{-14} cm^3/s [Sco74] yielding τ_{rad} = 1 ms at a majority carrier density n = 1 × 10^{17} cm^{-3}. In practice such a high lifetime is not measured in silicon devices, because the other recombination mechanisms are more effective (see the next sections). Using the connection between the injected carrier concentrations and the intensity of the recombination radiation, the radiation is used to investigate the inner operation of devices.

(b) Band-to-band Auger recombination: In Auger recombination, the energy released during the recombination event is transferred not to a light quant but to a third electron or hole, where participation of a phonon can be required for conservation of momentum. Therefore the recombination probability B in (2.54) has to be replaced by a factor which is proportional to the carrier concentrations. Hence the Auger recombination rate is

$$R_A = \left(c_{A,n} \cdot n + c_{A,p} \cdot p\right) \cdot \left(n \cdot p - n_i^2\right) \tag{2.55}$$

The coefficients $c_{A,n}$, $c_{A,p}$ determine the recombination rate for the cases that the third carrier taking the energy away is an electron, respectively a hole. Since the concentrations appear with power 3, the probability of this mechanism increases strongly with carrier concentration, and the lifetime *decreases* strongly. Therefore the Auger recombination is important mainly in highly doped regions. In an n^+ region with a small concentration of injected holes, with $p << n$ and $n{\cdot}p >> n_i^2$ (2.55) turns into $R_A = c_{A,n}{\cdot}n^2{\cdot}p$, and for the hole lifetime one obtains from (2.49):

$$\tau_{A,p} = \frac{p}{R_A} = \frac{1}{c_{A,n} \cdot n^2} \tag{2.56}$$

where the extremely small equilibrium concentration p_0 has been neglected. The formula for the electron lifetime in a p^+ region is formed analogously. The Auger coefficients in silicon are in the range of 10^{-31} cm^6/s, according to [Dzi77] their values are:

$$c_{A,n} = 2.8 \times 10^{-31} \mathrm{cm}^6/s, \quad c_{A,p} = 1 \times 10^{-31}\ \mathrm{cm}^6/s \tag{2.57}$$

They are approximately independent of temperature. For a doping density of 1 × 10^{19} cm^{-3}, the Auger electron lifetime in a p^+ region is $\tau_{A,n} = 1/(c_{A,p}{\cdot}p^2) = 0.1\ \mu s$, and the hole lifetime in an n^+ region is 0.036 μs. The small lifetime in highly doped regions is a constituent part of the h parameters via which the properties of these regions influence the characteristics of devices. This will be shown in Sect. 3.4.

Another case where the Auger recombination is of significance in devices is that of high concentrations of injected carriers in a weakly doped base region. With

neglect of the doping density the neutrality requires p ≈ n which inserted in (2.55) yields

$$R_{A,hl} = \left(c_{A,n} + c_{A,p}\right) \cdot p^3 \tag{2.58}$$

Hence, the high-level Auger lifetime is

$$\tau_{A,hl} = \frac{1}{\left(c_{A,n} + c_{A,p}\right) \cdot p^2} \tag{2.59}$$

At $p = n = 3 \times 10^{17}$ cm^{-3}, this relation together with (2.57) results in an Auger lifetime of 29 µs. At high current densities, the Auger recombination in the base region of high voltage devices becomes noticeable.

2.7.2 *Recombination at Recombination Centers Including Gold, Platinum and Radiation Defects*

The recombination via deep energy levels in the band gap caused by appropriate 'deep impurities' or lattice imperfections is the dominant mechanism of recombination in lowly and intermediately doped regions in silicon devices. By these recombination centers called also 'traps' the lifetime can be controlled over a wide range, which is commonly used to reduce the switching time and switching losses of devices for higher frequencies. The doping with deep impurities is carried out after the normal doping which determines the conductivity. In the history of device technology, at first gold was used as deep impurity for lifetime control in silicon. Meanwhile, many power devices are diffused with platinum, and most important is now the creation of lattice defects with deep levels by electron, proton or α-particle irradiation.

Unfortunately the reduction of switching time is associated with an increase of the on-state voltage drop and the high-temperature blocking current and often also

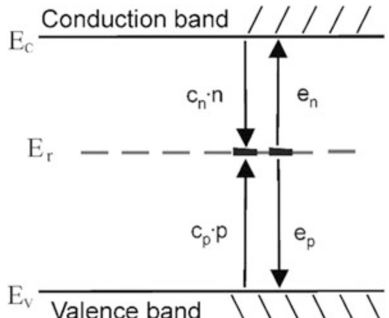


Fig. 2.17 Capture and emission of carriers at a recombination center

of the resistivity of the base region by compensation. Since dynamic and stationary device properties are hence intimately connected with the adjustment of the lifetime, this is a central point in the development and fabrication of power devices. The recombination at a deep impurity proceeds in two steps, the capture of a conduction electron which then occupies the deep energy level, and thereafter a falling down of the electron into an empty place of the valence band, meaning capture of a hole by the impurity (see Fig. 2.17). Vice versa, an electron-hole pair is generated by thermal emission of a valence electron first to the impurity level, i.e. emission of a hole from the impurity to the valence band, and then emission of the electron from the impurity level to the conduction band.

The energy released during capture of a carrier is transferred to lattice vibrations, and the energy required conversely for generation is taken up from the lattice. Because of the large band-to-level distance a series of phonons are emitted respectively absorbed during capture and emission. These multi-phonon processes of capture and emission, however, are considered as a whole and described by over-all capture and emission probabilities. Using this concept Shockley, Read [Sho52] and Hall [Hal52] have developed a theoretical model (SRH model) which describes how recombination and generation depends on the level position, the capture and emission probabilities, type and concentration of normal doping, the injection level (minority carrier concentration) and temperature. This model is discussed in the following in detail. Although most deep impurities have two or even more levels, first centers with one level are considered.

(a) Recombination centers with one deep level: We consider the recombination at a center R which can appear in a neutral and a negative charge state R^0, R^-. The impurity level is called in this case an acceptor level independent of its position in the bandgap. Similarly, if the charge state of the impurity atom changes from positive to neutral when the level is occupied by an electron, the level is called a donor level. The capture of electrons by centers R^0 causes an electron recombination rate $r_n = c_n \cdot n \cdot N_r^0$ where N_r^0 is the concentration of the neutral centers and c_n a constant called capture probability or capture rate. The electron *generation* rate, given by emission of electrons from centers R^- to the conduction band, is proportional to the concentration of negatively charged centers N_r^-, $g_n = e_n \cdot N_r^-$, where the constant e_n is the emission probability also called emission rate of electrons. Hence the net recombination rate R_n is

$$R_n = c_n n N_r^0 - e_n N_r^- \tag{2.60}$$

From thermal equilibrium whith $R_n = 0$, one obtains, indicating the concentrations by a ‘0’ and using (2.21), (2.6), the following relationship between emission and capture probability

$$e_n = c_n \frac{n_0 N_{r0}^0}{N_{r0}^-} = c_n n_r \tag{2.61}$$

with

$$n_r = N_c g \, \exp\left(-\frac{E_c - E_r}{kT}\right) \tag{2.61a}$$

Similarly, the capture of holes by R^- and emission (generation) of holes from R^0 result in the net recombination rate of holes

$$R_p = c_p p N_r^- - e_p N_r^0 \tag{2.62}$$

where c_p is the hole capture probability and e_p the emission probability of holes. A relationship between these quantities is obtained again from thermal equilibrium with $R_p = 0$ using (2.21)

$$e_p = c_p \frac{p_0 N_{r0}^-}{N_{r0}^0} = c_p p_r \tag{2.63}$$

with

$$p_r = \frac{N_v}{g} \exp\left(-\frac{E_r - E_v}{kT}\right) = n_i^2 / n_r \tag{2.63a}$$

The concentrations n_r, p_r which relate the emission probabilities to the corresponding capture probabilities are apart from the degeneracy factor g identical with the electron or hole density, respectively, which are present if the Fermi level coincides with the recombination level. Using the relation $n_r p_r = n_i^2$, Eqs. (2.61), (2.63) yield:

$$e_n e_p = c_n c_p n_i^2 \tag{2.64}$$

Resulting from thermal equilibrium of elementary reactions of the recombination process, the relations between capture and emission rates are called detailed balance equations. Since the capture rates are defined in a low-field environment, also the emission rates correlated with them are low-field emission rates. Up to a field strength of 1×10^5 V/cm, however, they seem to be constant [LuN87].

With the definition of n_r, p_r Eqs. (2.60), (2.62) take the form:

$$R_n = c_n (n N_r^0 - n_r N_r^-) \tag{2.65}$$

$$R_p = c_p (p N_r^- - p_r N_r^0) \tag{2.66}$$

As mentioned earlier, in the stationary case and generally if the time variation of charge on the deep level is negligible, one has $R_n = R_p$. Equating hence the right hand sides of (2.65) and (2.66) and using the total concentration $N_r = N_r^0 + N_r^-$, one can solve for N_r^- and N_r^0 to obtain:

$$N_r^- = \frac{c_n n + c_p p_r}{c_n(n + n_r) + c_p(p + p_r)} N_r, \quad N_r^0 = N_r - N_r^- \tag{2.67}$$

Inserting this into Eqs. (2.65) and (2.66) the net recombination rate is obtained as the following function of n and p:

$$\begin{aligned} R_n = R_p = R &= c_n c_p N_r \frac{n \cdot p - n_i^2}{c_n(n + n_r) + c_p(p + p_r)} \\ &= \frac{n \cdot p - n_i^2}{\tau_{p0} \cdot n + \tau_{n0} \cdot p + \tau_g \cdot n_i} \end{aligned} \tag{2.68}$$

τ_{n0}, τ_{p0} and τ_g are lifetime quantities defined as

$$\tau_{p0} = \frac{1}{N_r c_p}, \quad \tau_{n0} = \frac{1}{N_r c_n} \tag{2.69}$$

$$\tau_g = \frac{n_i}{N_r}\left[\frac{1}{e_n} + \frac{1}{e_p}\right] \tag{2.70}$$

These are the central equations of the SRH model. They are valid also for a donor level except that N_r^0 has to be replaced by N_r^+ and N_r^- by N_r^0 and the degeneracy factor g by 1/g. For the hole concentration $p_d = p_r$ of a donor level Eq. (2.63a) turns into Eq. (2.61a) with N_c replaced by N_v, the equation for p_d has the same form as the concentration $n_r = n_a$ of an acceptor level. The capture coefficients have typically values in the range 10^{-9} to 3×10^{-7} cm^{-3}/s at 300 K, they decrease mostly slightly with temperature. So, a concentration N_r of only 1×10^{13} cm^{-3} results in lifetime values τ_{n0}, τ_{p0} in the range 0.3–100 µs.

We consider first the consequences of the SRH model for the minority carrier lifetime in a neutral region, thereafter the generation rate in a space charge region. Together with the definition (2.49), Eq. (2.68) yields the lifetime as function of n and p. In an n-type region with $n_0 \gg p_0$ and $np - n_i^2 = n_0 \Delta p + p_0 \Delta n + \Delta n \Delta p$ the hole lifetime $\tau_p = \Delta p/R$ is obtained as follows:

$$\tau_p = \tau_{p0} + \tau_{n0}\frac{p}{n} + \tau_g \frac{n_i}{n} = \tau_{p0}\left(1 + \frac{n_r}{n}\right) + \tau_{n0}\left(\frac{p + p_r}{n}\right) \tag{2.71}$$

In the case of neutrality, the electron and hole concentrations are connected with each other by the condition $n = N_D^+ + p \approx n_0 + p$, if the charge on the traps can be neglected. This is often allowed in a first approximation because the trap

concentration is mostly about an order of magnitude smaller than that of the normal doping. According to Eqs. (2.61a) and (2.63a) n_r, p_r depend exponentially on the position of the recombination level and on temperature. Except the level lies near the middle of the bandgap the concentration referring to the more distant band, n_r or p_r, is negligible. At least at higher temperatures, the higher of the concentrations n_r, p_r has a considerable effect on τ assuming the concentration n_0 is not much higher than about 10^{14} cm^{-3}. The lifetime at low injection level ($p << n_0 \approx n$) is

$$\tau_{p,LL} = \tau_{p0}(1 + \frac{n_r}{n_0}) + \tau_{n0}\frac{p_r}{n_0} \tag{2.72}$$

and this becomes equal to τ_{p0} if the recombination level is located near the middle of the gap or the temperature is low ($n_O >> n_r$, p_r). Analogous holds for the electron lifetime τ_n in a p region. At high injection levels defined by the condition $n = p >> n_0$, p_0, both lifetimes turn into the high level lifetime

$$\tau_{hl} = \tau_{no} + \tau_{p0} + (\tau_{p0}n_r + \tau_{n0}p_r)/n \tag{2.73}$$

where also $\tau_{p0}n_r + \tau_{n0}p_r >> \tau_{n0}\ n_0$ was assumed. For a recombination level near a band edge and particularly at elevated temperatures, Eq. (2.73) showing a decrease of τ_{hl} with n applies to a considerable range of concentrations. As n increases further until also $n >> n_r$, p_r the high-level lifetime approaches the limiting value

$$\tau_{HL} = \tau_{n0} + \tau_{p0} \tag{2.74}$$

The lifetime τ_{HL} is larger than the low-level lifetime if $\tau_{p0}n_r + \tau_{n0}p_r < \tau_{n0}n_0$ or

$$c_n/c_p \cdot n_r + p_r < \ n_0 \tag{2.75}$$

The dependency on the injected hole concentration is very simple if the approximate neutrality condition $p = n - n_0$ can be used. Inserting this in (2.71) it is seen that the variation of τ_p between the low and high injection values is monotonous. This holds also if the trap charge is taken into account in the neutrality condition. With the equilibrium electron concentration $n_0 = N_D - N_r/(1 + n_r/n_0)$ (assuming further an acceptor level) formula (2.75) is the general condition for an increase of the minority carrier lifetime in an n-region with increasing p. If the left hand side is *larger* than n_0, the lifetime decreases. As can be shown by insertion of $n = N_D + p - N_r^-$ into (2.67), also N_r^-/N_r varies monotonously with p between its equilibrium value $n_0/(n_0 + n_r)$ at $p = p_0$ and the high injection value $c_n/(c_n + c_p)$ at $p \gg n_0, n_r, p_r$.

Besides the dependency on the injection level also the temperature dependence is connected with the quantities n_r, p_r. If one of these is comparable or larger than n_0 the variation of n_r or p_r with T results in a strong increase of τ with temperature. Furthermore a decrease with increasing doping concentration $N_D \approx n_0$ results. As follows from (2.72) the low-level lifetime decreases from

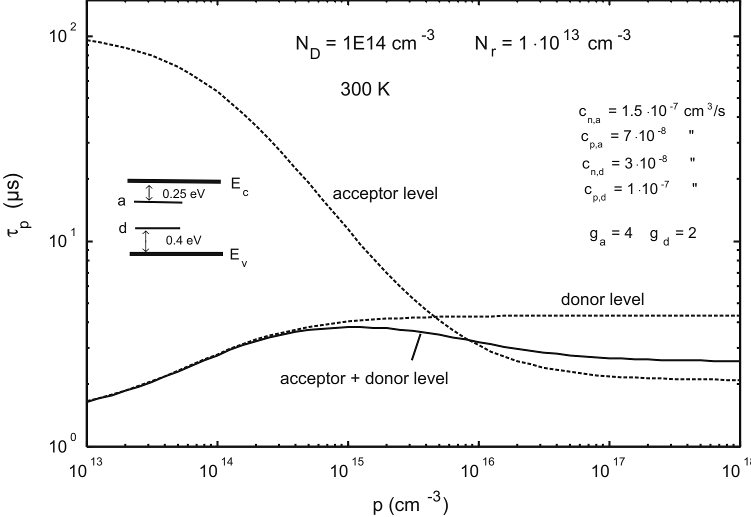


Fig. 2.18 Hole lifetime in n-type silicon for three exemplary recombination centers, (1) a donor trap with a level 0.4 eV above the valence band, (2) an acceptor trap with a level 0.25 eV below the conduction band (dashed lines) and (3) a trap possessing both levels for itself (solid line)

$$\tau_p = \frac{\tau_{p0} n_r + \tau_{n0} p_r}{n_0} \quad \text{at} \quad n_0 \ll \max(n_r, p_r) \tag{2.76}$$

to τ_{p0} for $n_0 \gg n_r$, p_r. Under the condition max(n_r, p_r) >> n_0 the lifetime $\tau_{p,LL}$, varies inversely proportional to n_0. By interchange of n and p the same expression results for the low-level electron lifetime in p-silicon except that n_0 is replaced by p_0. Hence the low-level minority carrier lifetime in n- and p-silicon with equal trap and doping density is equal under these conditions.

In Fig. 2.18 the lifetime $\tau_p(p)$ in n-Si is depicted for three examples, a donor trap with E_r = 0.4 eV above the valence band, an acceptor center with E_r = 0.25 eV below the conduction band and a center which possesses these two levels for its own (solid curve), which case will be discussed later. To distinguish between the capture rates for the donor and acceptor level they are indicated by an additional subscript 'd' respectively 'a'. The examples are chosen to illustrate how the lifetime depends on the position of the level in the bandgap. The capture probabilities are in the usual range. As degeneracy factors the usual values g_a = 4 for the acceptor and g_d = 2 for the donor level were used. The concentration of the recombination centers (1×10^{13} cm^{-3}) is significantly smaller than the normal doping density (1×10^{14} cm^{-3}). The charge on the recombination centers was taken into account in the calculation, but its effect is not large since N_r << N_D and additionally at low injection levels the centers are mostly neutral because the Fermi level lies below the acceptor level and above the donor level. As is seen, the lifetime for the donor trap increases with p as is expected from (2.71) and (2.75), since n_r is negligible and

according to Eq. (2.63a) (in the donor version, $g \to 1/g$) p_r is with 1.18×10^{13} cm^{-3} smaller than n_0. The lifetime caused by the acceptor center on the other hand decreases from a very high value at small p to a low high-level lifetime. Also this follows immediately from (2.71) since $n_r = 7.22 \times 10^{15}$ cm^{-3} >> n_0. At higher temperatures not included in the figure the lifetime particularly of the acceptor impurity is still much higher at small p and the decrease with increasing p correspondingly stronger. Similar temperature dependences are shown further below in the case of platinum.

What is the cause of the high low-level lifetime of the center with the level near the conduction band ($n_r >> n_0$)? The capture and emission rates of electrons are still nearly in thermal equilibrium, which means $R_n \approx 0$. With (2.65) this results in $N_r^- = n_0/n_r \cdot N_r^0 \approx n_0/n_r \cdot N_r << N_r$. Only very few centers are hence in the negative charge state in which they can capture a hole. Inserting N_r^- into Eq. (2.66), whose hole emission term is negligible against the capture term for $p \geq p_0$, one obtains $\tau_{p,LL} = 1/\left(c_p N_r^-\right) \approx \tau_{p0} \cdot n_r/n_0 \gg \tau_{p0}$. Also if the level lies near the valence band ($p_r >> \mathrm{n_0}$), the lifetime $\tau_{p,LL}$ is very high according to (2.72). In this case the capture and emission rate of *holes* are approximately in equilibrium, hence it follows from (2.66) $N_r^0 = p/p_r N_r$. Inserting this into (2.65) and neglecting the term with n_r, one obtains $R_p = R_n = c_n np/p_r \cdot N_r$ and together with (2.50) $\tau_{p,LL} = 1/(c_n N_r) \cdot p_r/n_0$. The *hole* lifetime is controlled here by the capture of *electrons* by the few R^0 atoms, which leaves the high number of capture and emission events of holes nearly in equilibrium. The general result is that an energy level near a band edge is little effective in recombination at low doping and injection levels with $n << n_r$ or $n << p_r$.

Both the very low ratio of high-level to low-level lifetime and the strong increase of the lifetime with temperature obtained for a near-band level are unfavorable for devices. The first property worsens the relation between on-state voltage drop and switching and recovery times. For thyristor's and IGBTs it is additionally unfavorable because the high low-level lifetime results in a high amplification of the generation current in the space charge region by the p$^+$np partial transistor [Cor74, Bal77]. In both cases the high-level to low-level lifetime ratio is particularly crucial at elevated temperatures, so the high *temperature* dependence comes into play here. A strong increase of the lifetime with temperature even at high injection levels can lead, furthermore, to a strong decrease of the on-state voltage with temperature, and this makes devices unsuitable for parallel connection [Lut00, Sie01] (see Sect. 5.4.6). *Hence a recombination level near a band edge is disadvantageous*. It can be made innocuous, however, by other levels, which lie nearer to the middle of the gap.

Equation (2.71) is also valid if the equilibrium is undercut, i.e. if $np < n_i^2$, provided $n >> p_0$. In a **space charge region** as found in reverse-biased pn junctions, however, usually *both* carrier concentrations are negligible. With $n = p = 0$ the generation rate is obtained directly from (2.68), (2.70) as:

$$-R = G = \frac{n_i}{\tau_g} = \frac{N_r}{1/e_n + 1/e_p} \tag{2.77}$$

Determining the generation, the time quantity τ_g is called a generation lifetime. Since it does not obey (2.49), however, it is strictly speaking not a carrier lifetime in the usual sense. τ_g and hence G depend strongly on the position of the level in the bandgap. Owing to the inverse interrelation between e_n and e_p (see Eq. 2.64), τ_g reaches its minimum and G its maximum at that energy $E_r = E_{r,m}$ where $e_n = e_p = \sqrt{c_n c_p}\, n_i$. Inserting this into (2.70) the generation lifetime at its minimum is obtained as $\tau_{g\,\min} = 2/(N_r\sqrt{c_n c_p}) = 2\sqrt{\tau_{n0}\tau_{p0}}$: the minimal value of τ_g is twice the geometrical mean value of τ_{n0} and τ_{p0}. The energy of maximal G is given by

$$E_{rm} = (E_c + E_v)/2 + kT/2 \cdot \ln(c_p N_v/(c_n N_c g^2))$$

which is not far from the bandgap middle. The generation rate can be written $G = G_{\max}/\cosh((E_r - E_{rm})/kT)$. A few kT away from $E_{r,m}$, G is proportional to the smaller one of the emission rates and hence decreases exponentially with increasing distance of the level from the more distant band. If short switching times of a device are required, but on the other side also a small generation in the space charge region and hence low blocking current, this can be reached by choosing a deep impurity whose recombination level is well distant from the middle of the bandgap, but not *too* close to one of the band edges.

Although the density of recombination centers is mostly significantly smaller than the doping of the weakly doped base region, this may be so only to a less extent in very fast devices. Here the charge on the traps can have undesirable effects, such as a compensation of the normal doping (reduction of the conductivity), a reduction of the breakdown voltage or a premature punch-through of the space charge region. The influence of the traps on the carrier concentration in thermal equilibrium depends upon the position of their level relative to the Fermi energy and likewise on the type of the level. For example an acceptor-like impurity in a neutral n base with a level a few kT above the Fermi energy will be neutral to a large part and hence will not affect the free electron concentration essentially. A donor trap with a level at the same position on the other hand is positively charged and leads to an *increase* of the electron concentration in a neutral n region. In a space charge region of a reverse biased pn-junction, the charge in steady state is given by the condition that the generation rate of electrons must equal to that of holes. For an acceptor level one has $e_n N_r^- = e_p N_r^0$, hence $N_r^-/N_r^0 = e_p/e_n$. To obtain a low concentration of charged deep acceptors (e.g. to avoid punch-through), a higher emission rate e_n than e_p is required. Generally, the density of charged traps in a stationary non-equilibrium state is given by (2.67).

During switching processes, the densities N_r^-, N_r^0 follow the variation of n and p not instantaneously but often rather slowly. Their time-dependence is described by the equation

$$\frac{dN_r^-}{dt} = c_n n N_r^0 - c_p p N_r^- - e_n N_r^- + e_p N_r^0 = -\frac{dN_r^0}{dt} \tag{2.78}$$

which follows immediately from the capture and emission events recharging the impurity. Together with differential equations for n and p following from (2.60), (2.62) and (2.48) the charged trap density as a function of time can be numerically calculated. Some effects of the temporary charge on deep impurities will come up in Sect. 13.3.

In above examples the distances of the acceptor level from the conduction band and the donor level from the valence band have been assumed constant calculating the concentrations n_a and p_d. Since the bandgap decreases with increasing temperature, at least one of the activation energies of a level, $\Delta E_n = E_c - E_r$ or $\Delta E_p = E_r - E_v$, varies also, mostly however both vary with T. This has a strong effect on the recombination. The variation of ΔE manifests itself in an 'effective' degeneracy factor g' which usually strongly deviates from the values 2 or 4 of the 'intrinsic' degeneracy factor g introduced in Sect. 2.5. As pointed out in Appendix B, the concentrations n_r, p_r can be calculated replacing in (2.61a) respectively (2.63a) the real activation energy $\Delta E(T)$ by a constant apparent experimental activation energy $\Delta E'$ and the degeneracy factor g by the effective degeneracy factor g′. In contrast to g, *the degeneracy factor* g'_n *for* n_r *is different from the degeneracy factor* g'_p *of* p_r. Written in full the concentration $n_r = n_a$ of an acceptor level and $p_r = p_d$ of a donor level can be calculated from

$$n_a = N_c g'_{n,a} \exp(-\Delta E'_{n,a}/kT) \tag{2.79a}$$

$$p_d = N_v g'_{p,d} \exp(-\Delta E'_{p,d}/kT) \tag{2.79b}$$

p_a and n_d are then given as n_i^2/n_a respectively n_i^2/p_d. In Appendix B it is shown how the degeneracy factors g' are made up and can be evaluated from the variation of a level position with temperature.

Actually all deep impurities in silicon used for lifetime reduction possess more than one level in the bandgap. The SRH model for two levels and its application to gold and platinum will be treated in following sections in some detail. A part of the exposition is possibly helpful also for a better understanding of recombination centers introduced by radiation.

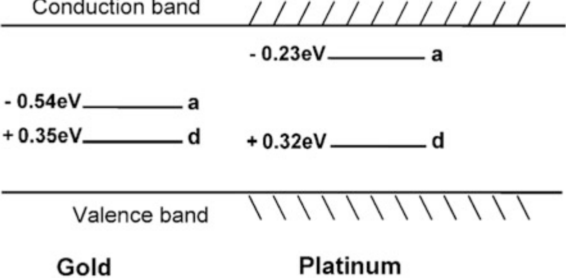


Fig. 2.19 Energy levels of gold and platinum in silicon. Numbers with a plus sign indicate the distance of the level from the valence band edge, numbers with a minus sign the distance from the conduction band. Donor levels are labeled with "d" and acceptor levels with "a". The specified values neglect their temperature dependence

(b) Recombination centers with two levels: Au, Pt and several other deep impurities in Si possess a donor level in the lower half of the bandgap and an acceptor level in the upper half [Pal74] (see Fig. 2.19). The two levels arise because these centers have three charge states, whose neighboring ones can pass into one another. The effective concentrations of the levels are coupled with another and vary with Fermi level, injection level and temperature. Equations describing this have been developed in several papers [Sah58, Kon59, Scz66, Abb84]. In the present section a general lifetime equation for two-level traps is introduced, a numerical example and significant limiting cases are discussed. The notation is adapted to Au and Pt, centers with a donor and an acceptor level.

Recombination centers of this type can be positively charged (R^+), neutral (R^0) or negatively charged (R^-). According to the definition given above, the donor level indicates the energies involved in transitions between a band and the impurity if the charge state changes between R^+ in R^0, the acceptor level the energies between the bands and the impurity for changes between R^0 and R^-. As in the case of one level, the transition can occur in each case by capture or emission of an electron from resp. to the conduction band and a similar exchange of a hole with the valence band (see Fig. 2.17). Because of the electrostatic energy gained during the capture of a conduction electron by R^+ but not by R^0, the donor level lies (generally) below the acceptor level. This differs from the familiar position of donor and acceptor levels of normal doping atoms (see Fig. 2.5).

To obtain a formula for the lifetime the total recombination rate is expressed as the sum of two SRH terms (2.68), one for each level. The trap density N_r is to be replaced in both cases by the concentration of occupied plus unoccupied states, hence for the donor level by $N_d = N_r^+ + N_r^0$ and for the acceptor level by $N_a = N_r^0 + N_r^-$. These concentrations are coupled via the constant total concentration $N_r = N_r^+ + N_r^0 + N_r^-$ and depend on n and p as is obtained in steady state setting $R_n = R_p$ for each level. The formulas can be found widely in [Kon59, Scz66, Abb84]. From the total recombination rate and the definition Eq. (2.49) one gets for the hole lifetime in an n-type semiconductor with $n_0 >> p_0$:

$$\tau_p^{(n-Si)} = \frac{1}{N_r c_{p,a}} \frac{\left(n + n_a + \frac{c_{p,a}}{c_{n,a}}(p_a + p)\right)\left(n + \frac{c_{p,d}}{c_{n,d}} p_d\right) + \left(n_a + \frac{c_{p,a}}{c_{n,a}} p\right)\left(n_d + \frac{c_{p,d}}{c_{n,d}} p\right)}{n c_{p,d}\left(\frac{n}{c_{p,d}} + \frac{n_a}{c_{p,a}} + \frac{p}{c_{n,a}} + \frac{p_d}{c_{n,d}}\right)} \tag{2.80}$$

n_a and p_a denote the concentrations n_r, p_r of the acceptor level introduced by Eqs. (2.61a), (2.63a), p_d and n_d the concentrations p_r, n_r of the donor level defined by these equations in the donor version where N_{r0}^-, N_{r0}^0 have to be replaced by N_{r0}^0 and N_{r0}^+, respectively, and g by 1/g. Actually, Eq. (2.79a, b) will be used often to calculate n_a, and p_d from the constant apparent activation energies $\Delta E'_{n,a}$, $\Delta E'_{p,d}$ and the effective degeneracy factors $g'_{n,a}$, $g'_{p,d}$. Mostly, if the acceptor level is well in the

upper half and the donor level in the lower half of the bandgap, the concentrations p_a and n_d are negligible. The electron concentration in a neutral n region is given by the hole density according to $n = N_D + p + N_r^+ - N_r^-$, where the concentration $N_r^+ - N_r^-$ representing the charge of the traps can be neglected often in a first approximation. In our calculations it is taken into account using formula (B17) given in Appendix B.2. The electron lifetime in a p-type region follows from (2.80) by interchanging n and p in letters and indices as well as indices a and d.

In Fig. 2.18 the lifetime following from (2.80) for a trap with the shown levels and capture rates is plotted as solid curve. The distances of the donor level from the valence band and of the acceptor level from the conduction band were assumed temperature independent. The same capture rates and degeneracy factors are used as for the above discussed traps which have only one of the two levels (dashed curves). As is seen the lifetime of the two-level center is up to about $p = 4\ 10^{14}$ cm^{-3} the same as if only the donor level would be present. The lifetime is strongly reduced in this range by the donor level. Responsible for this are the equilibrium concentration n_0 ($\approx 9 \times 10^{13}$ cm^{-3}) and the concentrations p_d (1.18×10^{13} cm^{-3}) and n_a which latter with 7.22×10^{15} cm^{-3} is two orders of magnitude higher than p_d and n_0. Under these conditions and for relatively low injection levels with $n_0 + p << n_a$ Eq. (2.80) reduces to

$$\tau_p = \frac{1}{N_r}\left(\frac{1}{c_{p,d}} + \frac{1}{c_{n,d}}\frac{p_d + p}{n}\right) \quad (\text{for} \quad n, p_d \ll n_a) \tag{2.81}$$

In the precondition for this equationratios of capture coefficients have been ignored (set equal to 1) for simplicity. In words we have obtained: If the donor level is more distant from the valence band than the acceptor level from the conduction band and the Fermi level in thermal equilibrium below the acceptor level, the hole lifetime in the range $n_0 + p <<$ n$_a$ is determined by the donor level. Equation (2.81) is identical with (2.71) with n_r neglected and $p_r \equiv p_d$. The unfavorable consequences of the near band position of the acceptor level are eliminated by the donor level. The trap atoms are mostly in a charge state (neutral) in which they cannot react via the acceptor level with minority carriers.

If with increasing injection level the precondition n << n_a of (2.81) becomes invalid, the acceptor level too influences the lifetime. Contrary to the case of one level this can lead to a maximum of the $\tau_p(p)$-function as shown by Fig. 2.18. The limiting value of the high-level lifetime for $p >> n_0$, p_0, n_a, p_d is obtained from (2.80) as

$$\tau_{HL} = \frac{1}{N_r} \cdot \frac{c_{n,a}/c_{p,a} + 1 + c_{p,d}/c_{n,d}}{c_{n,a} + c_{p,d}} \tag{2.82}$$

In the case of Fig. 2.18 τ_{HL} is higher than for the acceptor level alone although both levels contribute to it.

A simple limiting case appears if the Fermi level is well above the acceptor level so that $n_0 \gg n_a$. Then the low-level hole lifetime is obtained from (2.80) as

$\tau_{p,LL} = 1/\left(N_r c_{p,a}\right) \equiv \tau_{p0}^{(a)}$. Analogously, the low-level electron lifetime in p-silicon with $p_0 >> p_d$ is given by $\tau_{n,LL}^{(p-Si)} = 1/\left(N_r c_{n,d}\right) \equiv \tau_{n0}^{(d)}$ as follows from the equation for τ_n in p-Si obtained from (2.80) in the mentioned way. These cases are used to determine the capture probabilities $c_{p,a}$ and $c_{n,d}$.

Finally we consider a few limiting cases which we will meet again below. If under the condition $n_a >> p_d$ also the concentration n_0 is large against p_d, Eq. (2.80) yields for the low-level hole lifetime in an n-region:

$$\tau_{p,LL} = \frac{1}{N_r} \frac{1 + n_a/n_0}{c_{p,a} + c_{p,d} n_a/n_0} \quad (\text{for } \mathrm{n_0, n_a} \gg \mathrm{p_d}) \tag{2.83}$$

Here the low-level lifetime is influenced by both levels. Compared with (2.81) the temperature dependence is weaker, because the denominator counteracts the T-dependence of the nominator and, furthermore, the concentration n_a is less temperature dependent than p_d due to the smaller distance from the relevant band.

If on the other hand $n_a << p_d$, as is the case of gold, the low-level minority carrier lifetime is obtained from Eqs. (2.79a, b) *for arbitrary doping concentrations* as

$$\tau_{p,LL} = \frac{1}{N_r c_{p,a}} \left(1 + \frac{n_a}{n_0}\right) \tag{2.84}$$

where the precondition is strictly speaking $n_a \ll c_{p.a}/c_{n,d}\, p_d$. Equation (2.84) is identical with Eq. (2.72) applied to the acceptor level.

At higher temperatures and low doping both p_d and n_a are large compared with the equilibrium concentration n_0. Then one obtains from Eq. (2.80) for the low-level hole lifetime in n-Si:

$$N_r \tau_{p,LL} = \left\{\left(c_{p,a}/n_a + c_{n,d}/p_d\right) n_0\right\}^{-1} \quad \text{for } n_a, p_d \gg n_0 \tag{2.85}$$

This is the extension of (2.76) to the present two-level case. The lifetime in this range is inversely proportional to n_0. The same expression, with n_0 replaced by p_0, holds for the minority carrier lifetime in p-silicon. Hence as for one level the low-level minority carrier lifetimes in n and p silicon are equal, if doping as well as trap concentrations are equal in both cases.

(c) The recombination center gold: Gold is the most investigated deep impurity in silicon, although disagreements between the parameter values obtained remain. The two-level model as due to three charge states of the same impurity has been called into question [Vec76, Lan80], but was reconfirmed by other authors [LuN87]. The levels are shown in Fig. 2.19 with rough numerical positions neglecting the temperature dependence. The acceptor level is located only a little above the middle of the bandgap, which has the obvious drawback of causing a high generation current in reverse biased junctions. Although gold has lost therefore a part of its

applications to platinum and radiation induced traps, it is still applied, because in other respects it has advantages. Gold attracts attention also as a kind of model trap.

Since according to the position of the levels $n_a << p_d$, the low-level hole lifetime in n-Si is given by Eq. (2.84). Hence *the low-level minority carrier lifetime in n-Si as function of doping density and temperature is determined by the acceptor level*. For *arbitrary* injection levels one obtains from Eq. (2.80) using $n_a \ll p_d$

$$\tau_p = \frac{1}{N_r c_{p,a}} \frac{\left(n + n_a + \frac{c_{p,a}}{c_{n,a}} p\right)\left(n + \frac{c_{p,d}}{c_{n,d}} p_d\right) + \left(n_a + \frac{c_{p,a}}{c_{n,a}} p\right) \frac{c_{p,d}}{c_{n,d}} p}{n\, c_{p,d} \left(\frac{n}{c_{p,d}} + \frac{p_d}{c_{n,d}} + \frac{p}{c_{n,a}}\right)} \tag{2.86}$$

where also n_d and furthermore $c_{p,a}/c_{n,a}\, p_a$ against n_a are neglected. Although the acceptor level lies near the middle of the bandgap, the latter neglect is justified up to 450 K, because the degeneracy factor $g'_{n,a}$ of the acceptor level turns out to be high.

A complete set of capture rates have been determined by Fairfield and Gokhale [Fai65] and later by Wu and Peaker [WuP82] who determined also their temperature dependences and listed results published till then. Because these and later results scatter strongly and are not sufficiently in agreement with lifetime measurements, we adjust the parameters first as far as necessary to the following lifetime properties of gold:

I. The hole lifetime in an n region at room temperature increases between low and high injection level by a factor of about 5. In [Fai65] the ratio of high-level to low-level was obtained to be 6.3, Zimmermann measured a ratio 4.0 [Zim73] and Hangleiter 5.5 [Han87]. The doping concentration in [Zim73] was 1×10^{14} cm^{-3}, in [Fai65] and [Han87] higher doped devices with $N_D >> p_d = 1.31 \times 10^{14}$ cm^{-3} were used.
II. The electron lifetime in a p-region with a doping concentration $N_A >> p_d$ is independent of the injection level [Han87].
III. Both the high-level lifetime τ_{HL} and the low-level electron lifetime in a p-region with $N_A \gg p_d$ are proportional to T^2 over a wide range ($\tau_{n,LL}$ from 100 to 300 K [Scm82], τ_{HL} from 100 to 400 K [Sco76] (approximately confirmed in [Han87]).
IV. The electron emission rate of the gold acceptor level is an order of magnitude higher than the hole emission rate [Sah69, Eng75]. According to the detailed balance Eqs. (2.61), (2.63) this has to be taken into account for the choice of capture probabilities and degeneracy factor.

The capture rates of [Fai65] satisfy criterion I, but not II, the data of [WuP82] result in an unrealistic high-level to low-level ratio of 25.6. As is explained in detail in Appendix B, the above criteria together with published work lead us to the capture rates and degeneracy factors shown in Table 2.3. The 300 K-value of $c_{p,a}$ is adopted from [Fai65]. Using this and the emission rates of Sah et al. [Sah69] the value of $c_{n,a}$ is determined from the detailed balance Eq. (2.64). In agreement with [WuP82] the capture rates $c_{n,a}$ and $c_{n,d}$ are well an order of magnitude smaller than

the corresponding capture probabilities of holes. Hence according to Eq. (2.82) the high-level lifetime is determined essentially by the capture probability $c_{n,d}$ of the donor level. The increase of the lifetime between low and high level by a factor 5 to 6 quoted in criterion I, requires hence that $c_{n,d}$ is by about the same factor smaller than $c_{p,a}$, the capture rate determining the low-level lifetime in n-Si at 300 K. The temperature dependences are taken from literature considering criterion III. According to Eq. (B9) the effective degeneracy factor $g'_{n,a} = 10.9$ means, that the distance of the acceptor level from the conduction band decreases with temperature with the rate 0.795×10^{-4} eV/K. This is 29% of the rate of the bandgap decrease at 350 K, if calculated from Eq. (2.9) together with the parameters of Table 2.1.

Table 2.3 Capture probabilities, degeneracy factors and characteristic concentrations of the gold recombination center

Acceptor level	Donor level	Unit
$c_{n,a} = 0.56(T/300\,\mathrm{K})^{0.5}$	$c_{n,d} = 2.32(T/300\,\mathrm{K})^{-2}$	10^{-8} cm^3/s
$c_{p,a} = 11.5(T/300\,\mathrm{K})^{-1.7}$	$c_{p,d} = 28.0(T/300\,\mathrm{K})^{0.5}$	10^{-8} cm^{-3}/s
$\Delta E'_{n,a} = 0.5472$	$\Delta E'_{p,d} = 0.3450$	eV
$g'_{n,a} = 10.9$	$g'_{p,d} = 2.65$	
$n_a(300\,\mathrm{K}) = 2.00 \times 10^{11}$	$p_d(300\,\mathrm{K}) = 1.31 \times 10^{14}$	cm^{-3}
$n_\{a\}(400\,\mathrm{K}) = 6.26 \times 10^{13}$	$p_d(400\,\mathrm{K}) = 6.29 \times 10^{15}$	cm^{-3}

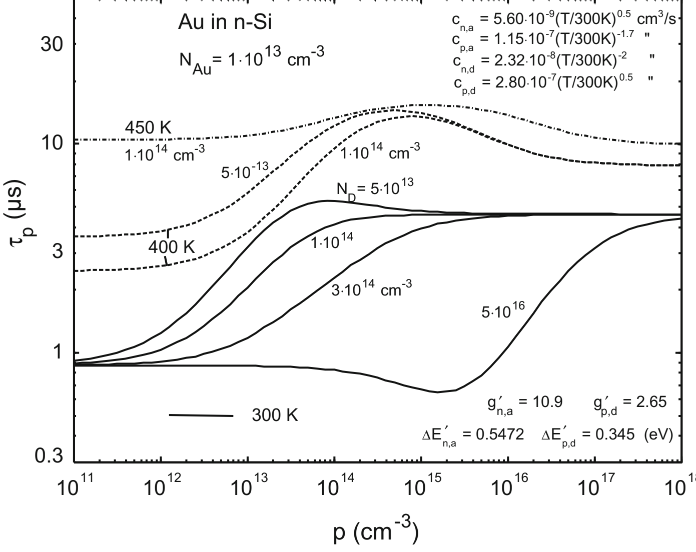


Fig. 2.20 Hole lifetime for the gold recombination center in n-type silicon in dependence of the injected hole concentration. Parameters are the doping concentration and temperature

With $g'_{p,d} = 2.65$ the distance of the donor level from the valence band is obtained to be nearly constant.
Using these data in Eq. (2.86) the calculated injection dependence of the hole lifetime in n-silicon is shown in Fig. 2.20 for several doping concentrations and temperatures. The dependence on doping concentration and temperature is strong. Owing to the choice of the capture parameters the lifetime at 300 K increases by a factor of about 5. Also the increase of the high-level lifetime from 300 to 450 K accords with the T^2 dependency (criterion III). The minimum of the $\tau_p(p)$ curve for $N_D = 5 \times 10^{16}$ cm^{-3} is not indicated by measurements (see [Fai65, Han87]), but since only weak it was possible not observable. Because of the good over-all agreement with measurements the minimum seems to be a necessary consequence of the theoretical model together with the criteria used for adjustment.

Since the criteria are reproduced, we can have some trust also in the other results. In Fig. 2.20 the $\tau_p(p)$-functions at 400 and 450 K show a maximum, as already seen in the case of Fig. 2.18. The acceptor level alone would yield an increase of the lifetime over the whole range, not only at 300 but also at 400 and 450 K, since the condition (2.75) is well satisfied. The shown decrease after the maximum at about $p = 5 \times 10^{14}$ cm^{-3} is caused hence by recombination via the donor level. As mentioned, the high-level lifetime is (nearly) solely determined by the donor level since $c_{n,a} \ll c_{p,a}$ and $c_{n,d} \ll c_{p,d}$, whereas the low-level lifetime by the acceptor level. The ratio of high-level to low-level lifetime for $N_D = 1 \times 10^{14}$ cm^{-3} is 3.0 at 400 K and about 1 at 450 K. *That the low-level lifetime is even at elevated temperatures not higher than the high-level lifetime is a significant advantage of gold* [see the discussion in section a)]. Since the concentration n_a in (2.84) is small up to nearly 400 K (see Table 2.3), the low-level lifetime shows a relatively weak temperature dependence compared for example with platinum (see section d).

The curves for $N_D = 5 \times 10^{16}$ and 1×10^{14} cm^{-3} show that in device regions with higher n-doping (5×10^{16} cm^{-3}) used as 'field stop layers' the lifetime is significantly smaller than in the weakly doped base region for a given absolute injection level (constant gold density assumed). From Eq. (B14) in Appendix B it follows that the small lifetime over a wide injection range follows from the high coefficient $c_{p,d}/c_{n,a}$ in the denominator, hence is caused by recombination via the donor level. Only at very low injection levels the lifetime at room temperature is independent of N_D and given by $\tau_{p,LL} = 1/\left(c_{p,a}N_r\right)$.

Also the minority carrier **lifetime in p-Si** is of interest for power devices, for example to simulate the recombination behavior of p-regions in thyristors and IGBTs. As stated in criterion II, at doping concentrations $N_A \gg p_d$, the case of the p-base of thyristors, the lifetime $\tau_n^{(p-Si)}$ is independent of the injection level. Rewriting Eq. (2.80) into the $\tau_n^{(p-Si)}$ version and using that $c_{n,a} \ll (c_{p,a},\, c_{p,d})$ and $c_{n,d} \ll c_{p,d}$ a simple approximation results under the condition $p_0 \gg c_{n,a}/c_{p,a}n_a$, which down to $N_A = 1 \times 10^{14}$ cm^{-3} and up to 400 K is well satisfied. The electron lifetime as function of injection level is then approximately given by:

$$\tau_n^{(p-Si)} \approx \frac{1}{N_r} \frac{1 + p/p_d}{c_{n,a} + c_{n,d} p/p_d} \tag{2.87}$$

Because $c_{n,a} < c_{n,d}$, the lifetime in p-silicon decreases with increasing injection $n \approx p - p_0$. For $p = p_0$ Eq. (2.87) is the counterpart to (2.83). The temperature dependence is only moderate, since the two p_d terms counteract each other. As is noted also, the low-level lifetime at a doping density of 1×10^{14} cm^{-3} is by a factor of about 9 higher than in n-silicon at room temperature, if the gold concentration is equal.

Of interest for devices are furthermore the concentrations of the different charge states under different conditions. In an *n region in thermal equilibrium* the gold atoms are completely in the negative charge state, because the Fermi level is located well above the acceptor level. The equilibrium free electron concentration is reduced to $n_0 = N_D - N_{Au}$. Hence gold leads to an increase of the resistivity (compensation). At *high injection levels* with $n = p \gg n_a, p_a, p_d$, however, one obtains from equations (B13) to (B15) in Appendix B.2:

$$N_{Au}^{+} : N_{Au}^{0} : N_{Au}^{-} = \frac{c_{p,d}}{c_{n,d}} : 1 : \frac{c_{n,a}}{c_{p,a}} = 12.1 : 1 : 0.049$$

where for the numerical values the capture rates at 300 K were used (see Table 2.3). Hence at high injection gold is largely in the positive charge state, $N_{Au}^{+} \approx N_{Au}$. In a *space charge region* under reverse bias on the other hand the generation rate in steady state $G = G_a = e_{n,a} N_{Au}^{-} = e_{p,a} N_{Au}^{0}$ yields $N_{Au}^{-} = e_{p,a}/e_{n,a} N_{Au}^{0} \ll N_{Au}^{0} \approx N_{Au}$. Because $e_{p,a}$ is more than an order of magnitude smaller than $e_{n,a}$, the gold centers in a space charge region are mostly neutral. The positive trap charge at high injection becomes noticeable during switching of a fast pin diode from forward to reverse bias, the negative charge in the n base in equilibrium during switching from zero to reverse bias. In the space charge region the Au^{+} and Au^{-} ions turn into the neutral state by emission of a hole respectively an electron. Because of the long time constant $1/e_{n,a}$ ($\approx$ 1 ms at room temperature) the decay of the concentration N_{Au}^{-} is rather slow.

(d) The recombination center platinum. The levels of platinum in silicon have been determined by a series of investigators [Con71, Pal74, Mil76a, Sue94]. With some spreading they obtain a donor level 0.32 eV above the valence band and an acceptor level 0.23 eV below the conduction band as shown in Fig. 2.19. Some authors found additional levels attributed partly to a second Pt site present to a certain share of the total concentration [Lis75, Evw76, Sue94]. An acceptor level at 0.42 eV above the valence band reported as the dominant Pt recombination level [Lis75] has not been observed however by later researchers [Bro82b]. Since accepted by most, the level scheme of Fig. 2.19 is assumed in what follows. The

Table 2.4 Capture cross sections of platinum in silicon determined in [Con71] and [Bro82a]. Also the exponents n of the power law temperature dependences $\sigma_{ij} \sim T^n$ and the effective degeneracy factors of [Con71] are given

	[Con71] (295 K)		[Bro82a] (77 K)	
	σ (10^{-16} cm^2)	n	σ (10^{-16} cm^2)	n
$\sigma_{p,a}$	55	−4.4	30	0
$\sigma_{n,a}$	63	0	8	0
$\sigma_{p,d}$	16.2	0	3	0
$\sigma_{n,d}$	18.8	−4.0	1	–
$g'_{n,a}$	1 ± 0.5			
$g'_{p,d}$	25 ± 8			

energy values are understood as apparent experimental activation energies $\Delta E'_{p,d}$, $\Delta E'_{n,a}$, to which a temperature dependence according to the observed effective degeneracy factors is superposed (see Appendix B and Eqs. 2.79a, b).

Capture probabilities of the two levels have been determined by Conti and Panchieri [Con71] and Brotherton and Bradley [Bro82a], who former made their measurements near room temperature, the latter around 77 K. The results are depicted in Table 2.4 in form of capture *cross sections* given in the papers. For conversion to capture probabilities the cross sections were multiplied with the following expressions for the mean thermal velocity of electrons respectively holes:

$$v_{th,n} = 2.25 \times 10^7 \sqrt{T/300K} \text{ cm/s}$$

$$v_{th,p} = 1.85 \times 10^7 \sqrt{T/300K} \text{ cm/s}$$

The constants are obtained using in Eq. (2.28) the effective masses $m_n = 0.27\ m_e$, $m_p = 0.4\ m_e$. In the table also the exponents n of the power law temperature dependencies $\sigma_{i,j} \sim T^n$ are shown, as far as determined. As is seen, the results are again very different. Whereas the cross section $\sigma_{p,a}$ according to [Con71] decreases with temperature as $1/T^{4.4}$, in [Bro82a] it is found to be constant. The degeneracy factors of [Con71] differ essentially from the intrinsic values, hence they indicate a significant temperature dependence of the distance of the levels from the respective band.

The injection dependence of the lifetime in n-type silicon calculated with the data of [Con71] shows at low doping a simple monotonous decrease as expected from measurements [Mil76b]. Against this the cross sections of [Bro82a], completed with reasonable degeneracy factors and the temperature dependence of $c_{n,d}$, result in a pronounced intermediate minimum in the $\tau_p(p)$-curves. Since the results of [Con71] are also more complete, they are taken in what follows as a basis.

A series of lifetime measurements concerning the dependence on temperature and doping have been published by Miller and coworkers [Mil82a]. Besides others they observed that the hole lifetime in n-Si with $N_D = 2.6 \times 10^{14}$ cm^{-3} increases from 300 to 400 K by a factor 10 and for a doping concentration of 2.9×10^{15} cm^{-3} by a factor 4.3. As function of doping density an inverse proportionality, $\tau_p \sim 1/n_0$, was

found at 400 K, at room temperature the variation with n_0 is approximately halved. The inverse proportionality agrees with Eq. (2.85), showing that the condition $n_a, p_d \gg n_0$ must be satisfied at 400 K up to $n_0 = 2.9 \times 10^{15}$ cm^{-3}. The data of [Con71] leading at 400 K to $n_a = 5.70 \times 10^{16}$ cm^{-3} and $p_d = 1.23 \times 10^{17}$ cm^{-3} fulfill this condition.

Because these measurements were carried out apparently at low injection levels, they have to be described in the SRH model by the equation

$$\tau_{p,LL} = \frac{1}{N_r c_{p,a}} \frac{(n_0 + n_a)\left(n_0 + \frac{c_{p,d}}{c_{n,d}} p_d\right)}{n_0\left(n_0 + \frac{c_{p,d}}{c_{p,a}} n_a + \frac{c_{p,d}}{c_{n,d}} p_d\right)}, \tag{2.88}$$

which follows for low injection from (2.80). Using the parameter set of [Con71] the calculated temperature dependence is 5–6 times stronger than measured, whereas the doping dependence is nearly as observed. For a better approach to the measurements it is necessary to lower the temperature dependences of $\sigma_{p,a}$ and $\sigma_{n,d}$. This is justified by the constant cross section $\sigma_{p,a}$ of [Bro82a] as well as by measurements on other deep impurities [Han87], which all show significantly smaller absolute n-values for attractive charge states than the values of [Con71]. Taking into account that the high-level lifetime τ_{HL} will probably increase with temperature, suitable values to reflect the present state of knowledge are $n = -1$ for $\sigma_{p,a}$ and -2.5 for $\sigma_{n,d}$. These values are used in the calculations presented below. With this modification the data of [Con71] result in a lifetime increase with temperature which is up to a factor 2.5 stronger than measured, whereas the doping dependence is again approximately reproduced.

Of course the temperature dependence is essentially influenced by the exponential dependencies of n_a and p_d on temperature. For given apparent activation energies, this can be changed via the effective degeneracy factors, see Eqs. (2.79a, b). Without violating the condition for the reciprocal dependency on n_0, the value of $g'_{p,d}$ could be lowered by a factor 10. This would enhance however the temperature dependence, because the condition $n_a \gg p_d$ holding then would lead to the approximation (2.81) characterized by the strong T-dependence of p_d and also $c_{n,d}$. With the choice of [Con71] the smaller temperature dependences of n_a and $c_{p,a}$ are effective in addition to p_d. Compared with this the approach can be improved somewhat reducing the value of $g'_{n,a}$ to 0.5. With this value the calculated temperature dependence is a factor 1.7 stronger than measured, instead of 2.5 for $g_{n,a'} = 1$. This difference is acceptable considering the significant error sources which lifetime measurements are usually subjected. In the present case possibly a less T-dependent recombination part at the junctions may have influenced the results.

In Fig. 2.21 the injection dependence of the hole lifetime in n-silicon calculated from Eq. (2.80) with the proposed data set is shown for some doping densities and temperatures. The data are depicted. The accordance with the measurements used for adjustment can be checked at the curves indicated with '2' and '3'. Between

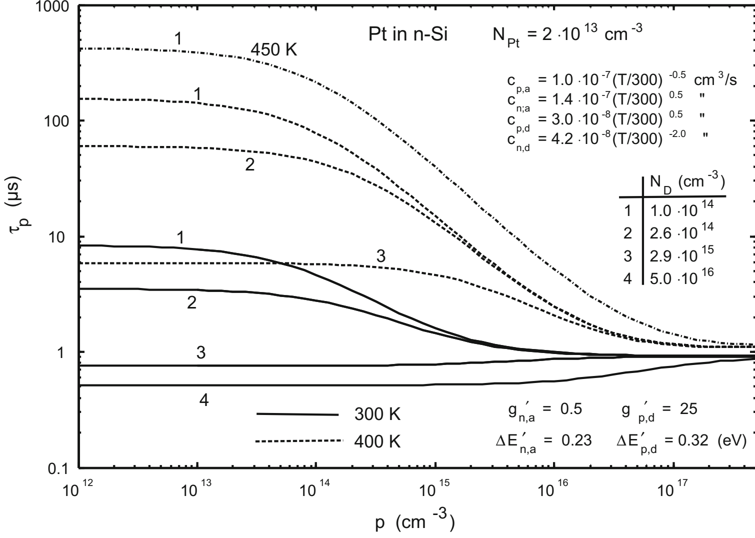


Fig. 2.21 Minority carrier lifetime for platinum in n-Si as function of injected hole concentration. Parameters are the doping concentration and temperature. The capture rates at 300 K are taken from [Con71]. The temperature dependences of $c_{p,a}$ and $c_{n,d}$ and the degeneracy factor of the acceptor level are reduced compared with [Con71] (see the text)

1×10^{14} and 2.6×10^{14} cm^{-3} the low-level lifetime is even at 300 K approximately inversely proportional to the doping concentration, because the condition $n_0 \ll n_a, p_d$ is satisfied in this doping range even at room temperature. At $N_D = 1 \times 10^{14}$ cm^{-3} the lifetime at 300 K decreases with injection level by a factor 9 and at 400 K by a factor 140. From 300 to 400 K the low-level lifetime for $N_D = 1 \times 10^{14}$ cm^{-3} increases by a factor 18.

The disadvantages of the strong decrease of the lifetime with injection level and of the increase with temperature have been pointed out on page 67. The latter property results in a decrease of the forward voltage drop of Pt diffused diodes with increasing temperature [Lut00], an unfavorable property for paralleling devices in modules. An example for the temperature dependence of the forward characteristic of a platinum diffused diode will be given later in Chap. 5 Fig. 5.11. The strong enhancement of lifetime with decreasing injection leads to a somewhat higher recovery time of platinum-diffused devices than in the case of gold. Since especially for thyristors and IGBTs the high low-level lifetime is disadvantageous, platinum is normally not used in fast thyristors and IGBTs.

For diodes these disadvantages associated with the position of the levels are overcompensated for most applications by the small generation current in the space charge region. In steady state the generation rate of the acceptor level is $G_a = e_{n,a}N_{Pt}^- = e_{p,a}N_{Pt}^0$, that of the donor level $G_d = e_{n,d}N_{Pt}^0 = e_{p,d}N_{Pt}^+$. Since $e_{p,a} \ll e_{n,a}$, $e_{n,d} \ll e_{p,d}$, it follows that $N_{Pt}^-, N_{Pt}^+ \ll N_{Pt}^0 \approx N_{Pt}$: platinum in the space charge region is neutral in steady state, similarly as gold. The total generation rate is

$$G = G_a + G_d = (e_{p,a} + e_{n,d})N_{Pt}$$
$$= \left(\frac{c_{p,a}}{n_a} + \frac{c_{n,d}}{p_d}\right) n_i^2 N_{Pt}$$

In our dataset with the large g'_{pd} one has $p_d > n_a$, so that the main part of G is contributed by the acceptor level. In agreement with measurements one obtains that the leakage current of Pt diffused diodes is even at 150 °C more than a decade smaller than obtained with gold. Therefore platinum diffusion or particle irradiation is now used in most fast diodes for reduction of carrier lifetime.

In contrast to gold platinum centers are also neutral in weakly doped n base regions in thermal equilibrium because the Fermi level is below the acceptor level. Hence the conductivity is not considerably reduced, an advantage if the diode is integrated together with a unipolar switching element. Furthermore switching of a diode from zero to reverse voltage is not interfered with recharging of platinum ions. At high injection levels the concentrations of Pt in the different charge states are obtained from equations from Appendix B.2 at 300 K as

$$N_{Pt}^+ : N_{Pt}^0 : N_{Pt}^- = \frac{c_{p,d}}{c_{n,d}} : 1 : \frac{c_{n,a}}{c_{p,a}} = 0.71 : 1 : 1.40$$

All charge states appear in significant fractions, although the negative predominates. The *net* concentration of negative charge, $N_{Pt}^- - N_{Pt}^+$, amounts to 22% of the total Pt concentration. The platinum charge involved in the switching process from forward to reverse bias is hence of the opposite sign and smaller than in the case of gold. These statements are subjected to the same errors which refer to the capture rates.

(e) Radiation induced recombination centers. Often irradiation with high-energy electrons, protons or He ions is used nowadays to generate recombination centers and reduce the carrier lifetime. The energy of irradiation is usually in the range 1 to 15 meV. While electron irradiation produces a homogeneous density of recombination centers, a narrow region of high trap concentration can be created with H and He ion irradiation. The radiation methods show a good reproducibility. As depicted in Fig. 2.22, mainly three independent centers with various levels are generated: the divacancy (VV), the A-center which is an oxygen-vacancy complex (OV), and the

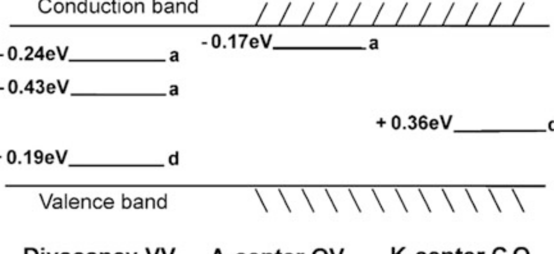


Fig. 2.22 Energy levels of some important radiation induced centers

K-center, probably an association of an interstitial carbon atom and an interstitial oxygen atom (C_iO_i) [Niw08]. The type of the levels and distances from band edges are indicated as in Fig. 2.22. The divacancy has a donor level and two acceptor levels, the upper of which refers to transitions between R^- and R^{2-}. In contrast, the A-center has a single acceptor level near the conduction band. The relative concentrations of the centers depend on the radiation energy as well as on tempering processes following the irradiation. For lifetime control in neutral regions, especially at high injection level, the A-center is considered to be most efficient because of high capture rates and high concentrations [Sie02, Sie06]. The carrier generation in space charge regions is determined by the divacancy whose level at 0.43 eV below the conduction band is most effective. Owing to this energy level, the lifetime control by radiation results in a significantly lower blocking current than obtained using gold, but it is higher than generated by platinum. The possibilities to calculate the lifetime behavior in the case of radiation induced centers are still very limited. More details on the radiation technique for lifetime control are given in Sect. 4.9.

2.8 Impact Ionization

Impact ionization is a mechanism of carrier generation which takes place at high electric fields and leads to avalanche multiplication. This generation limits the electric field which a blocking junction can sustain without producing a high current, hence it is fundamental for dimensioning of power devices. It determines the breakdown voltage in the whole range above about 6 V in silicon. At the lower end, below about 12 V, where the extension of the high-field region is small enough, quantum mechanical tunneling of carriers becomes significant for the current, and below about 6 V it is decisive for the usable blocking voltage [Hur92a, Hur92b]. Although many power devices contain low-blocking junctions, tunneling is not of high interest for them. The “critical” field strength which leads to avalanche breakdown is a property of the semiconductor. Via the width of the base region required for a given blocking voltage it determines static and dynamic limits of devices. In the present section, we deal with general features of impact ionization, the detailed consequences for power devices will be treated in later sections.

Impact ionization occurs if the electric field is high enough, such that a noticeable number of electrons or holes in the statistical distribution gain sufficient kinetic energy that they can lift a valence electron by impact into the conduction band. Each ionizing carrier generates a pair of a free electron and hole, which again can generate further electron-hole pairs, thus giving rise to an avalanche event. Impact and avalanche generation are therefore used synonymously. Because also the kinetic energy of the secondary particles as following from momentum conservation has to be provided, the ionizing energy which the primary carrier at least

must have is about 3/2 E_g [Moe69]. A few carriers in the thermal distribution have this energy already at zero field strength. However, the generation at zero field is taken into account already as Auger generation, $G_{Aug} = (C_{A,n}n + C_{A,p}p)n_i^2$, whose microscopic mechanism is identical with that of impact ionization. The latter is defined explicitly as the generation resulting from the *enhancement* of carrier velocities by the field. It is represented by impact ionization rates α_n, α_p defined as the number of electron-hole pairs generated per electron or hole per length of the path which the assembly travels with *drift* velocity v_n or v_p, respectively. The number of electron-hole pairs generated per unit of time, i.e. the avalanche generation rate G_{av}, is then given by

$$G_{av} = \alpha_n \cdot n \cdot |v_n| + \alpha_p \cdot p \cdot |v_p| = \frac{1}{q}\left(\alpha_n \cdot |j_n| + \alpha_p \cdot |j_p|\right) \tag{2.89}$$

On the right-hand side, the densities of the field currents are replaced by the total current densities, the diffusion currents are neglected because of the high fields.

Much experimental and theoretical work has been carried out to determine the ionization rates and their field dependency, which is very strong. Reviews have been given by Mönch [Moe69] and Maes et al. [Mae90]. Theoretically, Wolff predicted very early the relationship $\alpha \propto \exp(-b/\mathsf{E}^2)$ [Wof54]. Experimental ionization rates, on the other hand, were found by Chynoweth [Chy58] and later most other authors, to follow the relationship

$$\alpha = a\, e^{-\frac{b}{\mathsf{E}}}, \tag{2.90}$$

both for electrons and holes. With E we denote here the positive field strength in direction of the field. In a logarithmic plot versus $1/\mathsf{E}$, the values of α_n and α_p follow

Table 2.5 Coefficients of ionization rates in Eq. (2.90) as obtained from different publications. T = 300 K

Source	Electrons		Holes		Field for α_{eff} = 100/cm
Field range (E in 10^5 V/cm)		b (10^6 V/cm)	a (10^6/cm)	b (10^6 V/cm)	(10^5 V/cm)
Lee et al. [Lee64] 1.8 < E < 4	3.80	1.77	9.90	2.98	1.99
Ogawa [Oga65] 1.1 < E < 2.5	0.75[a]	1.39[a]	0.0188[a]	1.54[a]	1.88[a]
			4.65[b]	2.30[b]	1.80[b]
Overstraaten, De Man [Ove70] 1.75 < E < 4	0.703	1.231	1.582	2.036	1.66
Our choice [Sco91] 1.5 < E < 4	1.10	1.46	2.10	2.20	1.81

[a]directly measured [b]from α_{eff}, α_n

a straight line. Shockley derived this relationship [Sho61] using a simplified model for the manner the carriers reach the high ionization energy. As shown then by Baraff [Baf62], the field dependences of Wolff and of Chynoweth are limiting cases of a more general theory. The experimental determination is done usually by measuring carrier multiplication factors M_n and M_p, consisting of double integrals over the ionization rates (see Appendix C), which then have to be extracted. The measured field dependence obeys the relationship (2.90) down to low fields, as is not expected to this extent from theory [Mae90, Oga65]. Table 2.5 contains the constants a and b for silicon at room temperature as gathered from some often cited papers. From Ogawa's work [Oga65], two sets for α_p are given, the first one determined from multiplication factors like other values in the table. The second set describes the hole ionization rate determined in [Oga65] from measurements of the electron ionization rate α_n and the *effective* ionization rate

$$\alpha_{eff} = \frac{\alpha_n - \alpha_p}{\ln(\alpha_n/\alpha_p)} \tag{2.91}$$

which was determined from the breakdown voltage of pin diodes (see below). The extremely varying values of a (= $\alpha(E = \infty)$) in the table are mainly due to the extrapolation with different slopes in the plot versus 1/E owing to the different b-values. Nevertheless, also in the experimental range the diversity of the obtained ionization rates is large. For example, the α_{eff}– values obtained from [Lee64] and [Ove70] at a field strength of 2×10^5 V/cm differ by more than a factor 4. Although the differences in the *field* for a given α_{eff} are smaller because of the strong field dependence, they too are considerable. In the last column of Table 2.3, the field strengths belonging to α_{eff} = 100/cm are given. This field can be interpreted as the critical field E_c of a pin diode whose width of the intrinsic i-region is $w = 1/\alpha_{eff}$ = 100 μm, since the breakdown condition of a pin diode is $\alpha_{eff}(E_c) \cdot w = 1$ (see Eq. (2.93) below and Chaps. 3 and 5 on pn-junctions and pin-diodes). The breakdown voltage $V_B = w \cdot E_c$ calculated with the data of [Lee64] is 20% higher than obtained from [Ove70]. For p^+n junctions the difference is still larger (about 35% for V_B > 1000 V).

Calculations for a tight dimensioning of devices need ionization rates whose results are in close agreement with the measured blocking behavior. We have compared calculations using different α-sets with measurements on thyristors and diodes, whose weakly doped n base region was very homogeneously doped by neutron transmutation (see the chapter on technology), a method not used before 1974. Mainly devices in the blocking range 1200 to 6000 V were used for comparison, although data in the lower blocking range have also been considered. The ionization rates of Ogawa with the second set of α_p is found to be well suited for *diodes* with blocking voltage > 300 V. The blocking behavior of pn^-p-structures in thyristors, however, was found to be somewhat better described by a course of the hole ionization rate *between* the two strongly differing α_p-curves of [Oga65]. From these comparisons and by considering also measurements with low-blocking devices [Lee64, Ove70], the α parameters given in the last line of Table 2.5 are

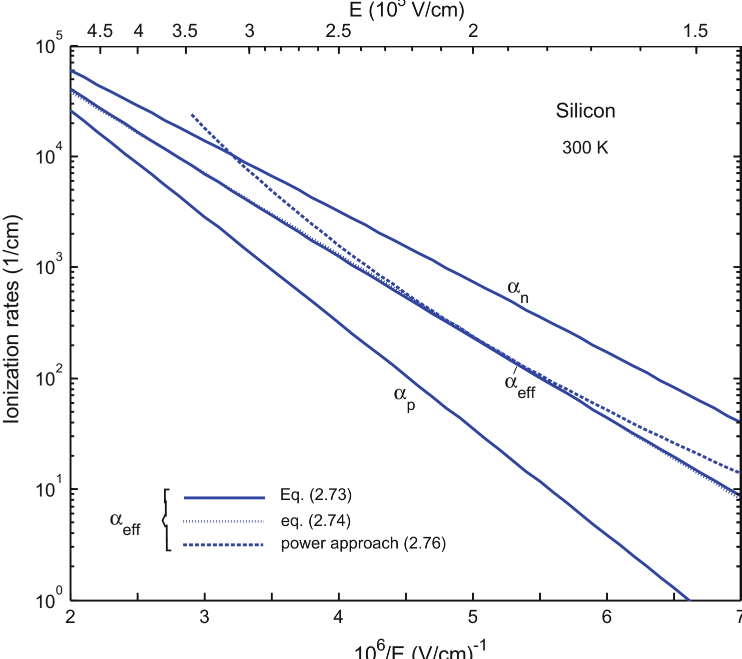


Fig. 2.23 Field dependence of ionization rates in silicon. The power approach (2.94) [Ful67, Shi59] is matched at the point $\mathsf{E}_0 = 2 \times 10^5$ V/cm using Eq. (2.95)

proposed for the simulation of power devices [Sco91]. They agree also well with a renewed determination of the ionization rates reported in a short paper of Valdinoci et al. [Val99], although these authors use a more complex fitting expression.

In Fig. 2.23 these ionization rates for silicon at 300 K (Table 2.5, last line) are plotted versus the inverse field strength. Within a relative small field range the α's vary by several orders of magnitude. The ionization rate of electrons is much larger than that for holes, and the effective ionization rate defined by (2.91) runs between $\alpha_n(\mathsf{E})$ and $\alpha_p(\mathsf{E})$. As is seen, α_{eff} too can be described with excellent approximation by a straight line in this plot and hence can be described by (2.90). Best agreement with (2.91) together with the parameter set in the last line of Table 2.5 is obtained with

$$\alpha_{eff} \approx 1.06 \times 10^6 \cdot e^{-1.68 \cdot 10^6/E} = a_{eff} \cdot e^{-b_{eff}/\mathsf{E}} \tag{2.92}$$

which up to a few percent agrees with the exact α_{eff} in the range 1.5×10^5 to 4×10^5 V/cm. Equation (2.92) differs only very little from the result of Ogawa [Oga65], who obtained

$$\alpha_{eff} = 1 \times 10^6 \exp(-1.66 \cdot 10^6/\mathsf{E})/\text{cm}.$$

The effective ionization rate (2.91) is defined in a manner that solely this α_{eff} determines the breakdown voltage of pn-junctions [Wul60, Oga65]. The breakdown condition which the field E(x) and hence voltage at breakdown must satisfy is

$$\int \alpha_{eff}(\mathsf{E}(x))dx = 1 \tag{2.93}$$

where the integration extends over the space charge region. The derivation is given in Appendix C. Although this holds exactly only if the ratio α_n/α_p is independent of E, it is a very good approximation also in most other cases (see Appendix C). To enable easy analytical integration for simple field shapes, Shields [Shi59] and Fulop [Ful67] approximated the field dependency by a power law, which normalized to a field E_0 reads

$$\alpha_{eff}(\mathsf{E}) \cong C\left(\frac{\mathsf{E}}{\mathsf{E}_0}\right)^n \tag{2.94}$$

Based on this approach, practicable and often used relationships for the breakdown voltage can be derived (see chaps. 3 and 5). If (2.94) together with its derivative is fitted to (2.92) at the point E_0, the constants are obtained as

$$n = b_{eff}/\mathsf{E}_0 \qquad C = a_{eff} \cdot \exp(-n) \tag{2.95}$$

where according to (2.92) $a_{eff} = 1.06 \times 10^6$ cm^{-1} and $b_{eff} = 1.68 \times 10^6$ V/cm. The value n = 7 used by Shields and Fulop is obtained for $E_0 = b_{eff}/7 = 2.40 \times 10^5$ V/cm. Matching at the field $\mathsf{E}_0 = 2 \times 10^5$ V/cm, Eq. (2.95) yields $n = 8.40$ and $C = 238$ cm^{-1}. Using these values, approximation (2.94) is plotted additionally in Fig. 2.20. Although the approximation is not very good over a wide range of E, very satisfying results can be obtained choosing the matching point E_0 near the maximum of the considered field distribution.

Whereas for diodes only α_{eff} is decisive for the breakdown voltage, for transistor structures such as the pn$^-$p structure in thyristors and IGBTs both ionization rates separately have an effect on the breakdown behavior. Of course, also the single ionization rates α_n and α_p can be approximated by a power law according to (2.94), (2.95). In this book, the Shields-Fulop approach will often be used to describe avalanche multiplication and blocking capability in an analytical way. If the *field* distribution $\mathsf{E}(x)$ in (2.93) does not allow an analytical integration, the power approximation loses its meaning.

The ionization rates decrease with increasing temperature because the mean free path between collisions with phonons decreases. As shown by Grant [Gra73] and Maes et al. [Mae90], the temperature dependence can be expressed by an increasing coefficient b in (2.90), while the pre-exponential factor a may be left constant. For electrons, the temperature coefficient db/dT was determined in [Gra73] to be

1300 V/(cmK), whereas from [Mae90] a mean value of 710 V/(cmK) is obtained; for holes Grant found db/dT = 1100 V/(cmK). Using a value of 1100 V/(cmK) in both cases, the temperature dependence of blocking behavior in the range −20 to 150 °C is well described according to our observations. Hence the following field and temperature dependences are found to be well suited for power devices:

$$\begin{aligned}\alpha_n &= 1.1 \times 10^6 \cdot e^{-\frac{1.46\times 10^6 + 1100(T-300K)}{E}}\ \mathrm{cm}^{-1} \\ \alpha_p &= 2.1 \times 10^6 \cdot e^{-\frac{2.2\times 10^6 + 1100(T-300K)}{E}}\ \mathrm{cm}^{-1}\end{aligned} \tag{2.96}$$

where the field is scaled in V/cm. Inserting this into (2.91) and using (2.92) at 300 K, the same T-dependence is obtained for α_{eff}:

$$\alpha_{eff} \cong 1.06 \times 10^6\ e^{-\frac{1.68\times 10^6 + 1100\cdot(T-300\ K)}{E}}\ \mathrm{cm}^{-1} = a_{eff} e^{-\frac{b_{eff}(T)}{E}} \tag{2.97}$$

The temperature dependence of the constants in the Shields approximation follows inserting the coefficient $b_{eff}(T)$ of the last equation into (2.95):

$$\begin{aligned}n(T) &= \frac{b_{eff}(T)}{E_0} = \frac{1.68 \times 10^6\ + 1100 \cdot (T - 300\ K)}{E_0} \\ C(T) &= \frac{a_{eff}}{e^{n(T)}} = \frac{1.68 \times 10^6\ /\mathrm{cm}}{e^{n(T)}}\end{aligned} \tag{2.98}$$

In the mentioned paper of Valdinoci et al. [Val99], the temperature dependency of the ionization rates in the range 300 – 670 K is described. The results [Val99] agree well with the above Eqs. (2.96). Singh and Baliga [Sin93] have determined the temperature dependence of the effective ionization rate in the range 77 – 300 K using the approach of Shields and Fulop. Although the temperature dependences of their constants differ partly strongly from (2.98), the ionization rate α_{eff} itself varies quite similarly (by a factor 2.4 from 300 to 100 K).

As mentioned in Sect. 2.1, SiC has the advantage of a very high critical field. Measurements of the ionization rates in 4H-SiC, the preferred polytype, have been performed by Konstantinov et al. [Kon98], Ng et al. [Ng03] and Loh et al. [Loh08]. The results were fitted by the following equations [Loh08], a modified form of (2.90):

$$\alpha_n = 2.78 \times 10^6 \exp\left[-\left(\frac{1.05 \times 10^7}{E}\right)^{1.37}\right]\ \mathrm{cm}^{-1} \tag{2.99}$$

$$\alpha_p = 3.51 \times 10^6 \exp\left[-\left(\frac{1.03 \times 10^7}{E}\right)^{1.09}\right]\ \mathrm{cm}^{-1} \tag{2.100}$$

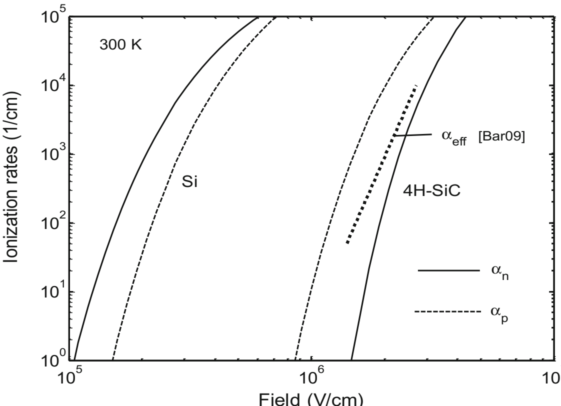


Fig. 2.24 Ionization rates of 4H-SiC and Si at 300 K. SiC-curves after [Kon98, Loh08, Ng03] (α_n, α_p) and [Bar09] (α_{eff})

where E has to be used in V/cm. These ionization rates together with those of silicon (2.96) are plotted in Fig. 2.24 versus E. As is seen, the field required for a given value of ionization rates is in 4H-SiC nearly an order of magnitude higher than in silicon. Contrary to silicon, the hole ionization rate α_p is larger than α_n in 4H-SiC.

An independent determination of the *effective* ionization rate has been published by Bartsch et al. [Bar09], who use measurements of the breakdown voltage of 4H-SiC $p^+n^-n^+$-diodes with different doping concentrations of the base region. From their measurements, which include temperature dependence, they obtain that

$$\alpha_{eff} \cong 2.18 \cdot 10^{-48}\, E^{8.03} \quad \mathrm{cm}^{-1}\ (E \text{ in V/cm}) \tag{2.101}$$

at 300 K. This dependency is plotted in Fig 2.24 additionally. The agreement with the other SiC curves on the whole is good. Nevertheless, considering (2.99), (2.100) and (2.91) on one side and (2.101) on the other the difference in the results is practically not unimportant. In the first case the effective α in the surrounding of 10^3/cm can be written in the power approach as $\alpha_{eff} \approx 3.43 \cdot 10^{-54} E^{9.02} cm^{-1}$. According to this, the field for the ionization rate 1000/cm is 1.82 MV/cm. According to Eq. (2.101) however the same ionization rate is reached only at 2.04 MV/cm. For a triangular field shape this means that the breakdown voltage according to (2.101) [Bar09] is 25% higher for the same doping concentration than according to (2.99), (2.100). This is probably the result of improvements in material quality of the SiC wafers and in SiC device manufacturing processes since the earlier measurements.

2.9 Basic Equations of Semiconductor Devices

The device operation depends on the processes with which the carriers and the electric field in the interior react on terminal currents and voltages. This is described by some basic equations which we will discuss now. A central part of these equations are the continuity equations of electrons and holes

$$-\frac{\partial n}{\partial t} = div\, \vec{\mathsf{J}}_{\mathsf{n}} + R_n \tag{2.102}$$

$$-\frac{\partial p}{\partial t} = div\, \vec{\mathsf{J}}_{\mathsf{p}} + R_p \tag{2.103}$$

Here $\vec{J}_n, \vec{J}_p$ are the vectors of particle current densities. The equations represent the time decrease of a carrier concentration (e.g.$-\partial n/\partial t$) as a flow of carriers out of the considered volume element ($div\, \vec{\mathsf{J}}_{\mathsf{n,p}}$) plus a disappearance of carriers with a rate R_n respectively R_p. In these exact mathematical equations one has to insert the previously derived physical relationships and models for the current densities and excess recombination rates R_n, R_p. The latter have to include generally, besides the thermal recombination-generation rates treated in Sect. 2.7 and called now $R_{n,p,th}$, the impact generation rate G_{av}: $R_{n,p} = R_{n,p,th} - G_{av}$. The electrical current densities have been given in one-dimensional form in (2.43a, 2.43b), (2.44). In two or three dimensions the particle current densities are

$$\vec{\mathsf{J}}_n = -\mu_n n\, \vec{\mathsf{E}} - D_n\, grad\, n \tag{2.104}$$

$$\vec{\mathsf{J}}_p = \mu_p p\, \vec{\mathsf{E}} - D_p\, grad\, p \tag{2.105}$$

The continuity equations include the law of conservation of charge

$$div\, \vec{j} + \frac{\partial \rho}{\partial t} = 0 \tag{2.106}$$

where $\vec{j}$ is the total electrical current density transported by the carriers (conduction current), $\vec{j} = q(\vec{J}_p - \vec{J}_n)$, and ρ the charge density or 'space charge'. Equation (2.106) is obtained by subtracting (2.102) from (2.103) and considering that the difference of the excess recombination rates $R_n - R_p$, if it is non-zero, signifies a recharging of impurities, usually of the recombination centers. Hence $R_n - R_p$ contributes to a change of charge density, which contains the charged flat and deep impurities in the form:

$$\rho = q(p - n + N_D^+ - N_A^- + N_r^+ - N_r^-) \tag{2.107}$$

A third differential equation on the level of the continuity equations results from the fundamental law

$$div\,\vec{D} = \rho \tag{2.108}$$

which states that an electric charge is a source of a displacement field $\vec{D}$. The latter is proportional to the electric field strength, $\vec{D} = \varepsilon\vec{E}$, where the permittivity constant ε splits up into the absolute permittivity (permittivity of vacuum) and the relative permittivity ε_r as a material constant, $\varepsilon = \varepsilon_r \cdot \varepsilon_0$. The ε_r-values of some materials are compiled in Appendix D. With (2.107) Eq. (2.108) can be written as follows:

$$div\,\varepsilon\vec{E} = \rho = q(p - n + N_D^+ - N_A^- + N_r^+ - N_r^-) \tag{2.109}$$

The further procedure now is to express the electric field in (2.104), (2.105) and (2.109) as the negative gradient of a potential V:

$$\vec{E} = -grad\,V \tag{2.110}$$

Equation (2.109) then turns into the Poisson equation:

$$div\,grad\,V = -\frac{q}{\varepsilon}(p - n + N_D^+ - N_A^- + N_r^+ - N_r^-) \tag{2.111}$$

where the semiconductor is assumed homogeneous and isotropic with regard to ε. For non-cubic crystals ε_r is a tensor, but in current simulation programs it is assumed to be a scalar. The differential operator on the left hand side of (2.111) is $div\,grad = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$, the Laplace operator.

The Poisson equation and the continuity Eqs. (2.102), (2.103) with (2.110) substituted for $\vec{E}$ in the current densities form a system of three partial differential equations with the unknown variables V, n and p. The doping structure and the density of recombination centers as well as boundary conditions at the surface and the characteristics of the external circuit are required to be known. The normal donors and acceptors are assumed usually to be completely ionized and the concentration of deep impurities to be small. The concentration of charged impurities is then given by the net doping concentration $N_D - N_A$, and the recombination rates R_n, R_p are approximately equal (see Sect. 2.7). The three differential equations are then sufficient to calculate the potential V and the carrier concentrations n and p as functions of space and time. The Eqs. (2.102)–(2.103) and (2.110), (2.111) with $N_D^+ = N_D$, $N_A^- = N_A$, $N_r^+ = N_r^- = 0$ and $R_n = R_p = R$ are called therefore the basic semiconductor equations. The behavior of this system is investigated in the book of Selberherr [Sel84]. Simulation programs based on these equations in one to three dimensions are on the market. The two- and three-dimensional programs are rather complex with all implications. In special one-dimensional cases the equations are used widely as starting point for analytical calculations.

In fast switching devices, the charge of deep traps is often considerable, and additionally the incomplete ionization of normal dopants can be a subject of interest, particularly in semiconductors like SiC. In these cases, (2.109), (2.110) and (2.111) have to be used in their general form, and for each impurity level, whose variable partial occupation has to be considered, one has to add an equation of the form (2.78) to the system. A program including these effects is Sentaurus^{TCAD} [SYN07].

So far we have assumed that the device has the same temperature at all points. Often this is a good approximation because the thickness of the semiconductor chips is relatively small and the thermal conductivity of silicon is high. On the other hand, phenomena of current constriction, thermal run-away and other processes responsible for destruction at heavy loads are connected often with a strongly inhomogeneous temperature distribution. For power devices, heat conduction is the primary criterion to be considered for the area and package of a device. To supplement the above system for inclusion of variable temperature and to provide a basis for thermal estimates, one has to add the equation of heat conduction. The heat current density, i.e. the heat energy transported through a surface element per unit area and time, is proportional to the negative gradient of temperature

$$J_{heat} = -\lambda \cdot grad\ T \tag{2.112}$$

where the proportionality factor λ is the thermal conductivity. The thermal energy per volume belonging to a temperature rise ΔT is $Q = \rho_m \cdot c \cdot \Delta T$, where ρ_m denotes the specific (mass) density and c the specific heat of the material. Using this, the continuity equation describing the conservation of thermal energy is obtained in the form

$$\rho_m c \cdot \frac{\partial T}{\partial t} = div\ (\lambda \cdot\ grad\ T) + H \tag{2.113}$$

where H signifies the heat generation rate per unit volume. This consists of the ohmic energy dissipation $\vec{E} \cdot \vec{j}$ and the heat produced by the net recombination of carriers, $R \cdot E_g$ [Sel84]:

$$H = \vec{E} \cdot \vec{j} + R \cdot E_g \tag{2.114}$$

where the non-degenerate case is assumed. R is defined here again as the total net recombination rate including the impact generation: $R = R_{th} - G_{av}$. By adding the heat flow equation Eq. (2.113) to the above system of partial differential equations, the temperature distribution can be calculated in addition to V, n and p. All quantities in these equations, such as the mobilities and the carrier lifetime, have to be used in their temperature dependent form. The equations of current densities may be unchanged for a first approximation only. A more detailed discussion can be found in [Sel84]. We note that in silicon $\rho_m \cdot c$ is nearly independent of temperature and has the value $\rho_m \cdot c = 2.0\ Ws/(\text{cm}^3\text{K})$, whereas the heat conductivity λ decreases

considerably with T. In the range 220–600 K, the temperature dependence of λ is described by the formula [Kos74]:

$$\lambda = \frac{320}{T - 82} \quad \frac{W}{\text{cmK}} \tag{2.115}$$

where T is scaled in Kelvin. The one-dimensional increase of temperature from the lower surface of a chip towards the interior can be simply estimated for a given load using these equations.

Till now the ***Maxwell equations*** as the general description of all electro-magnetic phenomena have not been used. The question arises, how far they are satisfied by the above calculation method. Law (2.108) is one of the Maxwell equations. Furthermore, the equation of conservation of charge (2.106) can be deduced from the first Maxwell equation

$$rot\ \vec{H} = \vec{j} + \frac{\partial \vec{D}}{\partial t} \tag{2.116}$$

by applying the div operator und using (2.108). Beyond this, however, the Maxwell equations are not taken into account. According to (2.116) the current is accompanied with a magnetic field $\vec{H}$ and a magnetic induction $\vec{B} = \mu\mu_0\vec{H}$. Fast changes of the current density $\vec{j}$ and the displacement current density $\partial\vec{D}/\partial t$ cause a corresponding variation of $\vec{H}$ and $\vec{B}$, and according to the second Maxwell equation

$$rot\ E = -\frac{\partial \vec{B}}{\partial t} \tag{2.117}$$

this leads to an induced electric field. Since $rot\ \vec{E} \neq 0$, this field cannot be represented by the gradient of a potential as in (2.110); hence it is not contained in the described differential equations. Effects of self-induction and current crowding (skin effect), which play an important part in metal wires, are therefore ignored. Due to the smaller current densities in semiconductors these effects are significantly smaller. In the current equations above we have omitted furthermore the Lorentz force $\pm\ q\ \vec{v} \times \vec{B}$ on the carriers. External magnetic fields as used for the Hall effect (see Sect. 2.4) are excluded therefore. Apart from this, the omission of the Lorentz force is in line with disregarding the second Maxwell equation, both neglecting the feedback of the induced magnetic field on the current distribution. Switching processes where these effects become significant have not been demonstrated however till now to our knowledge. For the great majority of possible applications the discussed basic device equations form an adequate basis for a realistic device simulation.

2.10 Simple Conclusions

We consider now a few simple but important consequences of the above equations. A homogeneous n-type semiconductor is assumed, and on one hand the decay of a small minority carrier density, on the other hand the decay of a small charge density is studied, both as a function of time after external generation and as function of the distance x from the surface in the case of a stationary generation at the surface.

2.10.1 Temporal and Spatial Decay of a Minority Carrier Concentration

A low injection hole density $\Delta p = p$ in a neutral n region is assumed, the hole charge is neutralized then by an equal excess electron concentration $\Delta n \equiv n - n_0 = \Delta p$.

(a) Temporal decay of a homogeneous hole concentration: This decay is obtained from (2.103) with $div\,\vec{J}_p = 0$ and has been given already by Eq. (2.50). The minority concentration decays with the time constant τ_p, the minority carrier lifetime.

(b) Spatial decay for a stationary generation at the surface: Since the current density $j_n + j_p = 0$, it follows from (2.43a), (2.43b) that a so-called "Dember field"

$$\mathsf{E} = -\frac{D_n - D_p}{\mu_n n}\frac{dp}{dx}$$

is necessary to maintain this condition. Since $p \ll n$, the hole current caused by this field is however small against the diffusion current, hence $J_p = -D_p dp/dx$. Using this in (2.103) together with $\partial p/\partial t = 0$ and $R_p = \Delta p/\tau_p$ one obtains

$$-D_p \frac{d^2 p}{dx^2} + \frac{p}{\tau_p} = 0$$

The solution is

$$p(x) = p(0) \cdot e^{-x/L_p}$$

with

$$L_p = \sqrt{D_p \tau_p} \tag{2.118}$$

This decay length by which the minority carrier concentration decreases is called **diffusion length** of the minority carriers (here holes), indicating that the spreading

takes place by diffusion. L_p can be adjusted by the minority carrier lifetime τ_p. With $D_p = kT/q \cdot \mu_p = 12$ cm^2/s (for small doping density) a lifetime of 1 μs results in a diffusion length of 34.6 μm. The diffusion length of electrons is somewhat larger for a given lifetime because of the higher diffusion constant D_n. The diffusion lengths in devices are chosen often with reference to the base width w_B. With the appropriate lifetime, diffusion lengths of several hundred microns are possible. Compared with a charge density considered in next section under b) the minority carriers gain a much larger distance from the point of generation.

2.10.2 Temporal and Spatial Decay of a Charge Density

A small charge density or space charge $\rho = -q \cdot \delta n$ with $\delta n << n_0$ is assumed, the deviation δn from the equilibrium n_0 being not neutralized by an equal hole concentration.

(a) Temporal decay of a homogeneous charge density: From the equation of conservation of charge (2.106) one obtains with $j = q \cdot \mu_n \cdot n_0 \cdot \mathsf{E}$ and $div\, \mathsf{E} = \rho/\varepsilon$

$$\frac{q \cdot \mu_n \cdot n_0}{\varepsilon}\rho + \frac{dp}{dt} = 0$$

This results in the time decay

$$\rho(t) = \rho(0) \cdot e^{-t/\tau_{rel}} \tag{2.119}$$

where the time constant is the **relaxation time**

$$\tau_{rel} = \frac{\varepsilon}{q \cdot \mu_n \cdot n_0} = \frac{\varepsilon_r \varepsilon_0}{\sigma} \tag{2.120}$$

τ_{rel} is inversely proportional to the electrical conductivity σ and very small; in silicon at $n_0 = 1 \times 10^{15}$ cm^{-3} ($\mu_n = 1350$ cm^2/Vs, $\sigma = 0.22$ A/Vcm) one obtains $\tau_{rel} = 4.8$ ps. A homogeneous space charge left to itself can exist only very shortly.

(b) Decrease of a space charge with distance x from the surface: The stationary decrease towards the inner is considered starting from a constant charge density $\rho(0) = -q \cdot \delta n(0)$ at the surface, caused for example by a positive voltage at an isolated gate electrode. Due to the nearly zero hole concentration the hole current is negligible, hence the total current density $j = j_n$. Since $d\rho/dt = 0$, one obtains from the equation of charge conservation (2.106) using (2.43a) together with the Einstein relation (2.44)

$$0 = \frac{dj}{dx} = q\mu_n\left(n\frac{d\mathsf{E}}{dx} + \frac{dn}{dx}\mathsf{E} + \frac{kT}{q}\frac{d^2n}{dx^2}\right)$$

The term $dn/dx \cdot \mathsf{E}$ as a product of two small quantities is negligible. Hence, using (2.108) and $\delta n = -\rho/q$, (2.118) turns into

$$\frac{n_0}{\varepsilon}\rho - \frac{kT}{q^2}\frac{d^2\rho}{dx^2} = 0$$

This yields again an exponential decay with distance x:

$$\rho(x) = \rho(0) \cdot e^{-\frac{x}{L_D}} \tag{2.121}$$

where now

$$L_D = \sqrt{\frac{\varepsilon_r\varepsilon_0 \cdot kT}{n_0 \cdot q^2}} = \sqrt{D_n\tau_{rel}} = 0.409\sqrt{\frac{T/300\,K}{n_0/10^{14}\text{cm}^{-3}}}\ \mu\text{m} \tag{2.122}$$

This length determining the spatial decay of a charge density is called **Debye length**. It is inversely proportional to the equilibrium carrier density. The numerical expression on the right-hand side is obtained for silicon with $\varepsilon_r = 11.7$. Even at $n_0 = 1 \times 10^{14}$ cm^{-3}, L_D amounts only to 0.41 μm at 300 K. Hence the space charge decreases very rapidly towards zero. The Debye length is typically two or three orders of magnitude smaller than the diffusion lengths. The relation between L_D and the relaxation time is similar to the relation between the minority carrier diffusion length and lifetime. The *majority* carrier concentration and mobilty appearing in the Debye length (2.122) and the relaxation time (2.120) show that the spreading and decay of the space charge is a majority carrier effect.

References

[Abb84] Abbas CC: A theoretical explanation of the carrier lifetime as a function of the injection level in gold-doped silicon. IEEE Trans. Electron devices, ED-31 pp. 1428–1432 (1984)

[Atk85] Atkinson CJ: "Power devices in gallium arsenide", IEE proceedings **132**, Pt.I, pp. 264–271 (1985)

[Baf62] Baraff, G.A.: Distribution functions and ionization rates for hot electrons in semiconductors. Phys. Rev. **128**, 2507–2517 (1962)

[Bal77] Baliga B.J., Krishna S.: Optimization of recombination levels and their capture cross section in power rectifiers and thyristors. Solid-State Electro-nics **20**, 225–232 (1977)

[Bar09] Bartsch W., Schoerner R., Dohnke K.O.: Optimization of bipolar SiC-diodes by analysis of avalanche breakdown performance. Proceedings of the ICSCRM 2009, paper Mo-P-56 (2009)

[Bla62] Blakemore, J.S.: Semiconductor statistics, 1st edn. Pergamon Press, Oxford (1962)

[Bro82a] Brotherton S.D., Bradley P.: Measurement of minority carrier capture cross sections and application to gold and platinum in silicon. J. Appl. Phys. **53**, 1543–1553 (1982)

[Bro82b] Brotherton S.D., Bradley P.: A comparison of the performance of gold and platinum killed power diodes. Solid-State Electronics **25**, 119–125 (1982)

[Cau67] Caughey, D.M., Thomas, R.E.: Carrier mobilities in silicon empirically related to doping and field. Proceedings IEEE **23**, 2192–2193 (1967)

[Chy58] Chynoweth, A.G.: Ionization rates for electrons and holes in silicon. Phys. Rev. **109**(5), 1537–1540 (1958)

[Con71] Conti, M., Panchieri, A.: Electrical properties of platinum in silicon. Alta Frequenza **40**, 544–546 (1971)

[Cor74] Cornu, J., Sittig, R., Zimmermann, W.: Analysis and measurement of carrier lifetimes in the various operating modes of power devices. Solid-State Electronics **17**, 1099–1106 (1974)

[Dan72] Dannhäuser, F.: Die abhängigkeit der trägerbeweglichkeit in silizium von der konzentration der freien ladungsträger – I. Solid-State Electronics **15**, 1371–1375 (1972)

[Deb54] Debye, P.P., Conwell, E.M.: Electrical properties of n-type Germanium. Phys. Rev. **93**, 693–706 (1954)

[Dzi77] Dziewior, J., Schmid, W.: Auger coefficients for highly doped and highly excited silicon. Appl. Phys. Lett. **31**, 346–348 (1977)

[Eng75] Engström, O., Grimmeis, H.G.: Thermal activation energy of the gold acceptor level in silicon. J. Appl. Phy. **46**, 831–837 (1975)

[Evw76] Evwaraye A.O., Sun E.: Electrical properties of platinum in silicon as determined by deep-level transient spectroscopy. J. Appl. Phys. **47**, 3172–3176 (1976)

[Fai65] Fairfield, J.M., Gokhale, B.V.: Gold as a recombination centre in silicon. Solid-St. Electronics **8**, 685–691 (1965)

[Fle57] Fletcher, N.H.: The high current limit for semiconductor junction devices. Proc. IRE **45**, 862–872 (1957)

[Fri06] Friedrichs P.: SiC power devices – recent and upcoming developments IEEE ISIE 2006, July 9-12, Montreal, Quebec, Canada (2006)

[Ful67] Fulop, W.: Calculation of avalanche breakdown Voltages of silicon pn-junctions. Solid State Electron. **10**, 39–43 (1967)

[Gil79] Gildenblat, G.Sh., Grot, S.A., Badzian, A.: Proc. IEEE **79**(5), 647–668 (1991)

[Gol01] Goldberg, Y., Levinshtein, M.E., Rumyantsev, S.L.: In: Levinshtein et al. (eds) Properties of advanced semiconductor materials GaN, AlN, SiC, BN, SiC, SiGe. pp. 93–148. Wiley, New York (2001)

[Gra73] Grant, W.N.: Electron and hole ionization rates in epitaxial silicon at high electric fields. Solid-State Electronics **16**, 1189–1203 (1973)

[Gre90] Green, M.A.: Intrinsic concentration, effective densities of states, and effective mass in silicon. J. Appl. Phys. **67**, 2944–2954 (1990)

[Hag93] Hagmann, G.: Leistungselektronik. Aula-Verlag, Wiesbaden (1993)

[Hal52] Hall, R.N.: Electron-hole recombination in germanium. Phys. Rev. **87**, 387 (1952)

[Han87] Hangleiter, A.: Nonradiative recombination via deep impurity levels in silicon: experiment. Phys. Rev. B **15**, 9149–9161 (1987)

[Hon15] Honea, J., Zhan Wang, Z., Wu, Y.: Design and implementation of a high-efficiency three-level inverter using GaN HEMTs. Proceedings of the PCIM Europe, pp. 486–492 (2015)

[Hur92a] Hurkx, G.A.M., Klaassen, D.B.M., Knuvers, M.P.G.: A new recombination model for device simulation including tunneling. IEEE Trans. on electron devices **39**, 331–338 (1992)

[Hur92b] Hurkx, G.A.M., de Graaff, H.C., Klosterman, W.J., Knuvers, M.P.G.: A new analytical diode model including tunneling and avalanche breakdown. IEEE Trans. on electron dev. **39**, 2000–2008 (1992)

[Ike08] Ikeda, N., Kaya, S., Jiang, L., Sato, Y., Kato, S., Yoshida, S.: High power AlGaN/GaN HFET with a high breakdown voltage of over 1.8 kV on 4 inch Si substrates and the suppression of current collapse. Proceedings of the ISPSD '08 pp. 287–290 (2008)

[Ish15] Ishida, M., Ueda, T.: GaN-based Gate Injection transistors for power switching applications. Japan-EU symposium on Power Electronics December 15-16, Tokyo (2015)

[Jac77] Jacobini, C., Canali, C., Ottaviani, G., Quaranta, A.: Review of some charge transport properties of silicon. Sol. State Electr. **20**, 77–89 (1977)

[Kan93] Kane, D.E., Swanson, R.M.: Modeling of electron-hole scattering in semiconductor devices. IEEE Trans. on Electron Dev. **40**, 1496–1500 (1993)

[Kla92] Klaassen, D.B.M.: A unified mobility model for device simulation—I. Model equations and concentration dependence. Solid State Electron. **35**(7), 953–959 (1992)

[Koh57] Kohn, W.: Shallow impurity states in silicon and germanium. In: Seitz, F., Turnbull, D (eds.) Solid State Physics, vol. 5, (1957)

[Kos74] Kokkas, A.G.: Empirical relationships between thermal conductivity and temperature for silicon and germanium. RCA Rev **35**, 579–581 (1974)

[Kon59] Kontsevoi Y.A.: Determination of capture cross sections in the case of recombination on multicharged centers. Sov. Phys. - JETP 1177–1181 (1959)

[Kon98] Konstantinov, A.O., Wahab, Q., Nordell, N., Lindefelt, U.: Study of avalanche breakdown and impact ionization in 4H silicon carbide. J. Electron. Mater. **27**(4), 335–341 (1998)

[Kra72] Krause, J.: Die Abhängigkeit der Träderbeweglichkeit in Silizium von der Konzentration der freien Ladungsträger – II. Solid-St. Electron **15**, 1377–1381 (1972)

[Kuz86] Kuzmicz, W.: Ionization of impurities in silicon. Solid-St. Electron **29**, 1223–1227 (1986)

[Lan80] Lang, D.V., Grimmeis, H.G., Meijer, E., Jaros, M.: Complex nature of gold-related deep levels in silicon. Phys. Rev. B **22**, 3917–3934 (1980)

[Lan79] Lanyon, H.P.D., Tuft, R.A.: Bandgap narrowing in moderately to heavily doped silicon. IEEE Trans. Electron Devices, vol. ED-26, pp. 1014–1018 (1979)

[Lar54] Lark-Horovitz, K.: The new electronics. In: The present State of Physics (American Assn. for the Advancement of Science), Washington, 1954

[Lee64] Lee, C.A., Logan, R.A., Batdorf, R.L., Kleimack, J.J., Wiegmann, W.: Ionization rates of holes and electrons in silicon. Phys. Rev. **134**, A761–A773 (1964)

[Lev01] Levinshtein, M.E., Rumyantsev, S.L., Shur, M.S.: Properties of advanced semiconductor materials. Wiley, New York (2001)

[Li78] Li, S.S.: The dopant density and temperature dependence of hole mobility and resistivity in boron doped silicon. Solid State Electron. **21**, 1109–1117 (1978)

[Lis75] Lisial K.P., Milnes A.G.: Energy levels and concentrations for platinum in silicon, Soli-State Electronics **18**, 533–540 (1975)

Same authors: Platinum as a lifetime-control deep impurity in silicon. J. App. Phys. **46**, 5229–5235 (1975)

[Loh08] Loh, W.S., Ng, B., Soloviev, K., et al.: Impact ionization coefficients in 4H-SiC. IEEE Trans. Electron Devices **55**, 1984–1990 (2008)

[LuN87] Lu, L.S., Nishida, T., Sah, C.T.: Thermal emission and capture rates of holes at the gold donor level in silicon. J. Appl. Phys. **62**, 4773–4780 (1987)

[Luo71] Luong, M., Shaw, A.W.: Quantum transport theory of impurity scattering—Limited mobility in n-type Si. Phys. Rev. B **4**, 2436–2441 (1971)

[LuS86] Lu, L.S., Sah, C.T.: Electron recombination rates at the gold acceptor level in high-resistivity silicon. J. Appl. Phys. vo. **59**, 173–176 (1986)

[Lut94] Lutz J, Scheuermann U: Advantages of the new controlled axial lifetime diode. Proceedings of the 28th PCIM, pp. 163–169 (1994)

[Lut00] Lutz, J.: Freilaufdioden für schnell schaltende Leistungsbauelemente. Dissertation, Techn. Univ. Ilmenau 2000, ISLE Publishing House

[Mad64] Madelung, O.: Physics of III-V Compounds. Wiley, New York (1964)

[Mae90] Maes, W., De Meyer, K., Van Overstraaten, R.: Impact ionization in silicon: a review and update. Solid-St. Electron. **33**, 705–718 (1990)

[Mas83] Masetti, G., Severi, M., Solmi, S.: Modeling of carrier mobility against concentration in Arsenic-, Phosphorus-, and Boron-doped Silicon. IEEE Trans Electron Devices, **ED-30**(7), 764–769 (1983)

[Mil76a] Miller M.D., Schade H., Nuese C.J.: Lifetime-controlling recombination centers in platinum-diffused silicon. J. Appl. Phys. **47**, 2569–2578 (1976)

[Mil76b] Miller, M.D.: Differences between platinum- and Gold-Doped silicon power devices. IEEE Trans. El. Dev., **ED-23**,12 (1976)

[Mna87a] Mnatsakanov T.T.: Transport coefficients and Einstein relation in a high density plasma of solids. Phys. stat. sol. (b), **143** 225–234 (1987)

[Mna87b] Mnatsakanov, T.T., Rostovtsev, I.L., Philatov, N.I.: Investigation of the effect of nonlinear physical phenomena on charge carrier transport in semiconductor devices. Solid-St. Electronics **10**, 579–585 (1987)

[Mna98] Mnatsakanov, T.T., Schröder, D., Schlögl, A.: Effect of high injection level phenomena on the feasibility of diffusive approximation in semiconductor device modeling. Solid-St. Electronics **42**, 153–163 (1998)

[Moe69] Mönch, W.: On the physics of avalanche breakdown in semiconductors. Phys. Stat. Sol. **36**, 9–48 (1969)

[Mol64] Moll, J.L.: Physics of semiconductors. McGraw Hill, New York (1964)

[Mon74] Monemar, B: Fundamental energy gap of GaN from photoluminescence excitation spectra. Phys. Rev. B. **10**(2), (1974)

[Mor14] Morita, T, Tanaka, K, Ujita, S, Ishida, M, Uemoto, Y, Ueda, T.: Recent Progress in Gate Injection Technology based GaN Power Devices. Proceedings of the ISPS, Prague, pp. 34–37, (2014)

[Mue93] von Münch, W.: Einführung in die Halbleitertechnologie. B.G. Teubner, Stuttgart (1993)

[Ng03] Ng, B.K., David, J.P.R., Tozer, R.C., et al.: Non-local effects in thin 4H-SiC UV avalanche photodiodes. IEEE Trans. Electron Devices **50**, 1724–1732 (2003)

[Niw08] Niwa, F., Misumi, T., Yamazaki, S., Sugiyama, T., Kanata, T., Nishiwaki, K.: A study of correlation between CiOi defects and dynamic avalanche phenomenon of PiN diode using he ion irradiation. Proceedings of the PESC, Rhodos (2008)

[Oga65] Ogawa, T.: Avalanche breakdown and multiplication in silicon pin junctions. Jpn. J. Appl. Phys. **4**, 473 ff (1965)

[Oka90] Okano, K., Kiyota, H., Iwasaki, T., Kurosu, T., Ida, M., Nakamura, T.: Proc. Second Int. Conf. on the New Diamond Science and Technology, Washington D.C., pp. 917–922 (1990)

[Ove70] Van Overstraeten, R., De Man, H.: Measurement of the Ionization Rates in Diffused Silicon p-n junctions. Solid State Electron. **13**, 583–608 (1970)

[Pal74] Pals J.A.: Properties of Au, Pt, Pd, and Rh levels in silicon measured with a constant capacitance technique. Solid-St. Electronics **17**, 1139–1145 (1974)

[Pog64] Poganski, S.: Fortschritte auf dem Gebiet der Selen-Gleichrichter. AEG-Mitteilungen **54**, 157–161 (1964)

[Qua08] Quay, R.: Gallium Nitride Electronics. Springer, Berlin Heidelberg (2008)

[Ras08] Rashid, S.J., Udrea, F., Twitchen, D.J., Balmer, R.S., Amaratunga, G.A.J.: Single crystal diamond schottky diodes – practical design considerations for enhanced device performance. Proceedings of the ISPS, Prague (2008)

[Sah58] Sah C.T., Shockley W.: Electro-hole recombination statistics in semiconductors through flaws with many charge conditions. Phys. Rev. **109**, 1103–1115 (1958)

[Sah69] Sah, C.T., Forbes, L., Rosier, L.I., Tasch Jr., A.F., Tole, A.B.: Thermal emission rates of carriers at gold centers in silicon. Appl. Phys. Lett. **15**, 145–148 (1969)

[Scf69] Scharfetter, D.L., Gummel, H.K.: Large-signal analysis of a silicon Read Diode oscillator. IEEE Trans. Electron Dev. ED-16, pp. 64–77 (1969)

[Scm82] Schmid, W., Reiner, J.: Minority carrier lifetime in gold-diffused silicon at high carrier concentrations. J. Appl. Phys. **53**, 6250–6252 (1982)

[Sco74] Schlangenotto, H., Maeder, H., Gerlach, W.: Temperature dependence of the radiative recombination coefficient in silicon. Phys. Stat. Sol. (a) **21**, 357–367 (1974)

[Sco76] Schlangenotto, H., Maeder, H., Dziewior. J.: Neue Technologien für Silizium-Leistungsbauelemente—Rekombination in hoch dotierten Emitterzonen. Research Report T 76–54, German Ministry of Research and Technology (1976)

[Sco91] Schlangenotto H: Script of lectures on Semiconductor power devices, Technical University Darmstadt, 1991 (in German)

[Scr94] Schäfer, W.J., Negley, G.H., Irvin, K.G., Palmour, J.W.: Conductivity anisotropy in epitaxial 6H and 4H SiC. In: Proceedings of Material Res. Society Symposium, Bd. 339, 595–600 (1994)

[Scz66] Schultz, W.: Rekombinations- und Generationsprozesse in Halbleitern. Festkörperprobleme Band V, pp. 165–219, Vieweg & Sons, Braunschweig, (1966)

[Sel84] Selberherr, S.: Analysis and simulation of semiconductor devices. Springer Vienna, (1984)

[Shi59] Shields, J.: Breakdown in Silicon pn-Junctions. Journ. Electron. Control, no. 6 pp. 132 ff (1959)

[Sho52] Shockley, W., Read Jr., W.T.: Statistics of the recombinations of holes and electrons. Phys. Rev. **87**, 835–842 (1952)

[Sho59] Shockley, W.: Electrons and Holes in Semiconductors, Seventh printing. D. van Nostrand Company Inc, Princeton (1959)

[Sho61] Shockley, W.: Problems related to p-n junctions in silicon. Solid-St. Electronics **2**, 35–67 (1961)

[Sie01] Siemieniec, R., Netzel, M., Südkamp, W., Lutz, J.: Temperature dependent properties of different lifetime killing technologies on example of fast diodes. IETA2001, Cairo, (2001)

[Sie02] Siemieniec, R., Südkamp, W., Lutz, J.: Determination of parameters of radiation induced traps in silicon. Solid-State Electr. **46**, 891–901 (2002)

[Sie06] Siemieniec, R., Niedernostheide, F.J., Schulze, H.J., Südkamp, W., Kellner-Werdehausen, U., Lutz, J.: Irradiation-induced deep levels in silicon for power device tailoring. J. Electrochem. Soc. **153**(2), G108–G118 (2006)

[Sin93] Singh, R., Baliga, B.J.: Analysis and optimization of power MOSFETs for cryogenic operation. Sol. State Electronics **36**, 1203–1211 (1993)

[Slo76] Slotboom, J.W., De Graaff, H.C.: Measurements of bandgap narrowing in Si bipolar transistors. Solid-State Electronics **19**, 857–862 (1976)

[Slo77] Slotbomm, J.W.: The pn-product in Silicon. Solid State Electron. **20**, 279–283 (1977)

[Smi59] Smith, R.A.: Semiconductors. University Press, Cambridge (1959)

[Spe58] Spenke, E.: Electronic semiconductors, 1st edn. McGraw Hill, New York (1958)

[Sue94] Südkamp, W.: DLTS-Untersuchung an tiefen Störstellen zur Einstellung der Trägerlebensdauer in Si-Leistungsbauelementen, Dissertation, Technical University of Berlin, (1994)

[SYN07] Advanced tcad manual, Synopsys Inc. Mountain View, CA. Available: http://www.synopsys.com (2007)

[Sze81] Sze, S.M.: Physics of semiconductor devices. Wiley, New York (1981)

[Sze02] Sze, S.M.: Semiconductor devices, physics and technology, 2nd edn. Wiley, New York (2002)

[Tac70] Tach Jr., A.F., Sah, C.T.: Recombination-Generation and optical properties of gold acceptor in silicon. Phys. Rev. B **1**, 800–809 (1970)

[Tho80] Thornber, K.K.: Relation of drift velocity to low-field mobility and high-field saturation velocity. J. Appl. Phys. **51** (1980)

[Thr75] Thurmond, C.D.: The standard thermodynamic functions for the formation of electrons and holes in Ge, Si, GaAs, and GaP. J. Electrochem. Soc., Solid State Science and Technology, **122**(8) (1975)

[Thu80a] Thurber, W.R., Mattis, R.L., Liu, Y.M., Filliben, J.J.: Resistivity-dopant density relationship for phosporous-doped silicon. J. Electrochem. Soc. **127**, 1807–1812 (1980)

[Thu80b] Thurber, W.R., Mattis, R.L., Liu, Y.M., Filliben, J.J.: Resistivity-dopant density relationship for boron-doped silicon. J. Electrochem. Soc. **127**, 2291–2294 (1980)

[Twi04] Twitchen, D.J., Whitehead, A.J., Coe, S.E., Isberg, J., Hammersberg, J., Wikström, T., Johansson, E.: High-voltage single-crystal diamond diodes. IEEE Trans. on Electron Dev. **51**, 826–828 (2004)

[Ued05] Ueda, D., Murata, T., Hikita, M., Nakazawa, S., Kuroda, M., Ishida, H., Yanagihara, M., Inoue, K., Ueda, T., Uemoto, Y., Tanaka, T., Egawa, T.: AlGaN/GaN devices for future power switching systems. IEEE International Electron Devices Meeting, pp. 377–380 (2005)

[Val99] Valdinoci, M., Ventura, D., Vecchi, M., Rudan, M., Baccarani, G., Illien, F., Stricker, A., Zullino, L.: Impact-Ionization in silicon at large operating temperature. Proc. SISPAD'99, pp. 27–30, Kyoto, (1999)

[Var67] Varshni, Y.P.: Temperature dependence of the energy gap in semiconductors. Physica **34**, 149–154 (1967)

[Vec76] Van Vechten, J.A., Thurmond, C.D.: Entropy of ionization and temperature variation of ionization levels of defects in semiconductors. Phys. Review B **14**, 3539–3550 (1976)

[Wfs60] Wolfstirn, K.B.: Hole and electron mobilities in doped silicon from radiochemical and conductivity measurements. J. Phys. Chem. Solids **16**, 279–284 (1960)

[Wof54] Wolff, P.A.: Theory of electron multiplication in silicon and germanium. Phys. Rev. **95**, 1415–1420 (1954)

[WuP82] Wu, R.H., Peaker, A.R.: Capture cross sections of the gold and acceptor states in n-type Czochralski silicon. Solid-St. Electronics **25**, 643–649 (1982)

[Wul60] Wul, B.M., Shotov, A.P.: Multiplication of electrons and holes in p-n junctions. Solid State Phys. in Electron. Telecommun. **1**, 491–497 (1960)

[Zim73] Zimmermann, W.: Experimental verification of the Shockley-Read-Hall recombination theory in silicon. Electron. Lett. **9**, 378–379 (1973)

Chapter 3
pn-Junctions

pn-junctions are a basic element of nearly all power devices. They are formed when the type of conductivity changes from p to n-type within the same crystal. pn-junctions are rectifying, they conduct current only in one direction of the applied voltage, called forward direction, whereas in the opposite direction, the blocking direction, the current is extremely small. The rectification of pn-junctions has been described first theoretically by Davidov [Dav38]. After some further work (see [Sho49]) the pn theory was developed in a more complete form by Shockley [Sho49, Sho50]. Together with the invention of the transistor this was a starting point of the enormous development of the semiconductor industry till now. The operation of the metal-semiconductor rectifier was known already at that time [Sch38, Sch39]. Today, even the early poly-crystal rectifiers are thought to function mostly by a pn-junction.

The rectifying effect can be simply understood qualitatively (see Fig. 3.1). If a positive voltage is applied to the p-region with respect to the n-region, then the free holes in the p-region and the free electrons in the n-region are driven towards the junction and are injected partly into the opposite region as excess minority carriers. Since there is no lack of carriers for the current flow, the pn-junction is conducting in this bias condition. If the voltage at the p-region is negative with respect to the n-region, then both types of majority carriers are withdrawn from the junction and cannot be supplied from the adjacent region of opposite conductivity, except for the few equilibrium minority carriers there. Hence only a very small current can flow, the pn-junction is biased in the reverse or blocking direction.

3.1 The pn-Junction in Thermal Equilibrium

First we consider a pn-junction in thermodynamic equilibrium, the case of zero external voltage and current. The formulae for this case can be transferred to a large extent to the case of applied voltage and hence are useful also for the I-V

© Springer International Publishing AG 2018

J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8_3

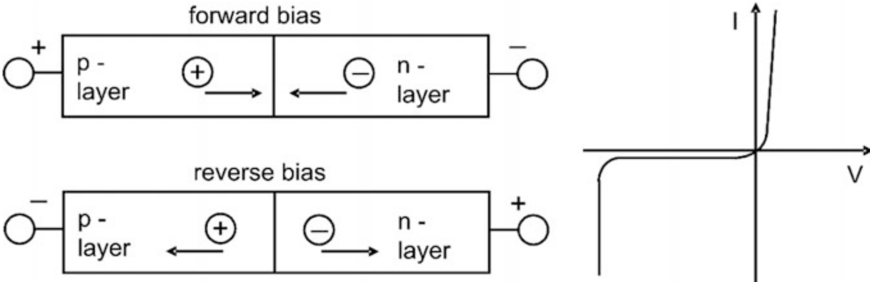


Fig. 3.1 pn-junction in forward- and in blocking direction

characteristics in forward and reverse direction. As discussed previously, the concentration of free holes in the bulk of the p-region is equal to the concentration of ionized acceptors, likewise deep in the n region the electron density is given by the donor concentration, thus the charge of the impurities is neutralized by the carriers. This is not the case, however, near the transition between the p and n region, as is indicated in Fig. 3.2. Here the hole density in the p region has a steep slope $-dp/dx$ resulting in a diffusion of holes towards the n-region. Since the fixed acceptors stay behind without compensation, a negative space charge arises in the p region near the junction. From the n-region, in the same way electrons diffuse towards the p region, so that a positive space charge of uncompensated donors remains in the n region near the junction. Between both space charges an electric field is built up, which drives the holes towards the p region and the electrons towards the n region, i.e. in both cases in opposite direction to the particle diffusion currents. The thermal equilibrium is reached when the field current compensates the diffusion current both for electrons and holes. The field over the space charge region results in a built-in voltage $V_{bi} = -\int_{x_p}^{x_n} \mathsf{E}\, dx$, where x_p, x_n are the boundaries of the space charge layer in the p and n region, respectively. The built-in potential holds the electrons in the n region and the holes in the p region. V_{bi} is often called also ‘diffusion voltage’, because its primary cause is diffusion.

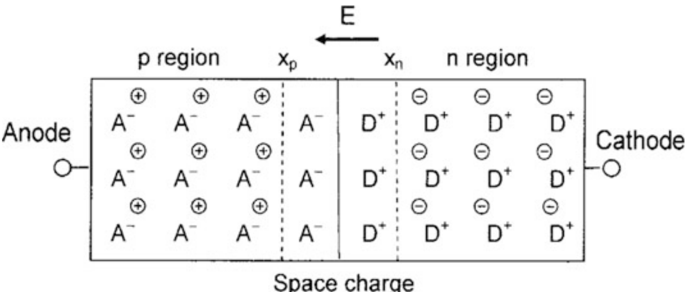


Fig. 3.2 pn-junction in thermal equilibrium

For the built-in voltage a simple general relationship can be derived. In Sect. 2.6, the Einstein relation (2.44) has been derived from the Boltzmann distribution (2.45). Inversely, using the Einstein relation in the current equations (2.43a), (2.43b)

$$j_n = q \cdot \mu_n \left(n \cdot \mathsf{E} + \frac{kT}{q} \frac{dn}{dx} \right) \tag{3.1}$$

$$j_p = q \cdot \mu_p \left(p \cdot \mathsf{E} - \frac{kT}{q} \frac{dp}{dx} \right) \tag{3.2}$$

and setting $j_p = 0$ in (3.2) one obtains the Boltzmann distribution by integrating the bracket term:

$$p(x) = p(x_p) \cdot e^{-\frac{q \cdot V(x)}{kT}}. \tag{3.3}$$

where the potential is given as $V(x) = -\int_{x_p}^{x} \mathsf{E}(x')dx'$. Similarly, one obtains from Eq. (3.1) for the electrons

$$n(x) = n(x_p) \cdot e^{\frac{q \cdot V(x)}{kT}} \tag{3.4}$$

By these equations the position dependent carrier densities in the space charge region are connected with the potential, which however is also not yet known as function of x. In Sect. 2.3 the relationship $n \cdot p = n_i^2$ has been derived for thermal equilibrium, without considering the presence of an electrical field. We see now that this relationship follows from (3.3), (3.4) at every point also in the space charge region where n and p are varying with x:

$$n(x) \cdot p(x) = n(x_p) \cdot p(x_p) = n_i^2 \tag{3.5}$$

The built-in voltage V_{bi} is obtained from (3.3) or (3.4) as $V_{bi} = V(x_n)$:

$$V_{bi} = \frac{kT}{q} \ln \frac{p(x_p)}{p(x_n)} = \frac{kT}{q} \ln \frac{p(x_p) \cdot n(x_n)}{n_i^2} \tag{3.6}$$

$$\cong \frac{kT}{q} \ln \frac{N_A(x_p) \cdot N_D(x_n)}{n_i^2} \tag{3.6a}$$

In the approximation (3.6a) complete ionization of the dopants is assumed together with neutrality at the boundaries of the space charge layer (see Sect. 2.5). According to this equation the built-in voltage is given by the doping densities at the boundaries of the space charge region.

The space dependence of the potential and carrier concentrations and the extension of the space charge layer into the p and n region can be calculated from the Poisson equation (2.111). Assuming as before complete ionization of the impurities and using (3.3), (3.4) one has for arbitrary doping profiles

$$\begin{aligned}\frac{d^2V}{dx^2} &= -\frac{\rho}{\varepsilon} = \frac{q}{\varepsilon}(n - p + N_{A,tot}(x) - N_{D,tot}(x)) \\ &= \frac{q}{\varepsilon}\left(n(x_p)\cdot e^{\frac{q\cdot V}{kT}} - p(x_p)\cdot e^{-\frac{q\cdot V}{kT}} - N(x)\right)\end{aligned} \tag{3.7}$$

where $N(x) \equiv N_{D,tot} - N_{A,tot}$. The donor and acceptor concentrations are indicated now additionally by a subscript '*tot*' to distinguish them from the net doping densities $N_A(x) = N_{A,tot}(x) - N_{D,tot}(x)$ in the p-region and $N_D(x) = N_{D,tot} - N_{A,tot}$ in the n-region which appear in Eq. (3.6a). The carrier concentrations at the p sided boundary of the space charge region are then $p_p(x_p) = N_A(x_p)$, $n(x_p) = n_i^2/N_A(x_p)$. Generally, the ordinary differential Eq. (3.7) has to be solved numerically, but for abrupt step junctions $V(x)$ can be expressed analytically by an integral (see Sect. 3.1.1). As will be seen, however, an approximate simpler calculation is often sufficient and even more useful. The exact solution of (3.7) will be used to correct the approximate formulae where it is necessary.

As follows from (3.6a) with $N_A(x_p)$, $N_D(x_n) >> n_i$, the built-in voltage is much larger than the thermal voltage kT/q. Hence a variation of the potential by a few kT/q at the boundaries of the space charge region towards the interior, connected with a decrease of the majority carrier concentrations towards zero, needs only a small distance. It seems reasonable therefore to neglect the carrier concentrations in the space charge region. Then (3.7) reduces to

$$\frac{d^2V}{dx^2} = -\frac{q}{\varepsilon}N(x) \tag{3.8}$$

This approach considering the space charge region as depleted from carriers is called depletion approximation, it will be used extensively in what follows.

3.1.1 The Abrupt Step Junction

The abrupt pn-junction or abrupt step junction is defined by a sharp, step-like doping transition between the p and n region and homogeneous doping within each of the two regions. Since N_A and N_D are independent of the positions x_p and x_n of the boundaries of the space charge layer, the built-in voltage is immediately known by (3.6a). In Fig. 3.3, the built-in voltage of abrupt pn-junctions in silicon is plotted for two temperatures versus the doping concentration N_D of the n region assuming a

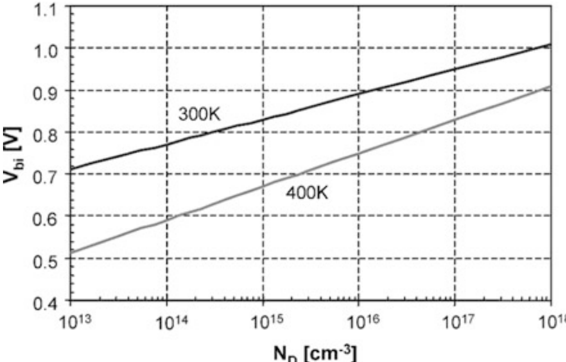


Fig. 3.3 Built-in voltage of abrupt pn-junctions in silicon as function of the doping concentration of the weakly doped side (N_D) for a fixed doping of 1×10^{19} cm^{-3} (N_A) of the highly doped region

fixed doping $N_A = 1 \times 10^{19}$ cm^{-3} of the p region. At 300 K, V_{bi} increases from 0.705 to 1.00 V in the range 1×10^{13}–1×10^{18} cm^{-3}, at 400 K from 0.508 to 0.905 V. The decrease of V_{bi} with increasing T is due to the strong increase of n_i. The decrease withtemperature is approximately linear, because the pre-exponential factor of n_i^2 in Eq. (2.6), $N_c \cdot N_v$, is larger than $N_D \cdot N_A$ in the application range of the formula and its logarithm nearly constant like E_g. At high doping concentrations both of the p and the n region the temperature dependence of V_{bi} is only weak.

We will now use the depletion approximation to calculate the potential and carrier concentrations in an abrupt junction as a function of x. Figure 3.4 illustrates the space dependence of (a) the doping and carrier concentrations, (b) the resulting charge density, (c) the electric field in the depletion approximation, (d) the potential and (e) the corresponding band diagram. These qualitative plots are substantiated now by the following calculations.

Placing the origin of x at the transition between the acceptor and donor doping, the metallurgical junction, the doping profile $N(x)$ is given by

$$\begin{aligned} N(x) &= -N_A = \text{const} \quad \text{for}\, x < 0, \\ N(x) &= +N_D = \text{const} \quad \text{for}\, x \geq 0 \end{aligned} \tag{3.9}$$

In the region with acceptor doping the Poisson equation

$$\frac{d^2V}{dx^2} = \frac{q}{\varepsilon} N_A \tag{3.10}$$

results in the field (see Fig. 3.4)

$$-\frac{dV}{dx} = \mathsf{E}(x) = -\frac{q}{\varepsilon} N_A \cdot (x - x_p), \quad \text{for}\, x_\text{p} \leq x \leq 0 \tag{3.11}$$

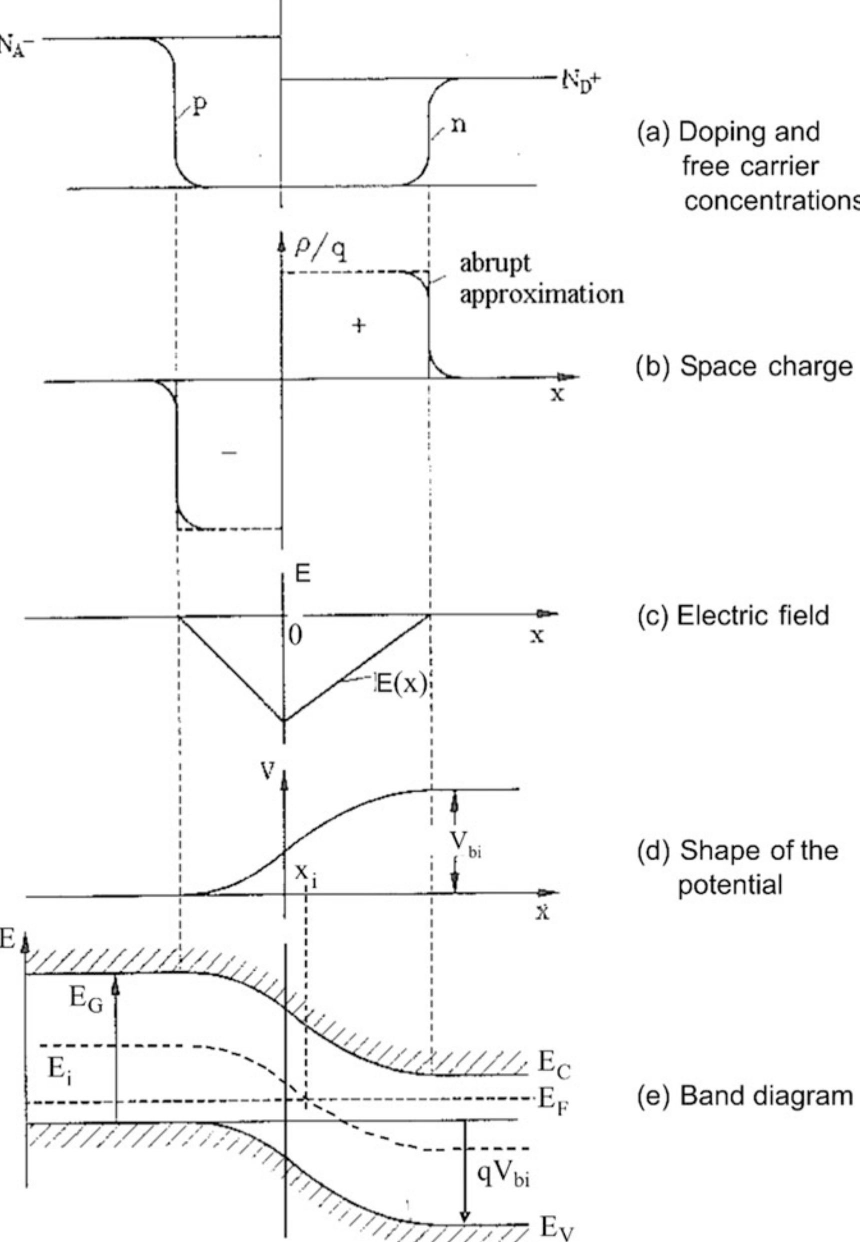


Fig. 3.4 Abrupt pn junction in the depletion approximation

For the metallurgical n-region, integration of the Poisson equation $d^2V/dx^2 = -q/\varepsilon N_D$ yields

$$-\frac{dV}{dx} = \mathsf{E}(x) = -\frac{q}{\varepsilon} N_D \cdot (x_n - x) \quad \text{for } 0 \leq x \leq x_\mathrm{n} \tag{3.12}$$

The continuity of E at $x = 0$ requires

$$N_A \cdot x_p = -N_D \cdot x_n \tag{3.13}$$

meaning that the charges on both sides of the junction are oppositely equal. Integration of (3.11), (3.12) leads to

$$V(x) = \frac{q}{2\varepsilon} \cdot N_A \cdot (x - x_p)^2 \quad \text{for } x_p < x \leq 0 \tag{3.14}$$

and

$$V(x) = \frac{q}{2\varepsilon}\left[-N_D \cdot (x_n - x)^2 + N_D x_n^2 + N_A x_p^2\right] \quad \text{for} \quad 0 \leq x < x_n \tag{3.15}$$

where the constant term in (3.15) is chosen to get continuity at $x = 0$. As is shown by (3.11), (3.12) and Fig. 3.4c, the field strength in each region is linear in x, the potential has a parabolic course. The ratio of the potential increase in the acceptor region to that in the donor region is obtained from (3.13) to (3.15) as

$$\frac{V_P}{V_N} = \frac{N_A \cdot x_p^2}{N_D \cdot x_n^2} = \frac{|x_p|}{x_n} = \frac{N_D}{N_A} \tag{3.16}$$

The penetration depths $-x_p$ and x_n can now be determined equating $V(x_n)$ to the built-in voltage as given by (3.6a). From (3.15), (3.13) one obtains

$$V_{bi} = \frac{q}{2\varepsilon} \cdot \left(N_D \cdot x_n^2 + N_A x_p^2\right) = \frac{q}{2\varepsilon} \cdot \left(N_D \cdot x_n^2 + \frac{N_D^2}{N_A} x_n^2\right) \tag{3.17}$$

$$x_n = \sqrt{\frac{2\varepsilon}{q} \cdot \frac{N_A/N_D}{(N_A + N_D)} \cdot V_{bi}}, \quad |x_p| = \frac{N_D}{N_A} \cdot x_n \tag{3.18}$$

The total thickness of the space charge layer is

$$w_{sc} = x_n + |x_p| = \sqrt{\frac{2\varepsilon}{q} \cdot \frac{N_A + N_D}{N_A N_D} \cdot V_{bi}} \tag{3.19}$$

The maximum absolute field strength $\mathsf{E}_m = |\mathsf{E}(0)|$ is obtained from (3.12), (3.18) as:

$$\mathsf{E}_m = \sqrt{\frac{2q}{\varepsilon} \cdot \frac{N_A N_D}{N_A + N_D} \cdot V_{bi}} = \frac{2 \cdot V_{bi}}{w_{sc}} \tag{3.20}$$

For asymmetric junctions with very different concentrations N_A, N_D, as found often in devices, the formulae simplify. Referring in the notation to a p$^+$n junction with $N_A \gg N_D$, Eq. (3.19) turns into

$$w_{sc} \approx x_n \approx \sqrt{\frac{2\varepsilon \cdot V_{bi}}{q \cdot N_D}}. \tag{3.19a}$$

whereas (3.20) yields

$$\mathsf{E}_m \approx \sqrt{\frac{2q}{\varepsilon} \cdot N_D \cdot V_{bi}} \tag{3.20a}$$

In the energy band diagram of Fig. 3.4e, the band edges vary inversely to the potential because the conduction band edge E_c represents the potential *energy* $-qV(x)$ of the electrons. As a general equilibrium condition the Fermi level E_F is constant across the pn-junction, which follows also from Eqs. (2.4) and (3.4). The built-in voltage multiplied with the elementary charge q is represented in the band diagram by the entire change of the band edges or of the intrinsic level E_i against the Fermi level. The point x_i where the intrinsic level crosses the Fermi level divides the region where $p > n$ from the region with $n > p$. This point differs generally from the metallurgical junction as is indicated in Fig. 3.4e. Instead of the above calculated voltage parts V_P, V_N in the metallurgical p and n region, the built-in voltage can be divided also into the potential difference $\Delta V_{p\text{-}i}$ between the neutral p region and the intrinsic point x_i on one side and the potential difference $\Delta V_{i\text{-}n}$ between the intrinsic point and the neutral n region on the other. Using (2.11), (2.10) these parts are obtained as

$$\begin{aligned} \Delta V_{p-i} &= \frac{1}{q}(E_i(x_p) - E_F) = \frac{kT}{q}\ln\frac{p(x_p)}{n_i} \cong \frac{kT}{q}\ln\frac{N_A}{n_i} \\ \Delta V_{i-n} &= \frac{1}{q}(E_F - E_i(x_n)) = \frac{kT}{q}\ln\frac{n(x_n)}{n_i} \cong \frac{kT}{q}\ln\frac{N_D}{n_i} \end{aligned} \tag{3.21}$$

If $N_A > N_D$ the voltage $\Delta V_{p\text{-}i}$ is larger than $\Delta V_{i\text{-}n}$, whereas $V_P < V_N$.

As numerical example we consider a pn-junction in silicon with $N_A = 2 \times 10^{15}$ cm^{-3}, $N_D = 1 \times 10^{15}$ cm^{-3}. The built-in voltage is in this case 0.604 V. Using this, Eqs. (3.18) with the permittivity of silicon $\varepsilon_r = 11.7$ yield the penetration depths $x_n = 0.725$ µm, $|x_p| = 0.363$ µm. The maximum value of the field according to (3.20) is $\mathsf{E}_m = 1.12 \times 10^4$ V/cm. The parts of the built-in voltage in the metallurgical acceptor and donor region are $V_P = N_D/(N_A + N_D) \cdot V_{bi} = V_{bi}/3$

= 0.201 V, V_N = 2 · V_P = 0.402 V. In contrast to this, Eq. (3.20) yields for the voltage increase in the region with $p > n_i$ a value $\Delta V_{p\text{-}I}$ = 0.311 V, while the voltage in the region with $n > n_i$ is $\Delta V_{i\text{-}n}$ = 0.296 V. The intrinsic point is shifted into the region with lower doping. Numerically (3.15) together with (3.21) yields x_i = 0.106 µm.

As will be seen, Eqs. (3.19a), (3.20a) have only a very limited applicability just for well abrupt asymmetrical junctions, whereas for diffused junctions they are very useful. Their accuracy can be tested by comparing with results following from the exact Poisson Eq. (3.7). Integrating 3.7 once,[1] the following relationship between the potential parts V_P, V_N in the acceptor and donor doped region, respectively, is obtained instead of (3.16):

$$\frac{V_P - kT/q}{V_N - kT/q} = \frac{N_D}{N_A} \tag{3.22}$$

Together with $V_P + V_N = V_{bi}$ it follows explicitly

$$\begin{aligned} V_P &= \frac{N_A - N_D}{N_A + N_D} \cdot \frac{kT}{q} + \frac{N_D}{N_A + N_D} \cdot V_{bi} \\ V_N &= V_{bi} - V_P \end{aligned} \tag{3.23}$$

For the maximal field the exact calculation yields:

$$\mathsf{E}_m = \sqrt{\frac{2kT}{\varepsilon} \cdot \left\{ N_A \cdot e^{-\frac{qV_P}{kT}} + N_D \cdot \left(\frac{qV_N}{kT} + e^{-\frac{qV_N}{kT}} - 1 \right) \right\}} \tag{3.24}$$

Equations (3.22)–(3.24) hold for arbitrary concentrations N_A, N_D. If V_P, $V_N >> kT/q$, (3.22) turns into (3.16), and in (3.24) only the linear term in V_N remains which with (3.23) leads to (3.20). For the doping concentrations this requires that N_A/N_D as well as N_D/N_A are small against qV_{bi}/kT. Since on the other hand $N_A/N_D \gg 1$ is required for (3.19a), (3.20a), these equations are roughly valid only in a small range around $N_A/N_D = 5$ in Si. In contrast to (3.20a), Eq. (3.24) yields an increase of E_m with N_A for constant N_D. For a *very* asymmetrical p^+n junction defined by $N_A/N_D \gg qV_{bi}/kT$ the voltage V_P tends to kT/q according to (3.23) (instead of zero after (3.16)), and now the N_D-term in (3.24) is negligible. Hence (3.24) turns into

$$\mathsf{E}_m = \sqrt{\frac{2kT}{\varepsilon} N_A / e} \quad \text{for } N_A/N_D \gg qV_{bi}/kT \tag{3.24a}$$

[1] After multiplication with $2 \cdot dV/dx$, Eq. (3.7) can be integrated analytically to obtain the field $\mathsf{E}(x)$ as function of $V(x)$ in both regions. From the continuity of potential and field at the metallurgical junction the potential parts V_P, V_N are then determined.

While according to the 'classical' formula (3.20a) the maximum field is determined by the doping concentration of the weakly doped region (and the built-in voltage), Eq. (3.24a) says that it is given only by the doping density of the highly doped region. Also numerically the results are extremely different: For the example $N_A = 5 \times 10^{18}$ cm^{-3}, $N_D = 1 \times 10^{14}$ cm^{-3} in silicon, one obtains from (3.20a) $\mathsf{E}_m = 4.82 \times 10^3$ V/cm, whereas Eq. (3.24a) and also (3.24) result in $\mathsf{E}_m = 1.21 \times 10^5$ V/cm.

The cause of this strong discrepancy becomes evident from Fig. 3.5, which shows the carrier distributions in a p$^+$n junction as calculated by twice integration of Eq. (3.7). It is seen that the hole concentration of the acceptor doping is not restricted to the N_A region but reaches beyond the metallurgical junction into the N_D region. In the first range of the latter the hole concentration is much higher than the doping density N_D and even the integrated hole charge in the N_D region is large compared with the integral charge of ionized donors in the space charge region. This explains why the field at the metallurgical junction is independent of N_D according to (3.24a). As will be shown in Sect. 3.5, the charge of mobile carriers in the space charge region has a strong influence on the capacitance of the junction.

The assumed exact abruptness is realized mostly by pn junctions in wide-gap semiconductors like SiC, where the doping profiles are not smoothed out by diffusion. In silicon, pn-junctions made by low-temperature epitaxy are very abrupt. For diffused junctions Eq. (3.24) is usually not applicable, because the doping concentration of the highly doped region decreases more slowly than the carrier concentration. The positive charge in the space charge region of a p$^+$n junction is then given solely by the donor doping. Hence Eqs. (3.19a), (3.20a) are approximately applicable for a wide class of diffused p$^+$n junctions, whereas they are not valid for abrupt junctions with $N_A/N_D \gtrsim qV_{bi}/kT$. This holds especially in the extended form which includes an applied voltage (see Sect. 3.3).

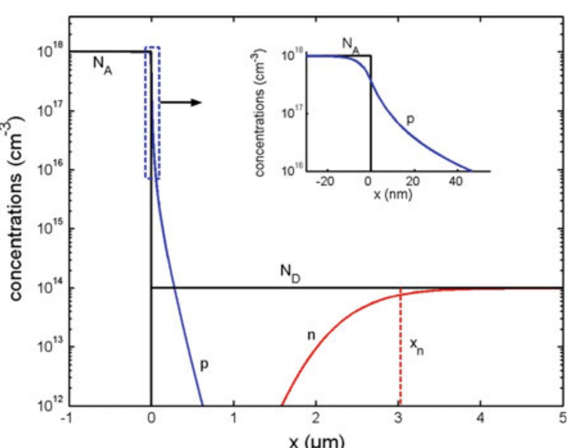


Fig. 3.5 Carrier distribution in an abrupt p$^+$n-junction. x_n is the boundary of the space charge region in the depletion approximation

3.1.2 Graded Junctions

Often the transition between the acceptor and donor doping is not abrupt, but occurs with a rather low gradient which then becomes significant for the properties of the junction. Particularly this is the case for pn-junctions produced by deep diffusion of impurities (for details of this technique see Sect. 4.4). If the diffused dopant is of the opposite type and has a higher surface concentration than the doping of the wafer, a pn-junction is formed at the point where the diffused impurity just compensates the wafer doping. The net doping density, $N(x) = N_D - N_A(x)$ for an acceptor diffusion, changes its sign at the junction. If the gradient of $N(x)$ is not too small, a significant space charge builds up as for abrupt junctions. In Fig. 3.6 the profile of a diffused acceptor with surface concentration 1×10^{18} cm^{-3} added to a homogeneous donor doping with density 1×10^{14} cm^{-3} is shown together with the resulting absolute net doping density and the electron and hole concentration calculated numerically for silicon. Near the pn-junction, the carrier concentrations are much smaller than the net doping density, why the space charge there is a large. The net doping density shown on an expanded linear scale in the inset varies nearly over the whole space charge layer. The depletion approach neglecting the carrier charge in the space charge region can be used also in this case as an approximate basis. Sufficiently near to the junction, the net doping concentration can be approximated by a linear dependence

$$N(x) = a \cdot x, \tag{3.25}$$

using the junction as origin of x. Assuming that this holds over the whole space charge region one obtains the *linearly graded junction* model which seems to be applicable to such doping profiles rather than the abrupt junction model. Under these conditions one obtains from (3.8), (3.25)

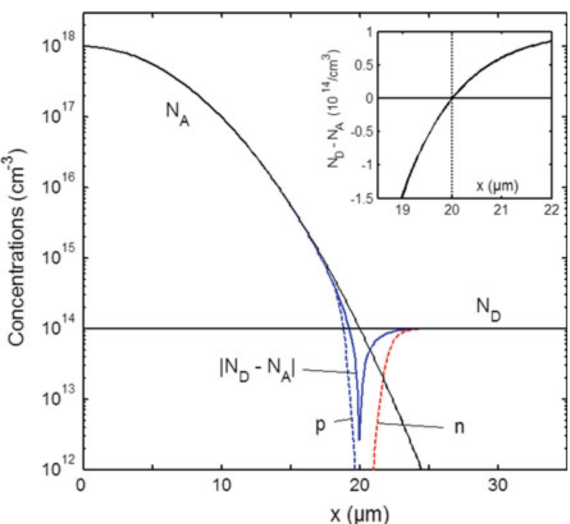


Fig. 3.6 Diffused p$^+$n-junction in silicon: Doping profile together with hole and electron distribution. A Gaussian function, $N_A \propto \exp(-(x/L_A)^2)$, is assumed as diffusion profile

$$\frac{d^2V}{dx^2} = -\frac{d\mathsf{E}}{dx} = -\frac{q \cdot a}{\varepsilon} \cdot x \tag{3.26}$$

$$\mathsf{E}(x) = \frac{q \cdot a}{2\,\varepsilon}\left(x^2 - w^2\right) \tag{3.27}$$

$$V(x) = \frac{q \cdot a}{2\,\varepsilon} \cdot \left(\left(w^2 \cdot x - \frac{1}{3} \cdot x^3\right)\right. \tag{3.28}$$

where w is the extension of the space charge layer in each of the two regions (half width of the total space charge layer) and the potential at the junction is set to zero.

To calculate the built-in voltage $V_{bi} = V(w) - V(-w)$ one has from (3.28) the following connection between w and V_{bi}:

$$w = \left(\frac{3\,\varepsilon\,V_{bi}}{2\,q\,a}\right)^{1/3} \tag{3.29}$$

On the other hand the built-in voltage is given by (3.6a):

$$\begin{aligned} V_{bi} &= \frac{kT}{q} \cdot \ln\left(\frac{-N(-w) \cdot N(w)}{n_i^2}\right) = \frac{kT}{q}\ln\left(\frac{(a \cdot w)^2}{n_i^2}\right) \\ &= 2\frac{kT}{q}\ln\left(\frac{a \cdot w}{n_i}\right) \end{aligned} \tag{3.30}$$

Inserting (3.29) one obtains an implicit equation from which V_{bi} can be determined by iteration as a function of the doping gradient a. The result for silicon at 300 and 400 K is shown in Fig. 3.7. By comparing with the exact numerical solution of (3.7) it is found [Mol64, Mor60] that the depletion approach assuming abrupt transition between complete depletion and neutrality is in this case only a rough approximation applicable mainly for impurity gradients

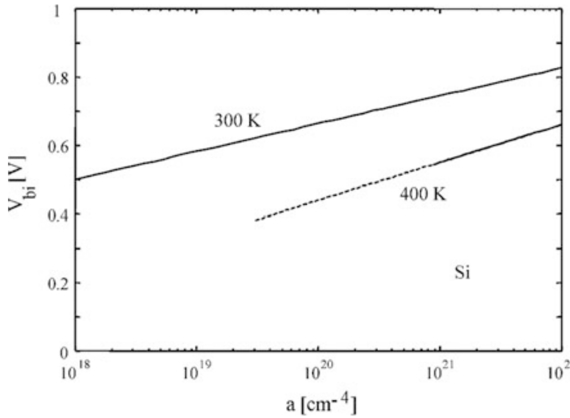


Fig. 3.7 Built-in voltage of a linearly graded junction in Si as function of the doping gradient

$$a \gtrsim 10^4 \frac{2n_i}{L_D} \tag{3.31}$$

where $L_D = \sqrt{\varepsilon kT/(2n_i q^2)}$ is the intrinsic Debye length. For silicon, the right hand side amounts to 7.6×10^{16} cm^{-3} at 300 K, but 8.8×10^{20} cm^{-3} at 400 K. Below this value, the line for 400 K in Fig. 3.7 may be somewhat inexact. The approach can be generalized to the case of an applied external voltage and is used to calculate the capacitance of graded junctions. For these topics we refer to [Mor60, Mol64].

Although far away from a junction and for a doping concentration large against n_i the semiconductor is essentially neutral, a spatially varying doping concentration produces a built-in field and built-in potential also here. Since the electron and hole currents are zero in thermal equilibrium, a field is necessary to compensate the diffusion current arising from the concentration gradient. From Eq. (3.2) one obtains replacing the hole density by the net acceptor concentration

$$\mathsf{E}(x) = \frac{kT}{q}\frac{d\,\ln p}{dx} = \frac{kT}{q}\frac{d\,\ln N_A}{dx} \tag{3.32}$$

$$V(x) - V(x_0) = -\frac{kT}{q}\ln\frac{N_A(x)}{N_A(x_0)} \tag{3.33}$$

For n-regions the sign is opposite. For an exponential profile $N(x) \sim e^{x/\lambda}$ the assumed neutrality holds exactly, since the built-in field is constant according to (3.32) and hence the space charge $\rho = \varepsilon\, dE/dx = 0$. The built-in potential of (quasi-) neutral diffused regions is mostly quite appreciable. In the case of Fig. 3.6, the potential difference between the surface of the p-region and the boundary of the space charge region, where $N_A = 3 \times 10^{14}$ cm^{-3}, amounts to 0.28 V at 400 K according to (3.33). If also a diffused n-region follows on the homogeneous n-base as in a pin diode, the whole built-in potential of the quasi-neutral regions comes near to that of the space charge region. For device characteristics, however, the built-in potential of the neutral regions is little significant, because it stays (nearly) constant if an external voltage is applied, whereas the voltage across the space charge region varies essentially, as will be seen. If not indicated else, the term “built-in voltage” refers therefore always to that of the space charge region.

The question arises, why does the built-in voltage not cause a current flow, if the *p* and *n* region are connected externally by a wire. From the constancy of the Fermi level throughout the contacted structure in thermal equilibrium, it follows that contact potentials between the contacted semiconductor regions and the metal exist, and these are in sum oppositely equal to the whole built-in voltage in the semiconductor.

As one of the results of the present section we mention again that the extension of the space charge layer in thermal equilibrium is very small compared with the usual thickness of the (quasi-)neutral p and n regions enclosing it. If an external voltage is applied, the situation can be changed.

3.2 Current-Voltage-Characteristics of the pn-Junction

Now a voltage V is applied to the p region with respect to the n region. If this voltage is positive, it is directed against the built-in voltage V_{bi}. Assuming that the caused current is small and the ohmic voltage drop over the neutral regions can be neglected, the voltage across the space charge region is now

$$\Delta V = V_{bi} - V \tag{3.34}$$

For building up this voltage step, the required charges in the dipole layer of the junction are smaller or larger than without the external voltage depending on the sign of the voltage. Since the charge density on each side is given approximately by the doping density, the thicknesses of the space charge layer in the p and n region decrease or increase. Most important, however, is that the hole concentration in the n region and electron concentration in the p region are raised above the equilibrium minority densities for $V > 0$ and lowered for $V < 0$. This can be concluded from Eqs. (3.3), (3.4) *assuming that the Boltzmann distribution is applicable also to cases away from thermal equilibrium*. This is a basic assumption of the whole device theory, justified by the fact that the deviation from equilibrium is usually weak, i.e. the field and diffusion currents compensate each other largely in the space charge region.

Explicating this, we assume the pn-junction as abrupt and the minority carrier concentrations in the neutral p and n region to be small compared with the doping concentrations. Hence from neutrality one has furthermore $p(x_p) = N_A, n(x_n) = N_D$. Replacing $V(x_n)$ in Eq. (3.3) by (3.34) one obtains then for the hole concentration p_n^* in the neutral n region at the boundary $x = x_n$ to the space charge region

$$p_n^* = N_A \cdot e^{-\frac{q \cdot (V_{bi} - V)}{kT}} \tag{3.35}$$

$$= p_{n0} \cdot e^{\frac{q \cdot V}{kT}} \tag{3.36}$$

Here the equilibrium hole density in the n region is denoted by p_{n0}, and the relationship

$$p_{n0} = N_A e^{-\frac{qV_{bi}}{kT}} = \frac{n_i^2}{N_D} \tag{3.37}$$

was used, which follows from (3.6). For (3.35), (3.36), we have not yet used, that the injection level in the n region is low. This is proposed however now to obtain the electron concentration n_p^* in the neutral p region at the boundary $x = x_p$ to the space charge region. Similarly as (3.36) one obtains:

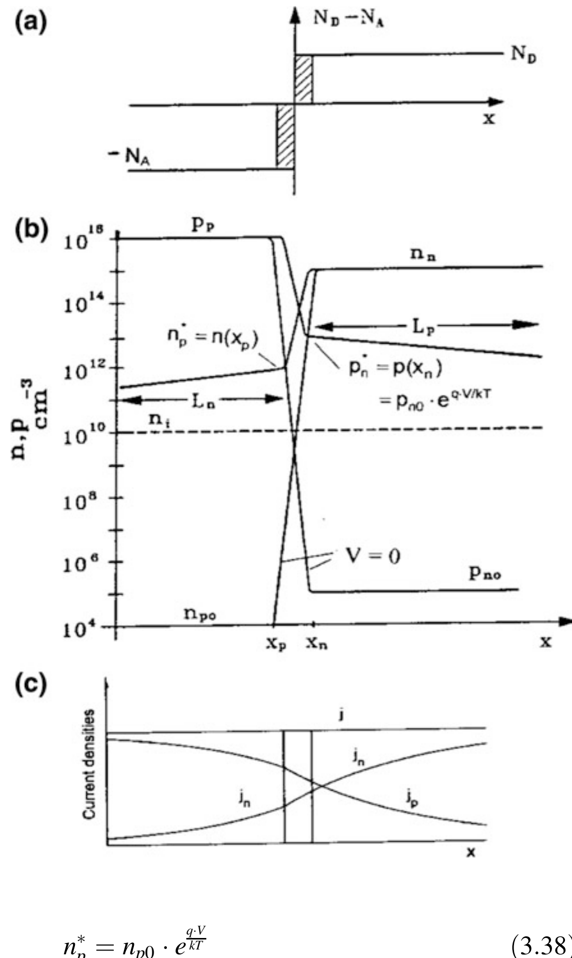


Fig. 3.8 Forward biased pn-junction. **a** Net doping density, **b** Carrier distribution for $V > 0$ and $V = 0$, **c** Hole and electron current densities

$$n_p^* = n_{p0} \cdot e^{\frac{q \cdot V}{kT}} \tag{3.38}$$

where n_{p0} is the equilibrium electron concentration in the neutral p region. The minority concentrations at the boundaries to the space charge region are raised respectively lowered by an exponential factor containing the applied voltage V, the Boltzmann factor. In Fig. 3.8 this is illustrated for forward bias. The applied voltage raises the concentrations p_n* and n_p* here by 8 orders of magnitude. On the chosen logarithmic scale the minority carrier concentrations decrease linearly with distance from the space charge layer. The figure anticipates here results of the following calculation.

For further visualization especially of the situation at reverse bias, the minority carrier densities are plotted in Fig. 3.9 on a linear scale. In the picture for reverse bias (lower part of the figure) the minority carrier concentrations at the boundaries to the space charge region are lowered already to zero. This is approached already for reverse voltages higher than a few times the thermal voltage kT/q, as follows

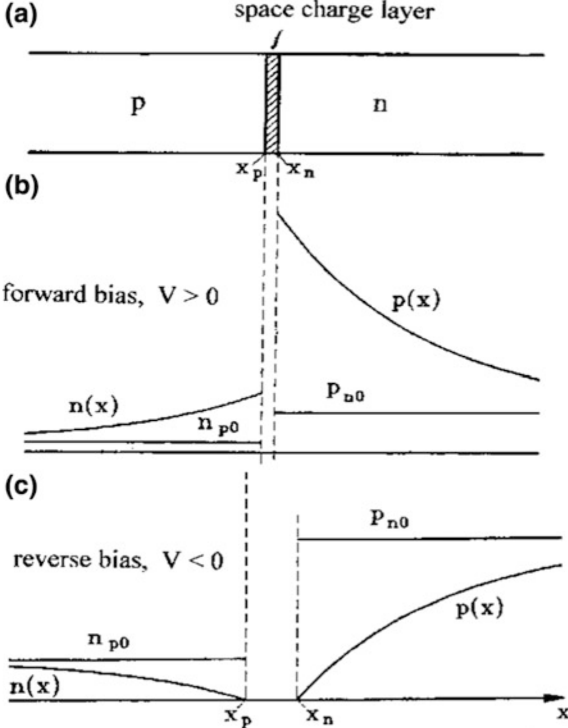


Fig. 3.9 Minority carrier distributions in a pn-junction (**a**) under forward bias (**b**) and reverse bias (**c**) on a linear scale

from (3.36), (3.38). The diffusion of holes out of the n-region and electrons out of the p-region which determines the reverse current can then no longer be enhanced by a further increase of the reverse voltage. Because the equilibrium densities $p_{n0} = n_i^2/N_D$, $n_{p0} = n_i^2/N_A$ are very small and this transfers to the concentration gradients of the minority carriers, the blocking current is also very small.

For the np-product in the space charge region one obtains from Eqs. (3.3), (3.4) together with (3.38)

$$n(x) \cdot p(x) = n(x_p) \cdot p(x_p) = n_{p0} \cdot e^{\frac{qV}{kT}} \cdot p_0 = n_i^2 \cdot e^{\frac{q \cdot V}{kT}} \tag{3.39}$$

As without bias, the np-product in the space charge layer is independent of x, but it is increased or decreased depending on the sign of the voltage V by the exponential voltage factor. The deviation from equilibrium leads to net recombination or generation, respectively.

The I-V characteristic is governed by the minority carrier currents in the neutral regions. To calculate them one uses the continuity equation which for the holes in the n-region according to Eq. (2.103) reads

$$\frac{dj_p}{dx} = -q \cdot R_p = -q \cdot \frac{p - p_{n0}}{\tau_p} \tag{3.40}$$

Here the stationary case is assumed and the excess recombination rate R_p is expressed, according to (2.49), by the minority carrier lifetime and the excess hole concentration. Since the hole density p is assumed small compared with n (low injection level), and additionally the field in the neutral region is small, the field term in the current Eq. (2.43b) can be neglected. Hence we are dealing with the case b of Sect. 2.10.1. Inserting

$$j_p = -q \cdot D_p \frac{dp}{dx} \tag{3.41}$$

into Eq. (3.40) one obtains

$$D_p \cdot \frac{d^2 p}{dx^2} = \frac{p - p_{n0}}{\tau_p} \tag{3.42}$$

The solution of this differential equation with the boundary condition $p(x_n) = p_n^*$ is

$$p(x) - p_{n0} = (p_n^* - p_{n0}) \cdot e^{-\frac{x - x_n}{L_p}} \tag{3.43}$$

where L_p is the hole diffusion length:

$$L_p = \sqrt{D_p \cdot \tau_p} \tag{3.44}$$

Inserting p_n^* from (3.36) one has:

$$p(x) - p_{n0} = p_{n0}\left(e^{\frac{qV}{kT}} - 1\right) \cdot e^{-\frac{x - x_n}{L_p}} \tag{3.45}$$

Using this hole distribution in (3.41) the hole current density at $x = x_n$ is obtained as

$$j_p(x_n) = j_{ps} \cdot \left(e^{\frac{q \cdot V}{kT}} - 1\right) \tag{3.46}$$

with

$$j_{ps} = q \cdot p_{n0} \cdot \frac{D_p}{L_p} = q \cdot \frac{n_i^2}{N_D} \cdot \frac{D_p}{L_p} \tag{3.46a}$$

Analogously the electron current density in the p region at $x = x_p$ is given by

$$j_n(x_p) = j_{ns} \cdot \left(e^{\frac{q \cdot V}{kT}} - 1\right) \tag{3.47}$$

with

$$j_{ns} = q \cdot n_{p0} \cdot \frac{D_n}{L_n} = q \cdot \frac{n_i^2}{N_A} \cdot \frac{D_n}{L_n} \tag{3.47a}$$

where L_n is the electron diffusion length:

$$L_n = \sqrt{D_n \cdot \tau_n} \tag{3.48}$$

As is seen, the minority carrier diffusion currents adopt the exponential voltage dependence of the minority carrier densities p_n*, n_p*. Apart from these quantities, the currents depend on the diffusion constants and the diffusion lengths L_p, L_n which are determined by the respective minority carrier lifetime in the neutral regions.

Whereas the recombination in the neutral regions is considered to be essential, *the recombination/generation in the thin space charge layer is neglected* in the ideal I–V characteristic. Since $dj_{n,p}/dx = \pm q \cdot R$, this means that the electron and hole currents are assumed constant across the space charge layer. With $j_n(x_n) = j_n(x_p)$ one obtains

$$j = j_n(x_n) + j_p(x_n) = j_n(x_p) + j_p(x_n) \tag{3.49}$$

Thus, the current density j is given by the sum of the minority carrier diffusion currents in the neutral regions at the borders to the space charge layer, by adding (3.46) and (3.47), the current-voltage-characteristics of the pn-junction is obtained as

$$j = j_s \cdot \left(e^{\frac{q \cdot V}{k \cdot T}} - 1 \right) \tag{3.50}$$

with

$$j_s = q \cdot n_i^2 \cdot \left(\frac{D_p}{L_p \cdot N_D} + \frac{D_n}{L_n \cdot N_A} \right) \tag{3.51}$$

Equation (3.50) with (3.51) is the ideal current-voltage-characteristic of the pn-junction as derived by Shockley [Sho49]. In Fig. 3.10 the characteristic is shown in the normalized form j/j_s versus qV/kT. The current increases exponentially with positive voltage, in the blocking direction it approaches quickly the saturation current which is very small. If for example

$$\mathrm{N_A} = 1 \times 10^{16}\,\mathrm{cm}^{-3}, \mathrm{N_D} = 1 \times 10^{15}\,\mathrm{cm}^{-3},$$
$$\mathrm{L_n} = \mathrm{L_p} = 50\,\mu\mathrm{m}, \mathrm{D_n} = 30\,\mathrm{cm}^2/\mathrm{s}, \mathrm{D_p} = 12\,\mathrm{cm}^2/\mathrm{s}$$

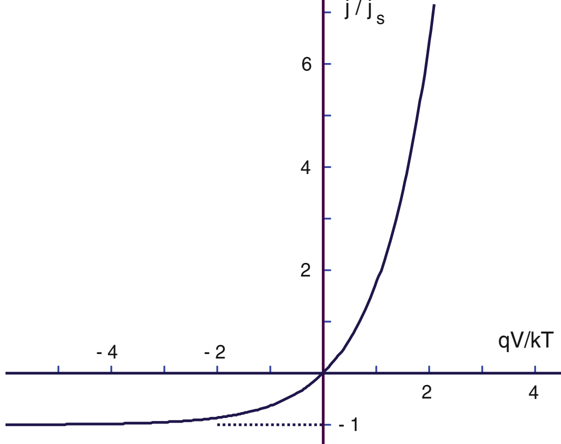


Fig. 3.10 Normalized ideal I-V characteristic of a pn-junction

one obtains for silicon with n_i (300 K) = 1.07×10^{10} cm^{-3}: $j_s = 5.5 \times 10^{-11}$ A/cm^2. The characteristic (3.50) in this general form has been derived first for the Cu/CuO_2 rectifier by Wagner [Wag31]. Also the later theory of the metal-semiconductor contact leads to the characteristic (3.50), the saturation current however differs essentially from (3.51). The exponential dependence of current on voltage is always based on the Boltzmann distribution.

In (3.51) the factor n_i^2 finally contains the dependence on the band gap and the main part of temperature dependence of j_s. In Fig. 3.11 the characteristics for several semiconductors at 300 K as given by (3.50), (3.51) are plotted on a larger scale than in Fig. 3.10, the reverse current is scaled logarithmically. For all semiconductors the same values of N_A, N_D and $D_{n,p}/L_{n,p}$ were used as in the above example for silicon. The saturation current density varies over many orders of magnitude. Up to a certain forward voltage, about 0.7 V for Si, the current remains very small on the scale of normal conduction current densities, whereas above that voltage it increases soon strongly. This threshold voltage can be defined with some arbitrariness as the voltage belonging to a current density j_{thr} of about 5 A/cm^2. Solving (3.50) for V

$$V = \frac{k \cdot T}{q} \ln\left(\frac{j}{j_s} + 1\right) \tag{3.52}$$

and inserting $j = 5$ A/cm^2 the threshold voltage V_{thr} in the case of Fig. 3.11 is obtained to be 0.26 V for Ge, 0.65 V for Si,[2] 1.09 V for GaAs and 2.77 V for 4H-SiC. These values are close to the built-in voltage V_{bi} calculated by Eq. (3.6a).

[2]The threshold voltage given in data sheets for power diodes is about 0.2–0.5 V higher. Because these diodes have a pin structure (see Sect. 5), a voltage drop across the base (i) region and at the i-n junction have to be added. As threshold voltage one uses there the onset voltage V_s in the simplified ohmic characteristic $V_F = V_s + R_{diff} I_F$.

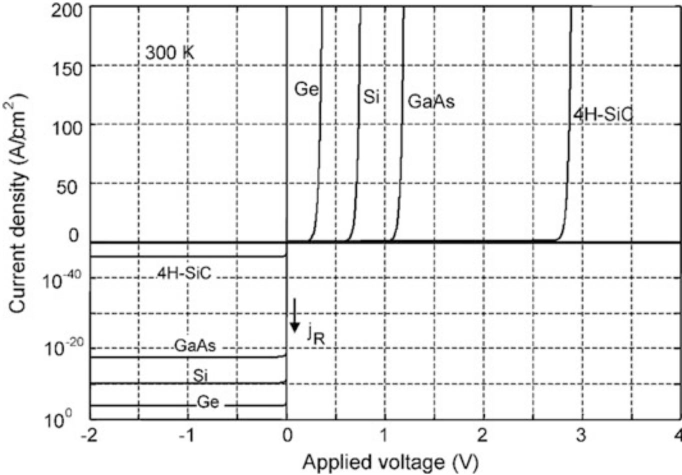


Fig. 3.11 Current density versus applied voltage for ideal pn-junctions in different semiconductors at 300 K

The threshold current, although near the lower end of normally used current densities, is already so high that the applied voltage must balance already largely the built-in voltage. The threshold voltage increases nearly linearly with band gap as follows from (3.52) if Eq. (2.6) is inserted. Experiments are well reproduced on the whole by this theory.

Of course a small threshold voltage is advantageous because of the forward losses. On the other hand, the high j_s required for small V_{thr} means a high reverse leakage current. In the case of Ge this disadvantage outweighs strongly the advantage of small threshold voltage. Already at 100 °C the leakage current grows so high that it results in a hardly controllable heating. For wide-gap semiconductors on the other hand, the large threshold voltage is the more weighing disadvantage. Here, the large threshold voltage of pn-junctions is often avoided by using a metal-semiconductor junction and for switching a unipolar field effect transistor without junction in the current path.

In practical devices, the pn-junctions are mostly very asymmetric. With $N_A \gg N_D$ Eq. (3.51) simplifies to

$$j_s = q \cdot n_i^2 \cdot \left(\frac{D_p}{L_p \cdot N_D} \right) \tag{3.53}$$

The saturation current is determined in these cases only by the minority carrier parameters of the weakly doped zone. Using this equation, we consider now the *temperature dependence* of the voltage V at a given forward current, which is often used to indicate and control the temperature in integrated power devices. Inserting Eq. (2.6) for n_i^2 one obtains from (3.53), (3.52) for $V > 3 \cdot kT/q$:

$$V(j,T) = \frac{E_g}{q} - \frac{kT}{q} \cdot \ln\left(\frac{q \cdot D_p \cdot N_C \cdot N_V}{L_p \cdot N_D \cdot j}\right) \tag{3.54}$$

Because the term $q \cdot D_p \cdot N_C \cdot N_V/(L_p \cdot N_D)$ represents a very high current density whose temperature dependence due to the logarithm has little influence, the forward voltage V decreases nearly linearly with increasing T. In Fig. 3.12, the dependency (3.54) with use of Eqs. (2.8) and (2.9) is plotted for two current densities for the example: $N_D = 1 \times 10^{16}$ cm^{-3} and $\tau_p(T) = 1\ \mu s \cdot (T/300)^2$. This temperature dependence is found often and leads to $D_p/L_p = \sqrt{(D_p/\tau_p)} \approx 3.3 \times 10^3\ (300/T)^{3/2}$ cm/s.

Regarding the conditions underlying the Shockley characteristic (3.50), (3.51), the assumption of negligible recombination in the space charge region is sometimes not satisfied in silicon even at forward bias where the thickness of the space charge layer is very small. This can be seen, if at small forward current densities, for example in the range of 100 js to 1 A/cm^2, the measured current is plotted on a logarithmic scale versus voltage. The observed slope is often not q/kT as required by (3.50) but essentially smaller. The measured characteristic can be described then by

$$j = j_s \cdot \left(e^{\frac{q \cdot V}{n \cdot k \cdot T}} - 1\right) \tag{3.55}$$

where the number n ranges often between 1.7–2. Because of the small current densities this cannot be attributed to a resistive voltage drop in the p and n region. The recombination in the space charge layer is given by the SRH Eq. (2.68) with $n \cdot p = const \gg n_i^2$ at forward bias [see Eq. (3.39)]. The nominator of this equation has approximately the same magnitude as in the neighboring parts of the neutral regions. However, if the recombination level lies near the middle of a not too small band gap and hence the concentrations n_r and p_r both are small, the denominator in (2.68) is on average small against the value in the neutral regions where n or p is large. Hence the recombination rate R is very high, and in spite of the small

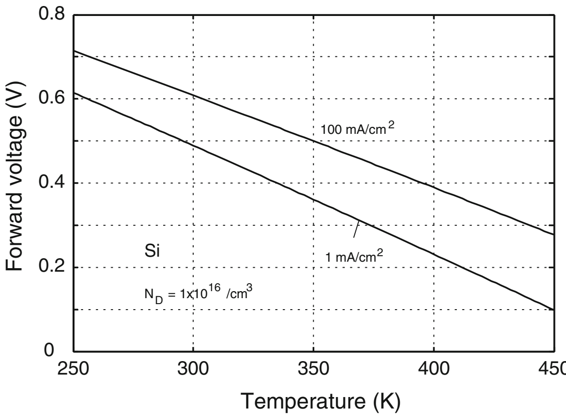


Fig. 3.12 Temperature dependence of forward voltage of a p$^+$n-junction at constant current densities

thickness of the space charge layer the integral recombination $\int R dx$ can be significant. Only if the recombination level is located in considerable distance from the middle of the band gap, the recombination in the space charge layer is negligible, and the ideal characteristic is measured.

At current densities, where the resistive voltage drop V_{drift} caused by the majority carrier currents in the p and n region is considerable, the voltage V in the above equations has to be interpreted as the junction voltage $V_j = V_F - V_{drift}$ where V_F is the total forward voltage. As long as the resistance is given by the doping concentration and thicknesses of the two semiconductor regions and hence is a constant, V_F as function of j is obtained by adding the drift voltage $V_{drift} = r \cdot j$ (r = resistance times area) to the right hand side of Eq. (3.52):

$$V_F = \frac{kT}{q} \cdot \left(\ln \frac{j}{j_s} + 1 \right) + r \cdot j$$

A current independent resistance is tied however to the condition of low injection levels, which was assumed also for the derivation of Eq. (3.50). This condition is satisfied over a wide range of current densities only for relatively high doping concentrations of the p and n region. For power diodes with a p^+nn^+ structure, the resistance of the weakly doped base region is strongly modulated at usual forward current densities by high injection of carriers. The characteristics of power diodes are treated in Chap. 5.

3.3 Blocking Characteristics and Breakdown of the pn-Junction

3.3.1 Blocking Current

Especially for reverse voltages, the Eqs. (3.50), (3.51) agree with measurements only in limited ranges of voltage and temperature if at all. Often the measured blocking current is much higher than the saturation current (3.51) because of large generation in the space charge region. A second effect not included in (3.50), (3.51) is avalanche multiplication which at high reverse voltages leads to enhanced blocking current and at a certain voltage to complete breakdown of the blocking ability. This effect will be treated in the next section, in the present section we consider the blocking current as made up of the saturation current (3.51) and the current caused by generation in the space charge region.

Since a reverse voltage results in a negative excess carrier concentration (deviation from equilibrium), the thermal generation rate $G(x)$ is everywhere ≥ 0. The total reverse current density j_r owing to thermal generation is obtained by integrating the continuity equation $dj_p/dx = -q \cdot R = q \cdot G$ from the left border of the p region ($x = -\infty$) to the right border of the n region ($+\infty$). This yields:

$$q\int_{-\infty}^{\infty} G dx = j_p(\infty) - j_p(-\infty) = -j_p(-\infty) = -j = j_r, \tag{3.56}$$

since the hole current in the p region at large distance from the junction equals the total current. For the calculation, one uses as before the abrupt depletion approximation with $n \approx p \approx 0$ in the space charge region and with the minority carrier distribution in the neutral regions as given for the holes in the n region by (3.45), where now $V < 0$ (see Fig. 3.9c). The integral over the neutral regions delivers the reverse current of (3.50), (3.51), as can be easily verified: The minority carriers diffusing out of the neutral regions are *generated* there due to the negative excess concentration. For the current caused by generation in the *space charge region,* the formula (2.68) applies. In order that the terms with n and p can be neglected on the base of (3.39) and (3.15), (3.18), the reverse voltage V_r is assumed to be $>3 \cdot kT/q$. The generation rate is then constant and given by (2.77), and the reverse current generated in the space region (boundaries x_p and x_n, thickness $w = x_n - x_p$) is

$$j_{sc} = q \cdot \int_{x_p}^{x_n} G dx = \frac{q \cdot n_i \cdot w}{\tau_g} = \frac{q \cdot N_r \cdot w}{1/e_n + 1/e_p} \tag{3.57}$$

Since the charge density on both sides of the junction is given by the doping concentration in the used approximation and is therefore independent of the voltage, the extension of the space charge region into the p and n region (and other relationships) are given by the same expressions as for zero bias except that V_{bi} has to be replaced by $V_{sc} = V_{bi} - V = V_{bi} + V_r$. Hence Eq. (3.19a) for an abrupt p^{+}-n-junction generalizes to

$$w = \sqrt{\frac{2 \cdot \varepsilon \cdot (V_{bi} + V_r)}{q \cdot N_D}} \tag{3.58}$$

Inserting this and adding (3.57) to (3.53) the total blocking current density of a p^{+}n-junction is obtained as:

$$j_r = j_s + j_{sc} = q \cdot \left(\frac{n_i^2}{N_D} \cdot \frac{L_p}{\tau_p} + \frac{n_i}{\tau_g} \cdot \sqrt{\frac{2 \cdot \varepsilon \cdot (V_{bi} + V_r)}{q \cdot N_D}} \right) \tag{3.59}$$

where D_p/L_p is written as L_p/τ_p to express the two current contributions in a similar manner by a concentration, a length and a lifetime. Also for the diffusion current the condition $V_r > 3 \cdot kT/q$ is used to be in saturation.

Via the width of the depletion region the blocking current increases now with voltage. Whereas the diffusion current is proportional to n_i^2, the space charge term increases only linear with n_i. Hence the diffusion term in proportion to the space charge current increases with intrinsic concentration, that is with decreasing band

gap and increasing temperature. Which part predominates depends, however, also on the lifetime τ_g as compared to τ_p as well as on the voltage V_r. If the generation lifetime τ_g is comparable with τ_p and w comparable with L_p, the space charge term in (3.59) predominates to the extent as $n_i \gg p_{n0} = n_i^2/N_D$ or $n_i \ll N_D$. However, as has been discussed in Sect. 2.7.2, the generation lifetime depends exponentially on the recombination level and becomes very large if E_r is not located near the middle of the band gap. According to the Eqs. (2.70), (2.61) (2.63) and (2.69) of the Shockley-Read-Hall model, the generation lifetime in the depletion region can be written

$$\tau_g = \frac{n_i}{N_r}\left(\frac{1}{c_n \cdot n_r} + \frac{1}{c_p \cdot p_r}\right) = n_i \cdot \left(\frac{\tau_{n0}}{n_r} + \frac{\tau_{p0}}{p_r}\right) \tag{3.60}$$

whereas the low-level minority lifetime τ_p according to Eq. (2.72) is given by

$$\tau_p = \tau_{p0} + (\tau_{p0} \cdot n_r + \tau_{n0} \cdot p_r)/N_D \tag{3.61}$$

n_r and p_r are (except for the degeneracy factor) equal to the carrier concentrations obtained assuming the Fermi level equal to the recombination level [see (2.61a, (2.63a)]. If E_r coincides with the intrinsic level E_i, one has $n_r = p_r = n_i$ and (3.60), (3.61) yield $\tau_g = \tau_{n0} + \tau_{p0}$ which is close to the minimum of τ_g. On the other hand, if the recombination level is distinctly distant from the middle of the band gap, either n_r or p_r is small against n_i and hence $\tau_g \gg \tau_p$ according to (3.60). In this case the space charge current can be small compared with the saturation current.

In Fig. 3.13 the reverse current density of a p^+n-junction in Si is plotted versus the reciprocal absolute temperature for an acceptor level $E_a = E_i$ (intrinsic energy) as well as for a donor level $E_d = E_V + 0.32$ eV (both calculated with g = 2). The reverse voltage is 1000 V. The case of $E_a = E_i$ reflects the behavior of gold-doped junctions since gold has an acceptor level very near the middle of the band gap (see Fig. 2.19). The level 0.32 eV above the valance band coincides with the donor level of Pt. The two levels represent extreme cases regarding the reverse current. A level on the intrinsic energy is most unfavorable because the space charge current is maximal. As shown by the figure (dotted line) it accounts for the whole reverse current, except for the upper temperature range. This holds even at a small reverse voltage. A level 0.32 eV above E_V (or below the conduction band) on the other hand represents an optimal case: The reverse current is much reduced, caused by a very strong reduction of j_{sc} and to some extent also by a lower j_s resulting from the larger low-level lifetime τ_p in the neutral n region. The reverse current is now however nearly totally made up by the saturation current j_s, so it cannot be further reduced via j_{sc}. By a further increase of the distance of the level from the middle of the gap the reverse current can be lowered only relatively weak by a decrease of the saturation current j_s caused by a very high lifetime τ_p. Since this is connected with an extremely strong decrease of the lifetime with injection, it is not advantageous.

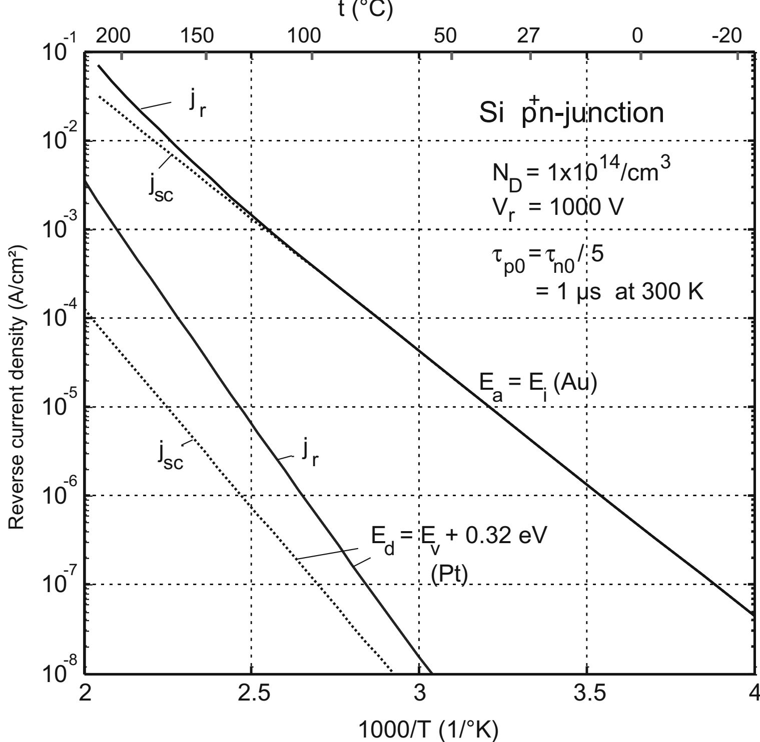


Fig. 3.13 Reverse current density of a p+n-junction in Si as function of temperature for two recombination levels. In both cases the lifetime parameters τ_{n0}, τ_{p0} are assumed as $\tau_{p0} = \tau_{n0}/5 = 1\ \mu s \cdot (T/300)^2$. Degeneracy factor in both cases g = 2

Actually the data in Sect. 2.7.2 d yielded, that the *acceptor* level of Pt (0.23 eV below E_C) accounts for the main part of the space charge current of Pt (see page 80). Nevertheless the j_r-curve for the 0.32 eV level in Fig. 3.13 provides a useful approximate depiction of the reverse current behavior of platinum doped diodes.

The heat dissipation during reverse bias, $j_r \cdot V_r$, is usually much smaller than at forward conduction. However, because j_r increases strongly with temperature, the heat generation can increase stronger than the heat conduction to the sink. Since this can lead to an uncontrolled temperature increase (thermal run-away) and destruction, the reverse current density must be small enough. As an upper limit for diodes of the 1500-V class a value of about 10 mA/cm^2 at 150 °C is typically used. Such a limit implies that the recombination center density has to be limited, which in the case of gold has the consequence that otherwise possible switching times cannot be attained.

3.3.2 Avalanche Multiplication and Breakdown Voltage

At a certain breakdown voltage, the reverse current increases abruptly and the blocking ability is lost. Except for very small breakdown voltages, smaller than 10 V in Si, this is due to impact ionization or avalanche multiplication as described in Sect. 2.8. By the kinetic energy gained in the electric field a carrier raises an electron from the valence to the conduction band and creates an electron-hole pair, and these secondary particles again create electron hole pairs, thus an avalanche process is initiated. The effect is significant only at high field strengths occurring under high reverse bias. With the avalanche generation rate (2.89) the continuity equations (2.102) and (2.103) in the one-dimensional stationary case take the form

$$\begin{aligned} \frac{dj_p}{dx} &= \alpha_p \cdot j_p + \alpha_n \cdot j_n + q \cdot G \\ \frac{dj_n}{dx} &= -\alpha_p \cdot j_p - \alpha_n \cdot j_n - q \cdot G \end{aligned} \tag{3.62}$$

where α_n, α_p are the field-depend impact ionization rates and G as before denotes the thermal generation rate. The total current density

$$j_n + j_p = j \tag{3.63}$$

is independent of x under stationary conditions (see Eq. (2.106)). The avalanche effect is expressed in (3.62) by the proportionality of the generation rate to the current densities which at the relevant field strengths are proportional to the carrier densities. The course of the field strength, of the ionization rates and current densities is shown for a p^+n-junction in Fig. 3.14 (see the legend). To have positive current densities and field strength at reverse bias, the n region is at the left and the p region at the right hand side. The strong variation of j_n, j_p in the high field region due to avalanche multiplication is illustrated qualitatively.

For a qualitative understanding we assume at first that the ionization rates are equal, $\alpha_n = \alpha_p \equiv \alpha$, as is found e.g. in GaAs. Then Eqs. (3.62) with (3.63) simplify to

$$\frac{dj_p}{dx} = \alpha \cdot j + q \cdot G = -\frac{dj_n}{dx} \tag{3.64}$$

Integrating over the space charge region with the boundaries $x_n = 0$ and $x_p = w$ one obtains

$$j - j_{ns} - j_{ps} = j \cdot \int_0^w \alpha \cdot dx + q \cdot w \cdot G$$

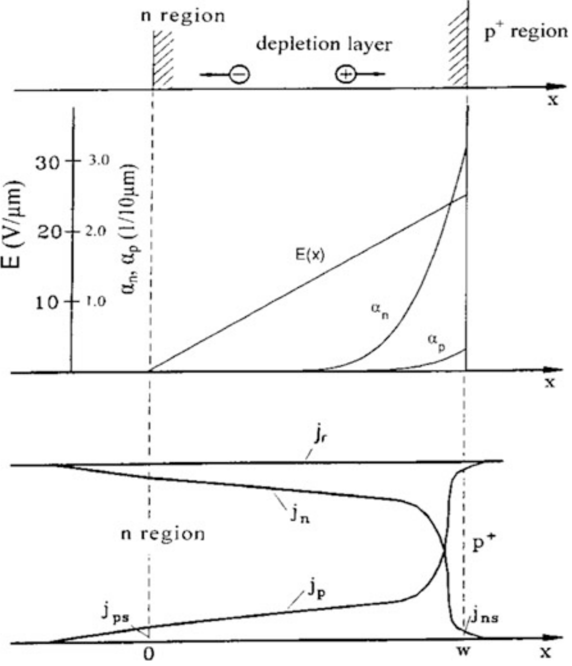


Fig. 3.14 Reverse biased p^+n-junction with avalanche multiplication: Field strength, ionization rates and reverse current densities as function of *x*. The ionization rates in dependency on the field strength refer to Si

since $j_p(x_p) = j - j_{ns}$ and $j_p(x_n) = j_{ps}$, where j_{ns}, j_{ps} are the minority carrier saturation current densities entering the depletion layer from the neutral regions. The current density j follows as

$$j = \frac{j_{ns} + j_{ps} + j_{sc}}{1 - \int_0^w \alpha \, dx} \tag{3.65}$$

where the previous notation $j_{sc} = q \cdot w \cdot G$ is used. As shown by this equation, the effect of avalanche can be expressed by a multiplication factor

$$M = \frac{1}{1 - \int_0^w \alpha \, dx} \tag{3.66}$$

which enhances the 3 components of the thermal current density. If with increasing voltage the field via the ionization rate satisfies the equation

$$\int_0^w \alpha(\mathsf{E}(x)) dx = 1 \tag{3.67}$$

the current increases to infinity. Hence this is the condition from which the breakdown voltage $V_r = V_B = \int_0^w \mathsf{E} \, dx$ can be calculated. Below the breakdown voltage, Eq. (3.65) describes the enhancement of the reverse current by avalanche multiplication.

The considered case $\alpha_n = \alpha_p$ allows a simple calculation because the position where an impact ionization initiates the avalanche process has no influence on the total generated avalanche charge in this case: After generation of an electron-hole pair at a point x the electron moves to the neutral n region and the hole to the p region, so both together cover the whole width of the space charge region (high field region). This holds also for the pairs generated by the secondary carriers at a point x'. Hence if electrons and holes are equally effective, the total generated avalanche charge is independent of x. This is also the cause why all three current components j_{ns}, j_{ps} and j_{sp} are multiplied by the same factor.

If $\alpha_n \neq \alpha_p$ as in the case of Si and most other semiconductors (see Figs. 2.23 and 2.24), the calculation is much more complicated. In Appendix C it is shown that the avalanche multiplication can be expressed generally by double integrals over the ionization rates. As expected, the three components of the thermal blocking current (nominator of (3.65)) are enhanced for $\alpha_n \neq \alpha_p$ each by a different multiplication factor and hence the leakage current density is generally given by

$$j = M_n \cdot j_{ns} + M_{SC} \cdot j_{SC} + M_p j_{ps} \tag{3.68}$$

From $\alpha_n > \alpha_p$ in the case of Si, it follows that $M_n > M_{SC} > M_p$. All three M's tend to infinity at the same voltage, the breakdown voltage. For an abrupt p^+n-junction in Si the multiplication factors, as calculated numerically from the formulae in the appendix using the ionization rates of Eqs. (2.96), are plotted in Fig. 3.15 versus the reverse voltage. The breakdown voltage for the assumed doping density 1×10^{14} cm^{-3} of the n region is 1640 V. Already at 1400 V, the current density j_{ns} is enhanced by $M_n = 5$, whereas the multiplication factor M_p amounts only to 1.17 at this voltage. In the case of a single p^+n-junction the higher M_n is not significant for the blocking current, since the electron saturation current j_{ns} is very small

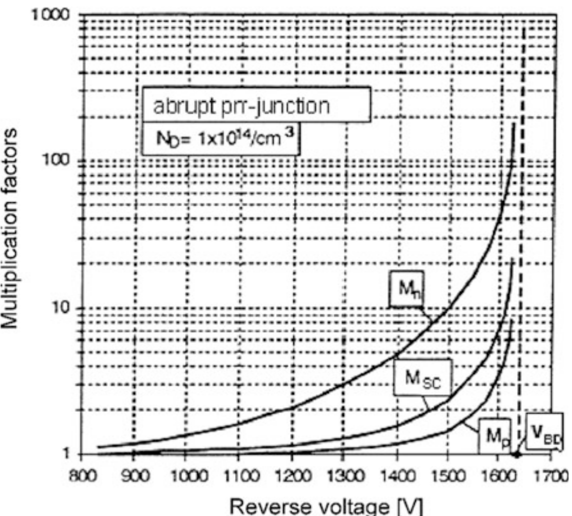


Fig. 3.15 Multiplication factors in dependence of the voltage for an abrupt p^+n-junction with $N_D = 1 \times 10^{14}$ cm^{-3}, $T = 300$ K

according to (3.51) due to the high N_A. For a transistor structure with a p^+n-junction between base and collector, however, the large M_n has a strong effect on the breakdown voltage, as will be seen.

Analytically, the multiplication factors as function of voltage are approximated often by the equation [Mil57]

$$M = \frac{1}{1 - (V/V_B)^m} \tag{3.69}$$

where the exponent m is used for fitting. To approximate the dependencies of Fig. 3.15, very different m-values are necessary for M_n and M_p. Fitting in both cases at $M_{n,p} = 2$, a value $m = 2.2$ is obtained for M_n and $m = 13.2$ for M_p. These values differ significantly from values in the literature where often m = 4 and 6 are used for M_n, M_p respectively. Although (3.69) is only a rough approximation, considering the whole voltage range, it is very useful for dimensioning of transistor and thyristor structures. The large difference between M_n and M_p has a significant influence on the blocking behavior.

If only the breakdown voltage of a (single) pn-junction is of interest, the complicated ionization integrals in the Appendix C can be avoided. As noted already in Sect. 2.8 and shown in Appendix C, the breakdown voltage can be calculated using the effective ionization rate

$$\alpha_{eff} = \frac{\alpha_n - \alpha_p}{\ln(\alpha_n/\alpha_p)} \tag{3.70}$$

which at breakdown satisfies the condition [Wul60, Oga65]

$$\int_0^w \alpha_{eff}(\mathsf{E}(x))dx = 1 \tag{3.71}$$

similarly as α in (3.67). This holds exactly only, if the ratio α_n/α_p is independent of E. However, also if this is not well fulfilled over a large field range, (3.71) is often a good approximation, because only a relative small field range contributes significantly to the integral.

The maximum field strength in the depletion layer, at which breakdown occurs, is called the critical field strength E_c. If the width of the high-field region is varied via the doping density according to (3.58) for a p^+n-junction, this must be compensated according to (3.71) by an inverse variation of α_{eff}. However, due to the very strong increase of the α's with E, this requires only a small change of the field strength. Hence E_c depends only weakly on the thickness w of the depletion region, and for rough estimates E_c is assumed often as constant.

Without any restriction we assume at first that the critical field (in dependence of w and N_D) is known and ask how the width and doping concentration of the weakly doped region(s) of a pn-junction are to be chosen to reach a desired breakdown

voltage V_B. Considering an abrupt p$^+$n-junction as shown in Fig. 3.14 the breakdown voltage is

$$V_B = \int_0^w \mathsf{E} dx = \frac{1}{2} w \mathsf{E}_c \tag{3.72}$$

Within the approximation of constant E_c the breakdown voltage is proportional to the width w of the space charge region. According to the Poisson equation

$$\frac{d\mathsf{E}}{dx} = \frac{\mathsf{E}_c}{w} = \frac{q \cdot N_D}{\varepsilon} \tag{3.73}$$

w depends inversely on the doping density N_D. Expressing the width w in (3.72) by N_D one obtains

$$V_B = \frac{\varepsilon \cdot \mathsf{E}_c^2}{2 \cdot q \cdot N_D} \tag{3.74}$$

Hence, within the approximation of constant E_c, the breakdown voltage is inversely proportional to the doping density of the n region. However, as mentioned, this is only a very rough approximation.

To calculate the critical field in dependence of w or N_D in analytical form, we use the power approach (2.94) in the form [Shi59, Ful67]

$$\alpha_{eff} = B \cdot \mathsf{E}^n \tag{3.75}$$

With the coordinates of Fig. 3.14, the field strength is

$$\mathsf{E}(x) = \frac{q\, N_D}{\varepsilon} x \tag{3.76}$$

Inserting (3.75) together with $dx = \varepsilon/(q \cdot N_D) \cdot d\mathsf{E} = w/\mathsf{E}_c \cdot dE$ into the condition (3.71) one gets

$$\frac{B\, w}{\mathsf{E}_c} \int_0^{\mathsf{E}_c} \mathsf{E}^n\, d\mathsf{E} = B\, w \frac{\mathsf{E}_c^n}{n+1} = 1 \tag{3.77}$$

$$\mathsf{E}_c = \left(\frac{n+1}{B \cdot w}\right)^{\frac{1}{n}} = \left(\frac{q \cdot (n+1) \cdot N_D}{B \cdot \varepsilon}\right)^{\frac{1}{n+1}} \tag{3.78}$$

The last expression follows from (3.77) with $w = \varepsilon \cdot \mathsf{E}_c/(q \cdot N_D)$ [Eq. (3.73)]. As expected, E_c decreases slightly with increasing w and increases slightly with N_D.

Inserting (3.78) into (3.74) the breakdown voltage as function of doping density is obtained as

$$\begin{aligned} V_B &= \frac{\varepsilon}{2 \cdot q \cdot N_D} \cdot \left(\frac{q \cdot (n+1) \cdot N_D}{B \cdot \varepsilon} \right)^{\frac{2}{n+1}} \\ &= \frac{1}{2} \cdot \left(\frac{n+1}{B} \right)^{\frac{2}{n+1}} \cdot \left(\frac{\varepsilon}{q \cdot N_D} \right)^{\frac{n-1}{n+1}} \end{aligned} \tag{3.79}$$

(3.72) and (3.78) yield

$$V_B = \frac{1}{2} \cdot \left(\frac{n+1}{B} \cdot w^{n-1} \right)^{\frac{1}{n}} \tag{3.80}$$

The width w of the depletion region as function of N_D at breakdown follows from (3.78) as

$$w = \left(\frac{n+1}{B} \right)^{\frac{1}{n+1}} \cdot \left(\frac{\varepsilon}{q \cdot N_D} \right)^{\frac{n}{n+1}} \tag{3.81}$$

Due to the variation of the critical field, the increase of V_B with w and with $1/N_D$ is sub-linear.

To adjust (3.75) to the Chynoweth law, the constants n and B are to be chosen according to (2.95), (2.94). With the numerical Eq. (2.92) for α_{eff} in Si at 300 K, one obtains in dependence of the field E_0 where the adjustment is carried out:

$$n = \frac{1.68 \times 10^6 \text{ V/cm}}{\mathsf{E}_0} \tag{3.82}$$

$$B = \frac{1.06 \times 10^6 \text{ V/cm}}{\mathsf{E}_0^n \cdot \exp(n)} \tag{3.83}$$

Following Shields [Shi59], most authors set the exponent to $n = 7$ in silicon. According to (3.82) this is obtained at $\mathsf{E}_0 = 2.4 \times 10^5$ V/cm. The corresponding B-value according to (3.83) is $B = 2.107 \times 10^{-35}$ cm^6/V^7. Using these constants, Eqs. (3.80), (3.81) can be written

$$V_B = \frac{1}{2} \cdot \left(\frac{8}{B} \right)^{\frac{1}{4}} \cdot \left(\frac{\varepsilon}{q \cdot N_D} \right)^{\frac{3}{4}} = 563 \text{ V} \cdot \left(\frac{4 \times 10^{14}/cm^3}{N_D} \right)^{\frac{3}{4}} \tag{3.84}$$

$$w = \left(\frac{8}{B} \right)^{\frac{1}{8}} \cdot \left(\frac{\varepsilon}{q \cdot N_D} \right)^{\frac{7}{8}} = 42.6\,\mu\text{m} \cdot \left(\frac{4 \times 10^{14}/\text{cm}^3}{N_D} \right)^{\frac{7}{8}} \tag{3.85}$$

In the numerical expressions on the right hand side, the doping density is related to the value 4×10^{14} cm^{-3} which is roughly the point at which the approximation is adapted and around which it will be very accurate. It is taken into account that the integral of α_{eff} (E) over the triangular field distribution (see Fig. 3.14) is best approximated choosing E_0 near $0.9 \cdot E_c$. Hence the choice $n = 7$, $E_0 = 2.4 \times 10^5$ V/cm means that the maximum field at the matching point is $E_c = 2.4 \times 10^5/0.9 = 2.667 \times 10^5$ V/cm. This field corresponds to the doping density $N_D = 4.36 \times 10^{14}$ cm^{-3} according to (3.78).

At much higher and lower voltages the accuracy suffers from the deviation of the power approximation from the Chynoweth law as illustrated in Fig. 2.23. In Fig. 3.16, the exponent n and the constant B are plotted versus the field E_0 where the fitting is carried out. n varies nearly by a factor 3 in the relevant field range. The extreme increase of B with E_0 is caused according to (3.83) by the decrease of n and the resulting rapid decrease of E_0^n. To take into account these variations, it is suitable to choose the matching point E_0 for each doping density separately as $0.9 \cdot E_c\ (N_D)$.

Inserting (3.82), (3.83) with $E_0 = 0.9 \cdot E_c$ into the right hand side of (3.78), one can solve the equation for E_c by iteration. With the critical field the breakdown voltage and width of the space charge layer are given by (3.74), (3.73) or can be calculated from (3.79), (3.81), since with E_c one has E_0 and then n and B. The breakdown voltage and width w at breakdown obtained in this way are plotted in Fig. 3.17 as functions of N_D. These results are very close to those calculated from the exact ionization integral given in Appendix C. In the figure also the $N_D^{-3/4}$ dependency according to Eq. (3.84) is shown (dotted line). As is seen, this simple formula is a good approximation over a wide range. For doping densities $N_D < 10^{14}$ and $>10^{15}$ cm^{-3} it underestimates the breakdown voltage. The used approach allows an analytical integration of (3.71) also in other cases where the field strength depends linearly on x. For $p^+n^-n^+$-diodes this is utilized in Chap. 5.

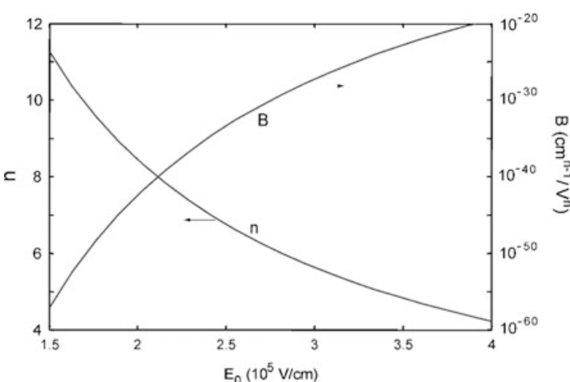


Fig. 3.16 Constants in the power law $\alpha_{eff} = B \cdot E^n$ as functions of the field strength at which it is fitted to $\alpha_{eff} = a/\exp(b/E)$. $a = 1.106 \times 10^6$ /cm, $b = 1.68$ V/cm (Si at 300 K)

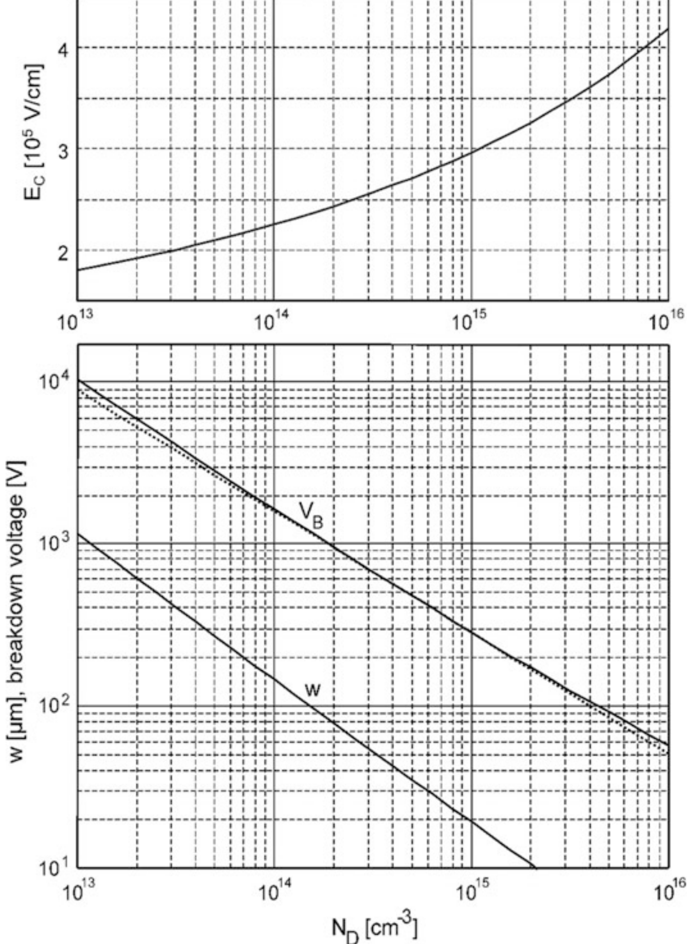


Fig. 3.17 Critical field strength, breakdown voltage and depletion width at breakdown as functions of the doping density N_D for abrupt p^+n-junctions in Si at 300 K. The dotted straight line represents the approximation (3.84)

Since often used, the explicit dependencies of N_D, w_B and E_c on the breakdown voltage are still noted. The inversion of (3.79) and (3.80) yields:

$$N_D = \frac{\varepsilon}{q}\left(\frac{n+1}{B}\right)^{\frac{2}{n-1}}\left(\frac{1}{2V_B}\right)^{\frac{n+1}{n-1}} \tag{3.86}$$

$$w_B = \left(\frac{B}{n+1}\right)^{\frac{1}{n-1}}(2V_B)^{\frac{n}{n-1}} \tag{3.87}$$

From Eq. (3.77), right hand part, one obtains inserting $w_B = 2V_B/\mathsf{E}_c$:

$$\mathsf{E}_c = \left(\frac{n+1}{2BV_B}\right)^{\frac{1}{n-1}} \tag{3.88}$$

These equations will be used later at several points in the book.

For diffused junctions with a very steep profile, the results for abrupt junctions can be used as an approximation. Generally, the approach (3.75) is useless for diffused junctions, because the integration of (3.71) or the ionization integral can be carried out even with this approximation only numerically. For a given doping density of the lightly doped region diffused junctions have a higher breakdown voltage than abrupt ones. This follows because (i) the maximum field at the junction is reduced and (ii) the field extends considerably into the highly doped diffused region near the junction. In Fig. 3.18 the numerically calculated breakdown voltage of diffused p^+n-junctions is plotted versus the background doping density N_D for two examples of steepness of the diffusion profile. The acceptor density around the junction at $x = 0$ is approximated here by an exponential decrease $N_A(x) \sim \exp(-x/\lambda)$ (orientation as in Fig. 3.6). Hence, λ is the decay length of the diffusion profile near the junction. Since $N_A(x) = N_D$ at $x = 0$, the net doping density in the space charge region is

$$N(x) = N_D - N_A(x) = N_D \cdot (1 - e^{-x/\lambda}) \tag{3.89}$$

and the doping gradient at the junction is $dN/dx(0) = -N_D/\lambda$. The case $\lambda = 3$ µm is typical for pn^-p structures in thyristors diffused with gallium or boron, the case $\lambda = 10$ µm is representative for high voltage pnp-structures realized with very deep aluminum diffusion.

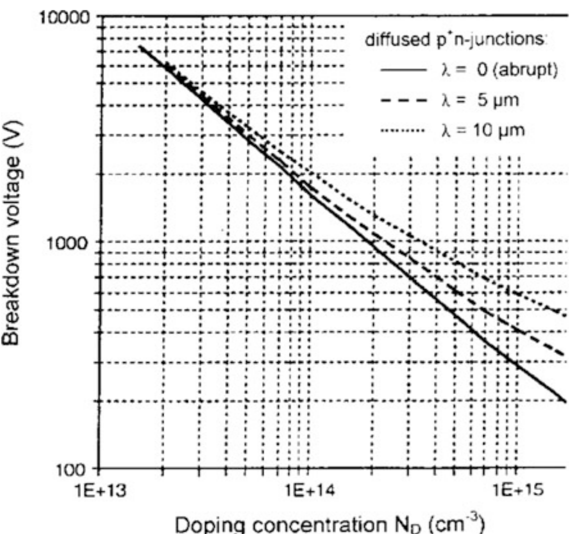


Fig. 3.18 Breakdown voltage of diffused p^+n-junctions in Si. λ is the decay lengths of the diffusion profile in the depletion layer

Temperature dependence
As described in Sect. 2.8, the ionization rates decrease with increasing temperature; for the effective ionization rate this is expressed numerically by (2.97). Hence the critical field strength and avalanche breakdown voltage increase slightly with T. Using again the power approach (3.75) to give an analytical description, the exponent n and constant B are needed as function of temperature. This is obtained from Eq. (2.98) together with (3.75):

$$
\begin{aligned}
n(T) &= \frac{b_{eff}(T)}{\mathsf{E}_0} = \frac{1.68 \times 10^6 + 1100 \cdot (T - 300)}{\mathsf{E}_0} \\
B(T) &= \frac{C(T)}{\mathsf{E}_0^{n(T)}} = \frac{a_{eff}}{\mathsf{E}_0^{n(T)} \exp(n(T))} = \frac{1.06 \times 10^6}{\mathsf{E}_0^{n(T)} \cdot \exp(n(T))}
\end{aligned}
\tag{3.90}
$$

where E_0 is scaled in V/cm. Inserting these dependences in Eq. (3.79), the breakdown voltage for a triangular field distribution is given analytically as function of temperature. If for example the doping concentration is $N_D = 1 \times 10^{14}$ cm^{-3}, a good choice for E_0 is 2.2×10^5 V/cm (see Fig. 3.17). Then at 300 K Eq. (3.90) yields $n = 7.64$, $B = 7.78 \times 10^{-39}$ cm^{n-1}/V^n. With these values Eq. (3.79) gives $V_B = 1630$ V. At $T = 400$ K one obtains $n = 8.14$, $B = 1.01 \times 10^{-41}$ cm^{n-1}/V^n, and (3.79) results in $V_B = 1835$ V, a 12.6% higher value than at 300 K. Due to a considerable decrease of V_B from room temperature downwards, one has to dimension so that the targeted blocking ability is maintained still at the lower limit of operation temperature (≈250 K).[3] This holds for diodes. For thyristors, the temperature dependency is weaker (and can be inverted), because the blocking voltage is supplied here by a transistor structure and depends not only on avalanche multiplication but also on the current amplification factor of a transistor, which increases with temperature.

3.3.3 *Blocking Capability with Wide-Bandgap Semiconductors*

The presented theory is of course applicable also to other semiconductors, particularly to those which have a wider band gap than Si. Owing to the higher energy required to carry an electron from the valence into the conduction band, the critical field strength increases with band gap. A summary of raw values of the critical field is given in Table 3.1. Using these values in (3.72), (3.74) a rough estimate can be obtained about the needed width w of the depletion region and the allowed maximum doping density for a given blocking ability.

[3] Some manufacturers data sheets specify V_B at 25 °C, and for application at lower temperatures one must be aware of said decrease of V_B with temperature.

Table 3.1 Raw values of the critical field strength for Si and semiconductors with a wider band gap

	E_C[V/cm]
Si	2×10^5
GaAs	4×10^5
4H-SiC	3×10^6
GaN	$>3 \times 10^6$
C (diamond)	$1–2 \times 10^7$

For 4H-SiC even an E_c-value of 3.3 MV/cm was reported recently for local avalanche [Rup14]. Using the value given in the table for SiC a p^+n structure blocking 10 kV requires a base width $w = 2V_B/E_c = 67$ µm and allows a base doping density $N_D = \varepsilon E_c^2/(2qV_b) = 2.40 \times 10^{15}$ cm^{-3}. For a more accurate calculation the decrease of the critical field strength with thickness of the depletion region or decreasing doping density is to be considered. The parameters for the effective ionization rate in (3.75) are reported for 4H-SiC as $n = 8.03$ and $B = 2.18 \times 10^{-48}$ cm$^{7.03}$V$^{-8.03}$ [Bar09]. Inserting these into (3.78) the critical field strength is obtained as

$$E_c = 2.58 \cdot \left(\frac{N_D}{10^{16}/\text{cm}^3}\right)^{0.111} \text{MV/cm} \tag{3.91}$$

For the blocking capability Eq. (3.79) gives:

$$V_B = 1770 \cdot \left(\frac{10^{16}/\text{cm}^3}{N_D}\right)^{0.78} \text{V} \tag{3.92}$$

The breakdown voltage calculated in this way for an asymmetrical junction in 4H-SiC is shown in Fig. 3.19 together with that of silicon. As can be seen, the maximum doping concentration for a given blocking voltage, for example 1000 V, is for SiC two decades higher than for Si, a consequence of the critical field strength being an order of magnitude higher. Owing to the much higher doping concentration and the smaller possible base width, SiC-devices show superior characteristics regarding conduction and switching losses, as will be shown in detail later.

From measurements at different temperatures with 4H-SiC $p^+n^-n^+$ diodes Bartsch et al. [Bar09] obtained the following temperature dependence of the parameters n and B

$$n = 6.78 + 1.25 \cdot \frac{T}{300\text{K}} \quad B = 3 \cdot 10^{-40} \cdot \exp\left(-18.74 \cdot \frac{T}{300\text{K}}\right) \tag{3.93}$$

At $T = 300$ K this leads to (3.91) and (3.92). Also mature SiC merged pin-Schottky diodes show a temperature dependence of blocking voltage which is in good agreement with these results; this is displayed later in Sect. 6.5 with Fig. 6.15.

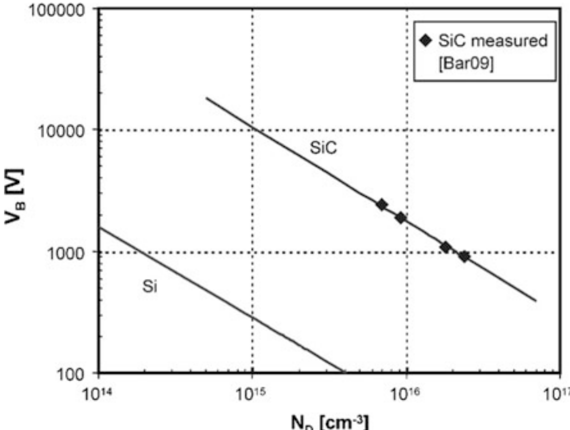


Fig. 3.19 Breakdown voltage of asymmetrical junctions in 4H-SiC and Si as function of the doping concentration of the weakly doped region

Several years it was practice in the design of SiC-devices to limit the electric field to a maximum value of about 1.5 MV/cm. The devices were specified for a lower blocking voltage than the breakdown voltage resulting from critical field strength, because the leakage current otherwise surmounts the allowed range. The reasons for the enhanced reverse current are crystal defects. In recent time, the SiC crystal quality has been improved strongly and former limits are overcome.

3.4 Injection Efficiency of Emitter Regions

In preceding sections mainly asymmetric pn-junctions have been considered. It turned out that to the extent that the doping ratio N_A/N_D is very large or small the influence of the properties of the highly doped region on the characteristics vanishes, in case of blocking as well as in forward direction. Since the density of injected minority carriers at forward bias is inversely proportional to the doping density according to (3.36), (3.37), mainly the highly doped region injects carriers into the weakly doped region and not vice versa. Because this is similar to the function of the emitter and base regions in transistors, the highly and weakly doped regions of asymmetrical junctions are also called emitter and base region, respectively. The injection efficiency, defined as the ratio of minority carrier current in the base region to the total current, is called emitter efficiency. A quantitative comparison of theory and experiment shows now that the measured emitter efficiency is essentially smaller than calculated on the base of the equations in Sect. 3.2. This manifests itself immediately in the current amplification factor of transistors. For power diodes with a p^+nn^+-structure, the current injected from the base into the emitter regions becomes much earlier significant with increasing injection level than according to the given 'classical' equations. A consequence is that the forward voltage drop of diodes and other power devices is significantly enhanced compared

with theoretical characteristics developed on the basis of above equations. The cause is that the band gap reduction in highly doped region(s) has been neglected till now.

In the present section we examine the injection properties of emitter regions with consideration of band gap narrowing. Furthermore, the injection level in the base region is allowed to be arbitrary and the high Auger recombination in the emitter is considered. According to the continuity Eq. (3.40), the minority carrier current entering the emitter is absorbed by recombination in the volume of the emitter and, if the thickness of the region is smaller than a few minority carrier diffusion length, at the contact. Although the contact recombination is utilized in novel devices to improve switching properties, we assume at first that only recombination in the volume is significant.

a) Characterization of emitter recombination:

The injection properties of a forward biased abrupt n^+p junction are examined first for the case of *arbitrary injection level in the base*, without considering already bandgap narrowing. The injection level in the emitter region is assumed to be small (see Fig. 3.20). According to Eqs. (3.3), (3.4), the ratio of carrier densities on one side to that on the other side of the space charge region are given by the Boltzmann factor containing the total potential difference between both sides:

$$\frac{p_n^*}{p_B^*} = \frac{n_B^*}{n^+} = e^{-q \cdot \Delta V/kT} \tag{3.94}$$

As illustrated in Fig. 3.20, the star indicates the carrier concentrations in the neutral regions at the boundaries to the space charge layer. The concentration n^+ in the emitter is equal to the donor density N_D due to the assumed low injection level there. For the minority carrier density in the emitter region (3.94) gives

$$p_n^* = \frac{n_B^* \cdot p_B^*}{n^+} = \frac{n_B^* \cdot (N_B + n_B^*)}{n^+} \tag{3.95}$$

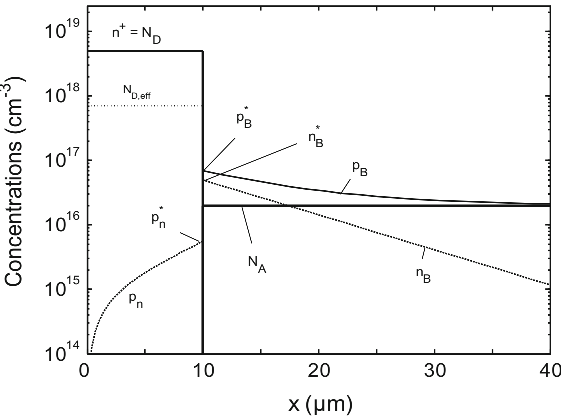


Fig. 3.20 Forward biased n^+p-junction with intermediate injection level in the p-base region

where the equation $p_B^* = N_B + n_B^*$ follows from neutrality. The base doping N_A is replaced here by N_B to include the case of an n$^+$n-junction, occurring in power diodes, by defining $N_B = -N_D$ in case of an n-base. For power diodes which have a low base doping, (3.95) is mainly significant for high injection levels in the base region where $p_B^* \approx n_B^*$, in bipolar transistors and thyristors the injection level in the p-base region is typically intermediate in the interesting current range. As shown earlier (see Eqs. (3.46) and (3.43)) the injected holes result in the minority current density in the n$^+$-region

$$j_p(x_n) = q \cdot \frac{D_p}{L_p} \cdot (p_n^* - p_{n0}) \tag{3.96}$$

Neglecting the very small equilibrium concentration $p_{n0} = n_i^2/N_D$ this can be written using (3.95)

$$j_p(x_n) = q \cdot h_n \cdot n_B^* \cdot p_B^* \tag{3.97}$$

where the properties of the n emitter are comprised in the constant:

$$h_n = \frac{D_p}{N_D \cdot L_p} = \frac{1}{N_D} \cdot \sqrt{\frac{D_p}{\tau_p}} \tag{3.98}$$

This parameter characterizes the emitter recombination which reduces the injection efficiency. It equals the saturation current density j_{ps} of the emitter defined in Eq. (3.46a) except that the factor $q \cdot n_i^2$ is omitted. It is useful to introduce the h-quantity, because the intrinsic concentration with its strong temperature dependence is not relevant for the injection efficiency for a given current. The h–parameter is nearly temperature-independent. The Eq. (3.97) defining the h-parameter by the proportionality between $j_p(x_n)$ and the product $n_B^* \cdot p_B^*$ is usable also for diffused emitter regions. The experimental h-values determined from (3.97) [Sco69, Bur75, Sco79] are not in close accordance with the saturation current density determining the I-V-characteristics at low current densities.

For the emitter efficiency of the n$^+$p-junction

$$\gamma = j_n(x_p)/j \tag{3.99}$$

one obtains with $j_n(x_p) = j - j_p(x_p) \approx j - j_p(x_n)$ by insertion of (3.97)

$$\gamma = 1 - j_p(x_n)/j = 1 - q \cdot h_n \cdot n_B^* \cdot p_B^*/j \tag{3.100}$$

As is seen, a small h_n–value is required to get a high emitter efficiency and current amplification.

According to (3.98) h_n can be made small by choosing a high doping density N_D. The explicit N_D dependence is however counteracted by the decrease of the minority carrier lifetime due to Auger recombination described in Sect. 2.7.1. For

$N_D > 5 \times 10^{18}$ cm^{-3} the lifetime is often solely determined by this recombination process, hence according to (2.56) one has $1/\tau_p = c_{A,n} \cdot N_D^2$. Inserting this, (3.98) yields the following "limiting" h-value for high N_D

$$h_{n,\mathrm{lim}} = \sqrt{D_p \cdot c_{A,n}} \tag{3.101}$$

which decreases however still slightly with increasing n$^+$ owing to the decrease of the diffusion constant D_p. Using the Auger coefficient of (2.57) and D_p = 1.9 cm^2/s for a doping concentration of 10^{19} cm^{-3}, (3.101) gives $h_{n,lim} = 7.3 \times 10^{-16}$ cm^4/s. Experimentally, the lowest h-values are an order of magnitude higher [Sco69, Bur75, Sco79, Coo83]. This is attributed to the band gap narrowing in the emitter region neglected till now.

b) Influence of bandgap narrowing:

In *thermal equilibrium* the ratio of electron concentrations in the weakly doped p region to that in the highly doped n$^+$ region is

$$\frac{n_{p0}}{N_D} = \frac{n_{i0}^2}{N_A N_D}$$

where n_{i0} is the intrinsic concentration in the p region (with low doping). The ratio of *hole* concentrations in the highly doped n$^+$ region to that in the p region is

$$\frac{p_{n0}}{N_A} = \frac{n_i^2}{N_A N_D} = \frac{n_{i0}^2 e^{\Delta E_g/kT}}{N_A N_D}$$

See Eq. 2.27 due to the bandgap narrowing the change of the hole concentration over the space charge region is enhanced compared with the change of electron concentration in the opposite direction:

$$\frac{p_{n0}}{N_A} = \frac{n_{p0}}{N_D} e^{\Delta E_g/kT} \tag{3.102}$$

If now a forward voltage V is applied, the associated potential change in the space charge region acts in the same manner on the holes as on the electrons, apart from the opposite direction of the force. Both sides of (3.102) are increased hence by the same factor $\exp(qV/kT)$. This is quite independent of the injection level, which in the p region is arbitrary. Hence with the notation of Fig. 3.20 it follows:

$$\frac{p_n^*}{p_B} = \frac{n_B^*}{N_D} e^{\Delta E_g/kT} \tag{3.103}$$

$$p_n^* = \frac{n_B^* p_B^*}{N_D} e^{\Delta E_g/kT} \tag{3.104}$$

where $p_B^* = N_A + n_B^*$. Defining an effective doping concentration $N_{D,eff} = N_D \cdot \exp(-\Delta E_g/kT)$, one has for the concentration ratios at opposite sides of the space charge layer for all doping densities $p_n^*/p_B^* = n_B^*/N_{D,eff}$ (see Fig. 3.20). This effective doping concentration increases only very little with increasing N_D. It serves only to illustrate the effect of band gap narrowing on the injection properties and has no significance at all for the majority carrier concentration $n^+ = N_D$. Inserting (3.104) into (3.96) the h_n quantity as defined by (3.97) takes the form

$$h_n = \frac{1}{N_D} \frac{D_p(N_D)}{L_p(N_D)} exp\left(\frac{\Delta E_g}{kT}\right) \tag{3.105}$$

Inserting ΔE_g from Eqs. (2.25) or (2.26) it is seen, that the h-value is strongly enhanced by the bandgap reduction. With increasing N_D the decrease of the first term holding for $\Delta E_g = 0$ will be overcompensated at higher N_D by the increase of the ΔE_g-term (see the later Fig. 3.21).

As mentioned, the highly doped region was assumed till now thick enough that recombination at the contact could be neglected. Often however the contact recombination is considerable or made large even intentionally. Fortunately, it can be taken into account simply by replacing the diffusion length L_p in (3.105) by a length $L_{p,eff}$ defined suitably for this purpose. Assuming a contact with infinite recombination velocity the minority carrier concentration at the surface is pressed down to the equilibrium value p_{n0} so that $p \approx p - p_{n0} = 0$. With this boundary condition at $x = 0$ the solution of the differential Eq. (3.42) is $p_n(x) = p_n^* \cdot \sinh(x/L_p)/\sinh(w_n/L_p)$ (w_n width of the emitter). The gradient of this carrier distribution at $x = w_n$ is

$$dp/dx(x_n) = p_n^*/(L_p \cdot \tanh(w_n/L_p)$$

which for $w_n \to \infty$ tends to the previous form p_n^*/L_p. Hence for small w_n the length to be used instead of L_p is

$$L_{p,eff} = L_p \cdot \tanh(w_n/L_p) \tag{3.106}$$

Thus the emitter parameter is finally given in this model by:

$$h_n = \frac{D_p}{N_D \cdot L_p \cdot \tanh(w_n/L_p)} \cdot e^{\Delta E_g/kT} \tag{3.107}$$

where the dependency on the doping and carrier concentration $N_D = n^+$ is contained apart from N_D itself in $L_p = \sqrt{(D_p \cdot \tau_p)}$, ΔE_g, and D_p. For very high N_D, where τ_p is determined by Auger recombination and $L_p = \sqrt{(D_p/c_{A,n})}/N_D$ is smaller than about $w_n/3$, Eq. (3.107) yields instead of (3.101)

$$h_n = \sqrt{D_p \cdot c_{A,n}} \cdot e^{\Delta E_g/kT} \tag{3.108}$$

Formulae for a p-emitter are obtained by exchanging n and p and N_D and N_A. Equation (3.107) for a p-emitter reads

$$h_p = \frac{D_n}{N_A \cdot L_n \cdot \tanh(w_p/L_n)} \cdot e^{\Delta E_g/kT} \tag{3.109}$$

For later use, we note also the formula for a small width of the p$^+$-region. If $w_p < L_n/3$, Eq. (3.109) turns into

$$h_p = \frac{D_n}{N_A \cdot w_p} \cdot e^{\Delta E_g/kT} \tag{3.110}$$

This equation holds also for position dependent emitter doping if $N_A \cdot w_p$ is replaced by the integral $\int N_A(x) \cdot dx$. For ΔE_g and D_n one has to use then suitable mean values. In Fig. 3.21 the h-parameter of a p$^+$-region as calculated from (3.109) is plotted versus the doping density N_A for a very small and a larger width w_p. The minority carrier lifetime τ_n was calculated from the Shockley-Read-Hall lifetime τ_{SRH} assumed to be 2 µs and Auger recombination according to $1/\tau_n = 1/\tau_{SRH} + c_{A,p} \cdot N_A^2$, where the Auger constant $c_{A,p}$ of (2.57) was inserted. As band gap narrowing $\Delta E_g(N_A)$ Eq. (2.25) of [Slo76] as well as Eq. (2.26) of Lanyon and Tuft [Lan79] were used. The minority diffusion constant $D_n(N_A)$ was calculated from the mobility μ_n using for simplicity the same dependence on doping concentration as in an n$^+$-region: $D_n(N_A) = kT/q \cdot \mu_n(N_D = N_A)$ (see Sect. 2.6 and Appendix A). As expected from Fig. 2.11, the ΔE_g of [Lan79] results in considerably smaller h_p below 10^{19} cm^{-3} than that of Slotboom and De Graaf. Experimental h-values obtained with power diodes are mostly in rough agreement with the $\Delta E_g(N)$-dependency of [Slo76]. At small N_A corresponding curves tend to the same value, since ΔE_g then approaches

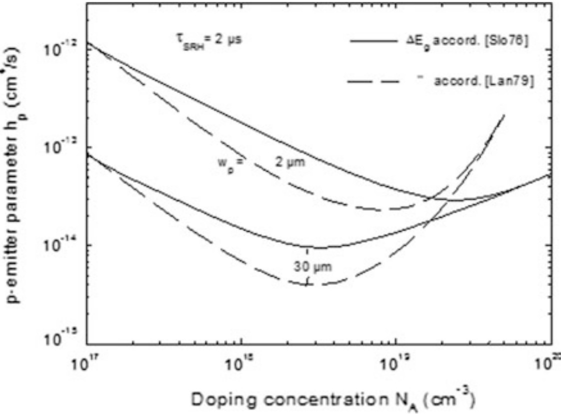


Fig. 3.21 h-parameter of a p$^+$-region as a function of doping concentration according to Eq. (3.109)

zero in both cases. As is shown already by the formulae, h can be strongly enhanced by choosing a small thickness of the emitter. The curves for w_p = 2 μm are described up to about 1 × 10^{19} cm^{-3} by (3.110) showing that recombination at the contact predominates. For w_n = 30 μm on the other hand, recombination in the emitter volume is significant in the whole range. Due to the decrease of the lifetime because of Auger recombination adding to the increase of ΔE_g with N_A, h_p has a minimum at 3 × 10^{18} cm^{-3}.

Concerning n-emitter regions, it can be seen, that due to smaller minority diffusion constant D_p and the higher Auger recombination constant $c_{A,n}$ the numerical h_n-values will be somewhat smaller than h_p under the same conditions. For a small emitter thickness the formula corresponding to (3.110) yields a h_n which is by the factor D_p/D_n smaller than h_p provided the integrated doping density is equal in both cases. The band gap narrowing is generally assumed to be the same in n^+ and p^+ regions. At very high doping densities the parameters h_n and h_p are very similar according to (3.108) since the products $D_n \cdot c_{A,p}$ and $D_p \cdot c_{A,n}$ differ not much. The bandgap narrowing is generally assumed to be the same in n^+ and p^+-regions

Although not all these consequences of the theoretical model have been verified precisely and the band gap narrowing itself is not yet completely understood quantitatively, the model has proved to be quite useful for describing the main influence of emitter layout on device behavior.

For the experimental determination of emitter properties, diodes with a p^+nn^+- or p^+pn^+-structure have been used in a forward current range where the injection levels in the base region is high. Hence (3.97) was used in the form

$$j_p(x_n) = q \cdot h_n \cdot n_B^{*2} \tag{3.111}$$

Typical highly doped emitter regions have h-values in the range 1 × 10^{-14} – 3 × 10^{-14} cm^4/s, and the lowest values observed are about 7 × 10^{-15} cm^4/s [Sco69, Bur75, Sco79, Coo83]. As will be treated in detail in Chap. 5, the h-values of the two emitter regions of pin-diodes and thyristors influence largely their forward voltage drop from moderate current densities upwards. For low-frequency power diodes which are designed to show a forward voltage drop as low as possible, small h-values are required. The injection efficiency of the junctions then is high. If h = 1 × 10^{-14} cm^4/s and $n_B^* = p_B^* = 1 \times 10^{17}/cm^3$ at a current density of 200 A/cm^2, one obtains from (3.100) γ = 0.92. On the other hand, the h-value of the p-emitter of fast power diodes and IGBTs is made often very high by choosing a small thickness and a relatively low doping density. In a typical example one may have h_p = 1 × 10^{-12} cm^4/s, and the carrier concentration in the base region at the p-emitter side may be 1.8 × 10^{16} cm^{-3} at a forward current density of 150 A/cm^2. In this case (3.100) yields γ = 0.65. The electron particle current which enters the p^+-region and ends at the contact by recombination, amounts then to more than a third of the total current.

An observation which is not expected from the above theory is that the minority carrier current flowing into the emitter regions contains besides the quadratic term

(3.111) also a significant part which increases proportional with the concentration n_B* [Sco79, Coo83]. Hence the minority carrier current density is actually given by

$$j_p(x_n) = q \cdot (h_n \cdot {n_B^*}^2 + s_n \cdot n_B^*) \tag{3.112}$$

This was found for all samples investigated regardless whether prepared by diffusion, epitaxy or alloying. The linear term in (3.112) is often higher than the recombination in the volume of the base. For gold doped pin-diodes with diffused junctions, Cooper [Coo83] found that the linear constant s_n in (3.112) is nearly proportional to the gold concentration. The linear recombination is caused possibly by an accumulation of recombination centers at the junctions. Its existence follows because the high-level lifetime in the base region determined from the carrier distribution is appreciably higher than the total linear recombination lifetime determined from stored charge measurements. See also Sect. 5.5.

3.5 Capacitance of pn-Junctions

In this section, we consider the capacitance of a pn-junction for small oscillations around a given stationary state at reverse bias or a small positive bias. The actual switching behavior of power devices consisting of large current and voltage changes between forward and reverse states will be discussed in later chapters. The small-signal capacitance in conjunction with an external stray inductance can be a source of disturbing oscillations in power electronic circuits (see Chap. 13). On the other hand, the capacitance is used as a tool for determining the doping concentration of the weakly doped base region of a junction. It is also used in different methods to investigate the properties of deep-level impurities which will be discussed in Sect. 4.9. In the present Section, however, the concentration of deep impurities is neglected against the density of the normal doping. The doping atoms are assumed to be completely ionized.

In a reverse biased pn-junction, the capacitance per unit area is defined as $c_j = dQ/dV_r$, where dQ is the variation of charge per unit area caused by an incremental change dV_r of the reverse voltage. Assuming a p$^+$n-junction the positive charge in depletion approximation , is given by $dQ = q \cdot N_D \cdot dw$, where N_D denotes the doping density of the base region and dw the change of the width w of the space charge layer caused by the voltage dV_r. The latter can be expressed according to Eq. (3.58) as $dV_r = q \cdot N_D \cdot w \cdot dw/\varepsilon$. This results in $c_j = dQ/dV_r = \varepsilon/w$, a well-known formula for a planar parallel plate capacitor. Substituting (3.58) for w, one obtains the capacitance per unit area as function of the reverse voltage:

$$c_j = \sqrt{\frac{q \cdot \varepsilon \cdot N_D}{2 \cdot (V_{bi} + V_r)}} \tag{3.113}$$

or

$$\frac{1}{c_j^2} = \frac{2 \cdot (V_{bi} + V_r)}{q \cdot \varepsilon \cdot N_D} \tag{3.114}$$

According to this equation, a plot of $1/c_j^2$ versus V_r yields a straight line, whose slope can be used to determine the doping concentration N_D. This measuring method has the advantage compared with other methods that it allows the direct determination of donor and acceptor doping concentrations down to 1×10^{13} cm^{-3}, without using any other quantities and material characteristics. Because the capacitance according to (3.113) refers to a depleted space charge region, it is called depletion capacitance.

If we assume for example $N_D = 1 \times 10^{14}$ cm^{-3}, $V_r = 600$ V and a junction area $A = 2$ cm^2, the capacitance $C_j = A \cdot c_j$ amounts to 235 pF. With a stray inductance $L_{par} = 50$ nH in series, this capacitance gives rise to oscillations with frequency $f = 1/(2\pi\sqrt{L_{par} \cdot C_j}) = 46.4$ MHz. Note that this far above the usual switching frequency in power electronic circuits.

As illustrated in Fig. 3.5, the space charge in abrupt asymmetrical junctions can be dominated by the charge of mobile carriers which is neglected in the depletion approximation for the capacitance (3.113). The exact expression for the maximum field strength in an abrupt junction is given by Eq. (3.24). Since from the Poisson equation the absolute charge of one sign in the space charge region is obtained as $Q = \int|\rho|dx = \varepsilon E_m$, Q follows immediately from (3.24). Assuming again a p$^+$n junction with $N_A \gg N_D$, the exponential term in V_N can be neglected. Expressing furthermore the linear term in V_N according to (3.22) by V_P one obtains:

$$Q = \sqrt{2\varepsilon kTN_A\left[\exp\left(-\frac{qV_P}{kT}\right) + \frac{qV_P}{kT} - 1\right]} \tag{3.115}$$

The voltage V_P over the N_A region is given by Eq. (3.23), where the built-in voltage has to be replaced at reverse bias by $V_{bi} + V_r$. Hence, Q is given as an explicit function of V_r which by differentiation yields the capacitance. From (3.23) one obtains $dV_P/dV_r = N_D/N_A$ and $qV_P/kT = 1 + \chi$ where

$$\chi \equiv \frac{N_D}{N_A} \cdot \frac{q}{kT} \cdot (V_{bi} + V_r) \tag{3.116}$$

Since $q(V_{bi} + V_r)/kT \gg 1$, the variable χ can be comparable with 1, although $N_D/N_A \ll 1$. The differential capacitance per unit area is then obtained from (3.115) as

$$c_j = \frac{dQ}{dV_r} = \sqrt{\frac{q\varepsilon N_D}{2(V_{bi} + V_r)}} \cdot \frac{1 - e^{-1-\chi}}{\sqrt{1 + e^{-1-\chi}/\chi}} \tag{3.117}$$

c_j decreases with increasing N_A, i.e. decreasing χ. For $\chi > 2$, a not too small N_D/N_A and a sufficiently high reverse voltage V_r, only the depletion meaning expression of (3.113) remains. With decreasing χ the capacitance decreases significantly and for $\chi << 1$ it is given by

$$c_j = 1.042\sqrt{\chi} \cdot c_{j,dpl} \tag{3.118}$$

where $c_{j,dpl}$ is now the depletion capacitance per area as given by (3.113). At appropriate N_D/N_A *the capacitance of abrupt* p$^+$n and n$^+$p *junctions at zero bias is up to two orders of magnitude smaller than the depletion capacitance*. In Fig. 3.22 the inverse capacitance of p$^+$n junctions in Si is plotted versus the square root of $V_{bi} + V_r$ for $N_D = 1 \times 10^{14}$ cm^{-3} and some values of N_A. For $N_A/N_D \leq 10^3$ the dependency passes soon into the straight line depicted for the depletion capacitance. For $N_A = 1 \times 10^{18}$ cm^{-3} and the more so for 4×10^{18} cm^{-3} the voltage dependence is over a wide or the whole range significantly weaker than that of the depletion capacitance. The mentioned method of determining the doping concentration N_D from the voltage dependence of c_j is hence not applicable in these cases.

The condition of sharp abruptness assumed in the formulae is realized often by junctions in SiC and other wide-gap semiconductors, in silicon junctions made by low-temperature epitaxy are very abrupt. In most diffused junctions the doping transition is on the contrary slow enough that carriers from the highly doped zone do not form a considerable part of the space charge; hence the depletion approach

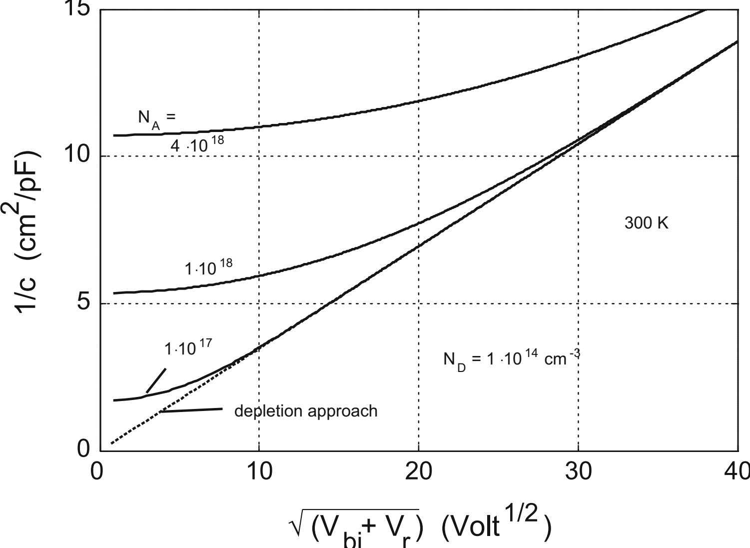


Fig. 3.22 Inverse capacitance of abrupt p$^+$n junctions in Si plotted versus square root of the voltage $V_{bi} + V_r$. The doping concentrations of the p$^+$ region is used as parameter

applies. Diffused junctions exhibit therefore a far higher capacitance than comparable abrupt high-low junctions, a noteworthy advantage of the latter.

Putting $V_r = -V_F$, the equations apply also to the capacitance at forward bias, if the voltage remains below several kT/q. At higher V_F the capacitance is dominated by a rearrangement of the minority carrier concentrations in the neutral regions. Concerning this capacitance contribution called 'diffusion capacitance' we refer to the literature [Sze02].

References

[Bar09] Bartsch, W., Schoerner, R., Dohnke, K.O.: Optimization of bipolar SiC-diodes by analysis of avalanche breakdown performance. In: Proceedings of the ICSCRM 2009, paper Mo-P-56 (2009)

[Bur75] Burtscher, J., Dannhäuser, F., Krasse, J.: Die Rekombination in Thyristoren und Gleichrichtern aus Silizium: Ihr Einfluß auf die Durchlaßkennlinie und das Freiwerdezeitverhalten. Solid St. Electron. **18**, 35–63 (1975)

[Coo83] Cooper, R.N.: An investigation of recombination in gold-doped pin rectifiers. Solid St. Electron. **26**, 217–226 (1983)

[Dav38] Davydov, B.: The rectifying action of semiconductors. Techn. Phys. UdSSR **5**, 87–95 (1938)

[Ful67] Fulop, W.: Calculation of avalanche breakdown voltages of silicon pn-junctions. Solid State Electron. **10**, 39–43 (1967)

[Lan79] Lanyon, H.P.D., Tuft, R.A.: Bandgap narrowing in moderately to heavily doped silicon. IEEE Trans. Electron. Devices **ED-26**(7), 1014–1018 (1979)

[Mil57] Miller, S.L.: Ionization rates for holes and electrons in silicon. Phys. Rev. **105**, 1246–1249 (1957)

[Mol64] Moll, J.L.: Physics of Semiconductors. McGraw Hill, New York (1964)

[Mor60] Morgan, S.P., Smits, F.M.: Potential distribution and capacitance of a graded p-n junction. Bell Syst. Tech. J. **39**, 1573–1602 (1960)

[Oga65] Ogawa, T.: Avalanche breakdown and multiplication in silicon pin junctions. Japan. J. Appl. Phys. **4**, 473–484 (1965)

[Rup14] Rupp, R., Gerlach, R., Kabakow, A., Schörner, R., Hecht, C., Elpelt, R., Draghici, M.: Avalanche behaviour and its temperature dependence of commercial SiC MPS diodes: Influence of design and voltage class. In: Proceedings of the 26th ISPSD, pp 67–70 (2014)

[Sch38] Schottky, W.: Halbleitertheorie der Sperrschicht. Naturwissenschaften **26**, 843 (1938)

[Sch39] Schottky, W.: Zur Halbleitertheorie der Sperrschicht- und Spitzengleichrichter. Zeitschrift für Physik **113**, 376–414 (1939)

[Sco69] Schlangenotto, H., Gerlach, W.: On the effective carrier lifetime in psn-rectifiers at high injection levels. Solid St. Electron. **12**, 267–275 (1969)

[Sco79] Schlangenotto, H., Maeder, H.: Spatial composition and injection dependence of recombination in silicon power device structures. IEEE Trans. Electron. Dev. **Ed-26** (3), 191–200 (1979)

[Shi59] Shields, J.: Breakdown in silicon pn-junctions. Journ. Electron. Control **6**, 132–148 (1959)

[Sho49] Shockley, W.: The theory of p-n junctions in semiconductors and p-n junction transistors. Bell Syst. Techn. J. **28**, 435–489 (1949)

[Sho50] Shockley, W.: “The Theory of pn Junctions in Semiconductors”, in Electrons and Holes in Semiconductors. D. van Nostrand Company Inc, Princeton (1950)

[Slo76] Slotboom, J.W., De Graaff, H.C.: Measurements of bandgap narrowing in Si bipolar transistors. Solid St. Electron. **19**, 857–862 (1976)

[Sze02] Sze, S.M.: Semiconductor Devices, Physics and Technology, 2nd edn. Wiley, New York (2002)

[Wag31] Wagner, C.: Zur Theorie der Gleichrichterwirkung. Physikaliche Zeitschrift **32**, 641–645 (1931)

[Wul60] Wul, B.M., Shotov, A.P.: Multiplication of electrons and holes in p-n junctions. Solid State Phys. Electron. Telecommun. **1**, 491–497 (1960)

Chapter 4
Introduction to Power Device Technology

In the following some basic aspects of power device production technology will be described. A considerable part is devoted to the different methods of doping, which form the heart of semiconductor technology.

4.1 Crystal Growth

The material for silicon devices must be of very high purity. Metallurgical silicon is converted to trichlorosilane $SiHCl_3$ which is liquid, and purified by fractional distillation. Especially chlorides of metals must be eliminated. By reduction of $SiHCl_3$ in a hydrogen atmosphere, polycrystalline rods of pure silicon are formed. For more details, see [Ben99].

The semiconductor material used for power device manufacturing must be monocrystalline. To produce such monocrystals there are two important processes.

In the *Czochralski process* (*CZ*) the crystal growth is performed in a crucible in which molten silicon is kept at a defined temperature. The required p- or n-type dopants are added to the melt. A small monocrystalline seed crystal is then dipped into the melt (Fig. 4.1). During slow rotation of the seed crystal, and of the crucible in the opposite direction, monocrystalline layers of silicon are deposited on the seed crystal maintaining the crystal structure of the seed. The growing bar is simultaneously pulled slowly upwards.

With the CZ process very large single crystals can be grown. Si cylinders of a length of several meters and a diameter of more than 30 cm for fabrication of 300 mm wafers are produced on an industrial scale, and wafers with diameter of 45 cm are introduced [Cla11, Lap08]. However, the purity and quality of single crystals is limited with the CZ-process, since the melt is in contact with the crucible during crystal growth. The oxygen content in CZ-silicon is typically $>10^{17}$ cm^{-3}, and also the content of impurity carbon is in the same range. CZ-wafers are mainly used as substrates for epitaxial wafer growth (see Sect. 4.3), from which integrated

© Springer International Publishing AG 2018

J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8_4

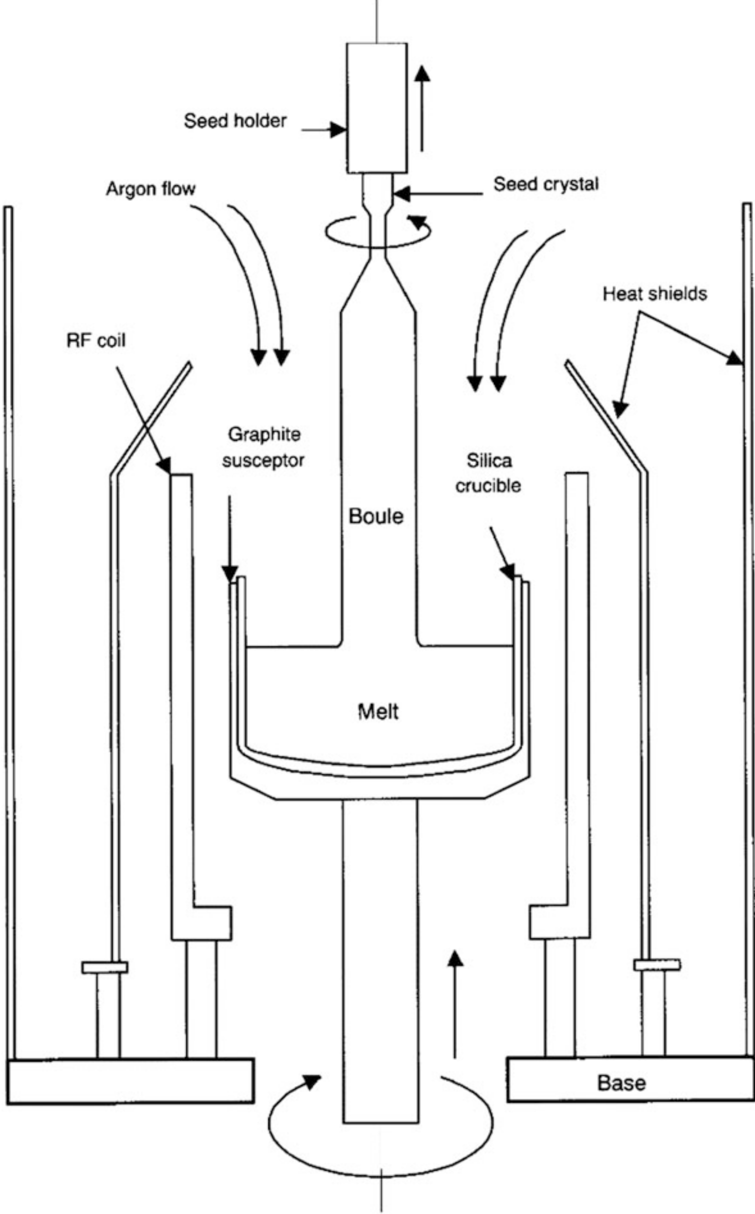


Fig. 4.1 Czochralski-process for production of single crystalline silicon rods. Figure taken from [Ben99]. Reprint with permission of John Wiley & Sons, Inc

circuits etc. are produced. Some power devices like MOSFETs are also produced from epitaxial wafers using CZ-substrates. For power devices in which the volume of the wafer is used, in most cases the purity of CZ-crystals was not sufficient. However, there is strong progress.

Meanwhile, 300 mm magnetic Czochralski Wafers are used for the fabrication of IGBTs [Scu16]. Regarding production effort, this is a large progress. Steady magnetic fields and alternating magnetic fields are used to damp convective flow and to stabilize the conditions of crystal growth [Gal02].

The *Float-Zone (FZ) process* is a crucible free crystal growth method. The starting material is a rod of high purity poly-silicon. A seed crystal is clamped in contact with the end of the rod (Fig. 4.2). An induction heating coil is placed around the rod, melting a small zone next to the seed crystal. While slowly moving the coil along the bar, the molten zone, starting at the interface between seed crystal and poly-silicon rod, passes slowly along the length of the bar. The poly-silicon melts and monocrystalline silicon is regrown with the same orientation as the seed crystal. Because of the absence of a crucible, crystals with higher purity and higher quality can be grown with the FZ process. Also since the solubility of the impurities is higher in the melted zone they tend to be accumulated there. The carbon content is $< 5 \times 10^{15}$ cm^{-3} and the oxygen content is $< 1 \times 10^{16}$ cm^{-3}. With the FZ method, monocrystalline silicon rods with a diameter of up to 20 cm can be produced.

Power devices, which use the whole volume of the wafer, have been mostly fabricated from FZ silicon over a long time because of the higher crystal quality and excellent purity. Most power devices which are manufactured with epitaxial techniques are using CZ silicon as material for the substrates.

In the FZ as well as in the CZ process, the doping of the crystal is done by adding dopants to the melt. Power devices normally require a thick low-doped middle region to support a large voltage. Most often an n-type region is preferred. The value of this doping depends on the voltage for which the power device is designed; see further Eq. (3.84) and Fig. 3.17. The doping processes will be discussed in more detail below.

After crystal growth follows a sawing process to slice the rod in single wafers. To remove the surface damage caused by sawing and to obtain a clean and undisturbed surface, the wafer surface layer is lapped mechanically and in some cases additionally etched chemically. For some semiconductor processes also wafers with single side polished surface are necessary.

4.2 Neutron Transmutation for Adjustment of the Wafer Doping

The homogeneity of the doping concentration is extremely important for power devices. If variations in doping concentration (or local defects) exist in the wafers, the current may be unevenly distributed, especially at avalanche breakdown. This can in

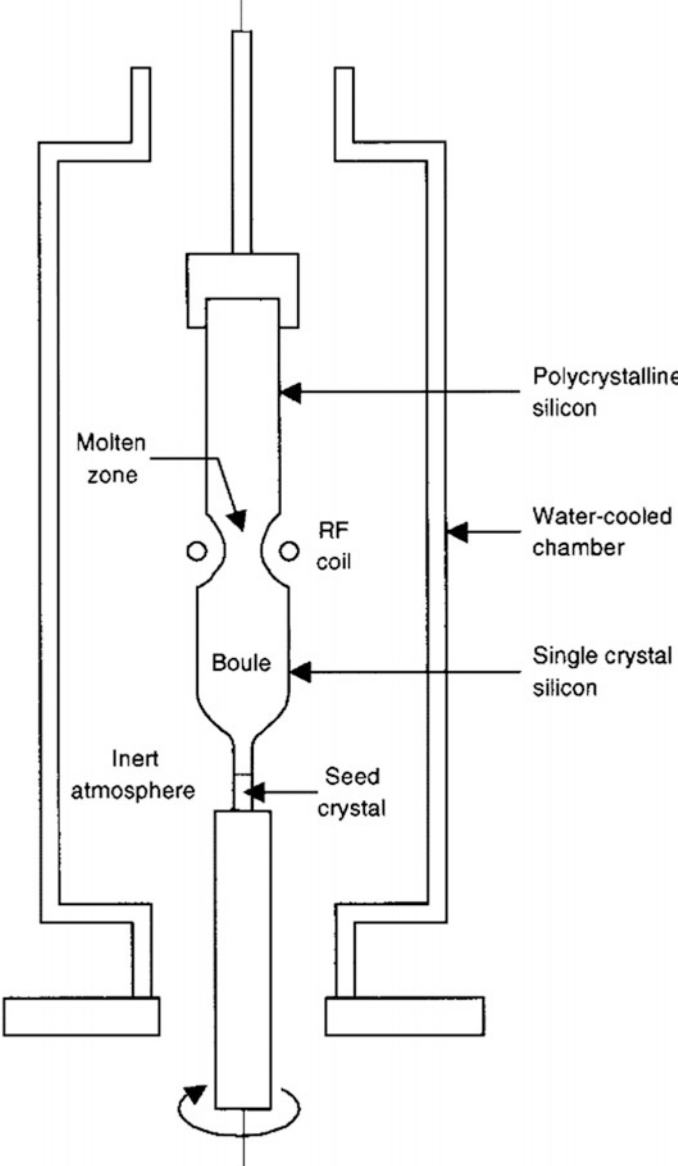


Fig. 4.2 Float-Zone (FZ) process for crystal growth. Figure taken from [Ben99]. Reprint with permission of John Wiley & Sons, Inc

turn result in local overheating and device failure. However, even with the FZ process, which alone is suited for the preparation of the extremely pure silicon crystals needed for power devices, fluctuations in doping cannot be avoided. The obtained wafers

show periodic concentric rings with doping variations called "striations". An example for the resistivity profile measured across a wafer is shown in Fig. 4.3

According to Eq. (2.34), the resistivity variations correspond to doping fluctuations in an inverse manner. Since the doping directly determines the blocking voltage, it is not possible to produce high voltage power devices with the required narrow design, if the doping of the starting material is fluctuating in the amount as shown in Fig. 4.3a. Only with the method of neutron doping (Fig. 4.3b), it became possible to realize for instance thyristors with a blocking voltage of > 2000 V.

The neutron transmutation doping (NTD) of silicon was proposed and investigated first by Lark-Horovitz [Lar51] and Tannenbaum and Mills [Tan59, Tan61]. For the fabrication of 5 kV-thyristors with an acceptable recovery time, the method was reinvented and introduced on industrial scale in 1973/74 [Haa76, Jan76, Scl74]. It is based on the radioactive transmutation of the silicon isotope ${}^{30}_{14}$Si to phosphorous ${}^{31}_{15}$P by thermal neutrons. This isotope is contained in natural silicon to 3.09%, besides ${}^{29}_{14}$Si with a share of 4.67% and the main isotope ${}^{28}_{14}$Si with 92.23%. As starting material a crystal bar is used which is grown by the FZ method and doped very low, typically below 5×10^{12} cm^{-3}. If the silicon rod is placed near the core of a nuclear reactor, the ${}^{30}_{14}$Si atoms capture partly a thermal neutron and turn under emission of a γ quant into the unstable isotope ${}^{31}_{14}$Si which by emission of an electron (β particle) decays then into the stable phosphorous isotope ${}^{31}_{15}$P:

$$ {}^{30}_{14}\mathrm{Si}(\mathrm{n},\gamma) \rightarrow {}^{31}_{14}\mathrm{Si} \rightarrow {}^{31}_{15}\mathrm{P} + \beta $$

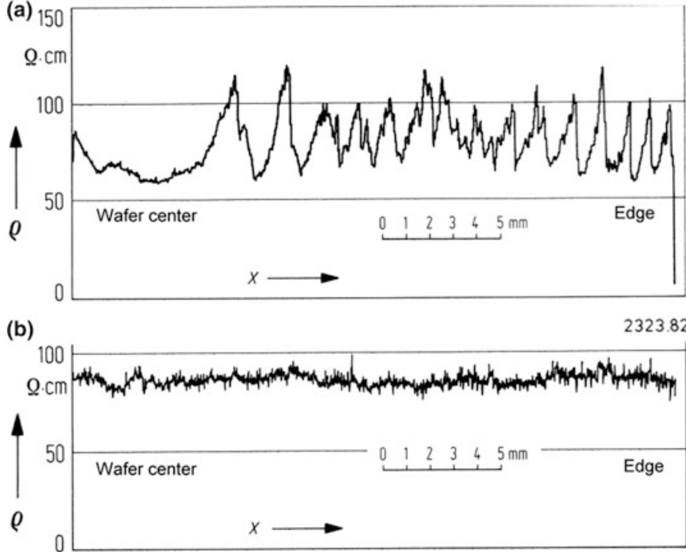


Fig. 4.3 Radial resistivity profile achieved by conventional doping (top) and by neutron transmutation doping (bottom). Figure from [Sco82]

The decay of ^{31}Si to ^{31}P has a half-time of 2.6 h. Thermal neutrons exhibit a decay length in Si of 19 cm, while the diameter of the bars is usually 10 to 20 cm. By rotation of the bar during the irradiation, however, the mean neutron flux in the Si bar becomes nearly homogeneous, the difference between the value at the axis to that at the periphery being less than 4% even for a rod diameter of 15 cm [Jan76]. The final phosphorous concentration obtained by these processes is given by

$$N_{Phos} = c \cdot \Phi \cdot t \tag{4.1}$$

where Φ denotes the neutron flux in cm^{-2}s^{-1} and t the time. With the cross section data given in [Jan76], the constant c for natural silicon is obtained to be $c = 2.0 \times 10^{-4}$ cm^{-1}. The typical flux density Φ of thermal neutrons at the used positions is in the range 10^{13}–10^{14} cm^{-2}s^{-1} [Amm92]. At a flux of 2×10^{13} cm^{-2}s^{-1}, for example, an irradiation time of 3.5 h is necessary according to (4.1) to obtain a phosphorous doping of 5×10^{13} cm^{-3}. The phosphorous concentration can be adjusted by this method very exactly within limits of only ±3%.

After the neutron irradiation a storage period is necessary for the on-going transmutation of ${}^{31}_{14}$Si into ${}^{31}_{15}$P and particularly for the decay of the activity due to the β emission. The time for the decay of the activity to an undetectable level is 3–5 days. As secondary process, the reaction of the generated ${}^{31}_{15}$P atoms with the neutrons and the decay of the resulting ${}^{32}_{15}$P according to

$${}^{31}_{15}\mathrm{P}(\mathrm{n},\gamma) \rightarrow {}^{32}_{15}\mathrm{P} \rightarrow {}^{32}_{16}\mathrm{S} + \beta$$

has to be considered. The decay of ${}^{32}_{15}$P into ${}^{32}_{16}$S has a half time of 14.3 days and will prolong the time required for the activity decay if the phosphorous concentration generated is higher than about 1×10^{15} cm^{-3} [Jan76]. To eliminate lattice defects induced by the radiation, the crystal bar is annealed at a temperature of 800 °C, before it is further processed to wafers.

Because the neutron doping method is rather costly there have been major efforts of silicon wafer suppliers to improve the FZ crystal growth process, and essential advances have been reached in the last 20 years. The doping tolerances of wafers with a diameter of 200 mm are decreased now to ±12% [SIL06]. Nevertheless, the NTD method is still indispensable for high voltage devices.

4.3 Epitaxial Growth

If a Si device requires a lowly-doped layer with a thickness below 50–100 µm, total wafer is thin and the wafer handling is challenging, especially regarding the breakage rate in the production process, and also the wafer must have a low bow to realize fine pattern structures with photolithography. Therefore, epitaxial growth of a thin n^{-} layer was usual. In 1997, however, so so-called “thin-wafer technology” was introduced [Las97] for a wafer thickness with 100 µm and below. Even 40 µm

thin-wafer technology was introduced for 400V IGBTs and diodes [Boe11]. The thin-wafer technology competes successfully with epitaxial growth. For unipolar devices with n-layers as thin as possible, epitaxial growth is still standard.

Epitaxial growth is an alternative method to produce high purity single crystalline layers. As a starting material or substrate for the epitaxial growth process, a single side polished CZ-wafer is used. On this wafer a layer with higher purity and higher crystal quality is then grown in the same crystal orientation as the substrate, see Fig. 4.4. The epitaxy process is done significantly below the melting temperature of the semiconductor material.

The growth of epitaxial layers from silicon is made in an enclosed chamber, a reactor, using a vapor phase process. There are different possible processes which use one of the following chemical reactions [Ben99]:

$$\begin{array}{llll} SiCl_4 + 2H_2 & \rightarrow & Si + 4HCl & \text{at } 1150 - 1220\,^{\circ}C \\ SiHCl_3 + H_2 & \rightarrow & Si + 3HCl & \text{at } 1100 - 1175\,^{\circ}C \\ SiH_2Cl_2 & \rightarrow & Si + 2HCl & \text{at } 1025 - 1100\,^{\circ}C \\ SiH_4 & \rightarrow & Si + 2H_2 & \text{at } 1150 - 1220\,^{\circ}C \end{array}$$

Prior to the growth an intensive mechanical and chemical cleaning of the wafer surfaces is done. The substrates are also cleaned and etched in the process reactor at a temperature of 1140–1240 °C with HCl. The growth is executed in a H_2-atmosphere. For doping, PH_3 (phosphine) or B_2H_6 (diborane) for n-type or p-type, respectively, are added to the hydrogen gas in controlled proportions.

Epitaxial layers in silicon can be fabricated with high purity, especially with very low impurity content of carbon and oxygen.

For the production of SiC monocrystals a high effort is necessary. The SiC substrate crystal growth is done at a temperature of 2300 °C. It was very difficult to achieve monocrystals of low defect density, however there is strong progress. Nitrogen-doped substrates as 150 mm wafer with resistivity of 15–28 mΩ cm are available. Former SiC substrates suffered of a high number of crystal defects, especially dislocations. Screw dislocation propagating through the bulk at the

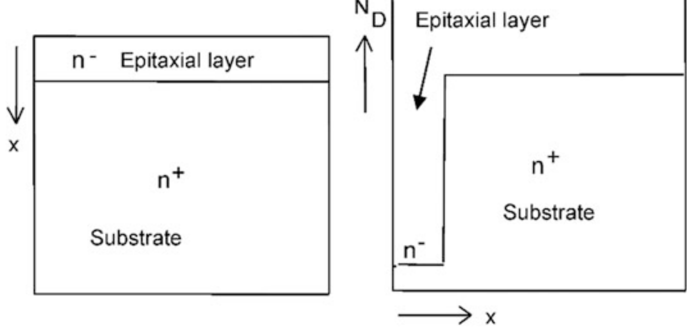


Fig. 4.4 Epitaxially grown wafer in cross section. Example n^+-substrate with n^- epitaxial layer

crystal growth formed micropipes which continue in subsequent epitaxial layers. The former high micropipe density is now significantly improved and micropipe-free wafers are found. The specification at manufacturers is meanwhile < 1 micropipe/cm^2. 6-inch wafers are available, the possibility of 8-inch wafer diameter was demonstrated.

For SiC power devices, a following epitaxy process is always necessary to reach the desired material quality. This epitaxial process is done at 1400–1600 °C under H_2-atmosphere. Commercial SiC CVD processes typically use silane SiH_4 and light hydrocarbons, such as propane or ethylene, and hydrogen as a carrier gas. Growth rates are usually 6–7 μm/h [Cho11]. Using $SiCl_2$ as primary growth species, growth rates of 30–100 μm/h have been obtained, 5–15 times higher than most conventional growths. The deposition temperature was 1750 °C, a good crystal quality and minority carrier lifetime in the order of 0.75 μs were achieved [Cho11]. Also a carrier lifetime in the order of 1–2 μs is reported. Basal plane dislocation (BPD) densities are an obstacle for device fabrication, they are today in the order of 1–2 cm^{-2} and limit the yield for large area devices. For bipolar devices, BPDs are critical defects, since they replicate from the substrate into the epitaxial layer. During bipolar operation the BPDs expand to form stacking faults which degrades the bipolar current capability (bipolar degradation). Much work is done to reduce the BPD density [ECP16]. However, the current methods used to reduce BPDs and to increase lifetime are an high effort, and particularly large area SiC wafers of high quality are therefore very expensive compared to Si.

Due to the lack of GaN large-size monocrystals, GaN is grown by epitaxy on other substrates – Si, SiC and sapphire. Sapphire is an insulator with limited thermal conductivity and only used for microwave devices, not for power devices. Due to cost reasons, Si substrates are preferred, even if devices from SiC substrates show better characteristics [Hil15].

Heterostructures are grown by metal organic chemical vapor deposition (MOCVD) and plasma induced molecular beam epitaxy (PIMBE). In the MOCVD process Triethylgallium $Ga(C_2H_5)_3$ and Trimethylalluminium $(CH_3)_3Al$ are used as source and ammonia as precursor [Amb99]. With moderate growth temperature of 740–780 °C and high ammonia flux, wafers with GaN on Si (111) substrates can be grown with high resistivity [Tan10]. By the share of Triethylgallium and Trimethylalluminium, pure GaN, mixed composition $Al_xGa_{1-x}N$ ore pure AlN can be produced, where x denotes the share of Al in relation to GaN. For more details, see Sect. 4.10.

4.4 Diffusion

The most important way to create n- and p-layers of defined thickness is by diffusion of the impurities into the solid at high temperature. This method to control conductivity was introduced in the early 1950s, and since then various techniques have been explored to improve uniformity and reproducibility. To a high degree,

these requirements can be met carrying out the diffusion with an impurity layer deposited before by ion implantation (see Sect. 4.5). Providing the wafer with a masking oxide pattern (see Sect. 4.6), the diffusion can be restricted to windows in the layer. Due to these opportunities the diffusion is a key process in manufacturing of devices.

For diffusion, semiconductor wafers are placed in a furnace, where they are brought in contact with the dopant, if not only a drive-in of pre-deposited impurities is intended. Often an inert gas which contains the impurity as gaseous molecules, e.g. BCl_3 or PH_3, flows through a tube with the wafers, at which surface the doping atoms are released in a reaction. However, also other techniques are in use [Mue93]. The temperature used for dopants like B and P in silicon ranges from about 1000–1250 °C. If the surface is covered with a masking pattern, the diffusion at the edges of a window proceeds not only normally to the surface but also laterally under the oxide to a certain extent; it is a two-dimensional or, near a corner, a three-dimensional process. Since both the vertical and lateral diffusion are decisive for the design of devices, a precise knowledge of impurity distribution in dependence of diffusion conditions is a necessity. On atomic scale, diffusion in crystals is a complicated process involving crystal defects such as vacancies (vacant lattice sites) and interstitial atoms located between lattice points. Nevertheless the macroscopic description is normally the same as for diffusion in gases and fluids. The differential equation of diffusion is a special case of the equations for electrons and holes developed in Sect. 2.9. In the next subsection, we derive the diffusion equation however separately, before using it to calculate some basic and practically important diffused impurity distributions.

4.4.1 Diffusion Theory, Impurity Distributions

If the impurity concentration N is inhomogeneous, a diffusion current with density

$$\vec{J} = -D \cdot gradN(x, y, z, t) \tag{4.2}$$

results, as described already in Sect. 2.6.3. D is the diffusion constant or 'diffusivity' which for most atoms in silicon becomes appreciable only above 800 °C. If also $\vec{J}$ varies in space, this leads in turn to a variation of N with time, as is illustrated for one dimension in Fig. 4.5: If the current density J_2 flowing out of a volume element of thickness dx is higher than the current J_1 flowing in, i.e. if $\partial J/\partial x > 0$, then the particle concentration in the volume element decreases with time (Fig. 4.5b). Here we use that, different from electrons and holes, impurities cannot be created and not disappear (zero generation/recombination). Hence the net flow out of the volume element, $\partial J/\partial x$, yields directly the decay rate of particle concentration: $-\partial N/\partial t = \partial J/\partial x$ or in general form:

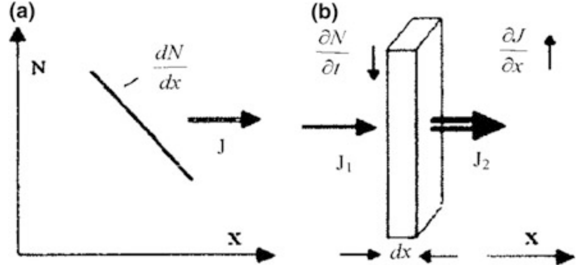


Fig. 4.5 Illustration of the relation between the **a** gradient of the particle concentration and particle current density (first Fick's law) and **b** divergence of particle current density and change of particle concentration with time (continuity equation, second Fick's law)

$$-\frac{\partial N}{\partial t} = div\,\vec{J} \tag{4.3}$$

Inserting (4.2) one obtains the 'diffusion equation':

$$\frac{\partial N}{\partial t} = div(D \cdot gradN) \tag{4.4}$$

Equations (4.2) and (4.4) are called Fick's first and second law, respectively, since formulated first by A. Fick in 1855. Together with the initial and boundary conditions Eq. (4.4) describes how a spatial doping distribution develops with time. For not too high concentrations, the diffusivity D is independent of N and the space coordinates. Then Eq. (4.4) takes the form

$$\frac{\partial N}{\partial t} = D \cdot \Delta N \tag{4.5}$$

with the Laplace operator $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2$, if two dimensions are involved. The impurity level up to which this applies depends on the intrinsic concentration n_i at the diffusion temperature, which at 1100 °C for example is about 3×10^{19} cm^{-3} (see Sect. 4.4.3). Most analytical calculations, including the following, deal with the solution of Eq. (4.5).

We consider now a few important cases of diffusion into a half-space $x \geq 0$.

(a) Drive-in diffusion of a thin impurity layer deposited in the semiconductor near the surface: Diffusion processes of this kind are often used in device fabrication. The impurity layer can be introduced by ion implantation or a short prior diffusion. The layer is assumed to be infinitesimal thin and located at $x = 0$, described by a δ-function with integral value S, the deposition dose per area. The solution of the one-dimensional form of (4.5)

$$\frac{\partial N}{\partial t} = D \cdot \frac{\partial^2 N}{\partial x^2} \tag{4.6}$$

must satisfy the initial condition N(x, 0) = 0 for $x > 0$ and the condition

$$\int_0^\infty N(x,t)dx = S = \text{const} \tag{4.7}$$

for all times $t \geq 0$. Equation (4.7) follows, since after the prior deposition no further impurities are introduced and hence their total number remains constant, assuming that out-diffusion into the atmosphere does not occur. The solution of (4.6) which satisfies these conditions is

$$N(x,t) = \frac{S}{\sqrt{\pi \cdot D \cdot t}} \cdot \exp\left(-\frac{x^2}{4Dt}\right) \tag{4.8}$$

As function of x, (4.8) represents a so-called Gaussian distribution. By integration one can verify the condition (4.7). The decrease of the surface concentration $N_0 = S/\sqrt{\pi Dt}$ with time is compensated by the increase of the depth x at which the normalized distribution $N(x, t)/N_0$ takes a given value. With the diffusion length

$$L = 2 \cdot \sqrt{D \cdot t} \tag{4.9}$$

equation (4.8) takes the form

$$N(x,t) = \frac{2}{\sqrt{\pi}}\frac{S}{L}\exp\left(-\left(\frac{x}{L}\right)^2\right) \tag{4.10}$$

A diffusion profile of this type is shown in the earlier Fig. 3.6. From (4.10) the deposition dose and diffusion length needed for the surface concentration and junction depth of that figure can be easily determined. How the distribution (4.10) develops with time is shown in Fig. 4.6 by the solid curves. The x-value at which a certain concentration $N_B < N_0$ is reached, is identical with the junction depth x_j of a pn-junction which is formed if the wafer possesses a background doping of opposite polarity with density N_B. As is shown by the figure, x_j increases somewhat weaker with L than proportional. From (4.10) one obtains

$$x_j = L \cdot \sqrt{\ln\left(\frac{2S}{\sqrt{\pi} \cdot L \cdot N_B}\right)} = L \cdot \sqrt{\ln\left(\frac{N_s}{N_B}\right)} \tag{4.11}$$

Within a limited range of L, the root expression can be considered approximately constant. For the range $10^3 < 2S/(\sqrt{\pi} \cdot L \cdot N_{vol}) < 10^5$ one obtains as a rule of thumb $x_j \approx 6 \cdot \sqrt{D \cdot t}$.

In reality, the thickness Δx of the pre-deposition layer is of course not zero, but may be for example 200 or 600 nm. In the surface region, the solution (4.7) is then valid only for times with $L_D \gg \Delta x$, i.e. $t \gg \Delta x^2/(4D)$. For impurity layers

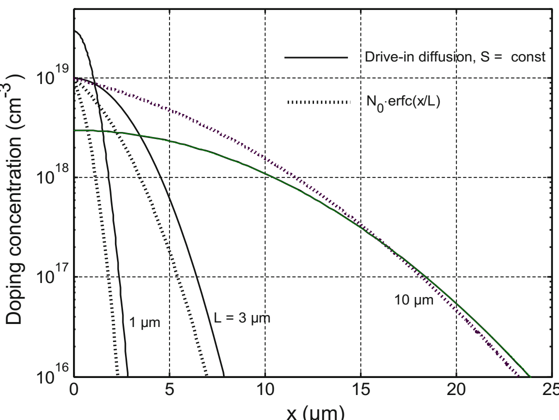


Fig. 4.6 Doping profiles for diffusion with constant integral number $S = 2.66 \times 10^{15}$ cm^{-2} per area (solid curves) and constant surface density (dotted curves). Assuming a diffusion constant $D = 1 \times 10^{-12}$ cm^2/s = 0.36 µm^2/h (for P and B at 1200 °C, see Fig. 4.11), the three L-values correspond to diffusion times $t = 0.694$, 6.25 and 69.4 h, respectively

produced by ion implantation, however, Eq. (4.10) can be modified in a simple manner to include the distribution in the layer at small t (see Sect. 4.5).

(b) Diffusion with constant surface concentration: A constant impurity concentration at the semiconductor surface during the diffusion process can be achieved by supplying the dopant as admixture to an inert gas flowing around the wafers or from vapor phase of the dopant. Another possibility is to deposit a solid layer with higher dopant concentration than its solubility in the semiconductor. The surface density is then maintained at the solubility value. The solution of (4.6) which satisfies the boundary condition $N(0,t) = N_s = \text{const}$ for $t > 0$ and the initial condition $N(x,0) = 0$ for $x > 0$ is given by

$$N(x,t) = N_s \cdot erfc\left(\frac{x}{2\sqrt{D\,t}}\right) \tag{4.12}$$

with the complementary error-function defined as[1]

$$erfc(x) = 1 - \frac{2}{\sqrt{\pi}} \int_x^x e^{-x'^2} dx' \tag{4.13}$$

[1]This function and the normal error function erf(x) = 1 – erfc(x) appear also in other diffusion processes (see next section, case c). Often the following analytical approximation for x $\geq$ 0 is sufficient:

$$erfc(x) \approx \exp(-1.14x - 0.7092x^{2.122})$$

Its maximum error in the range 10^{-7} < erfc < 1 is 2‰. For negative arguments appearing in Eq. (4.16) below the approximation can be used considering that erf(–x) = –erf(x) = erfc(x) – 1.

As is shown by (4.12), the doping density is only a function of x/L in this case. Profiles given by (4.12) are shown in Fig. 4.6 as dotted curves, using the same L-values as for the curves with constant number S of impurities discussed in case a). For equal surface concentration and L-value, the *erfc*-profile penetrates less deep into the crystal than the Gaussian distribution, as shown for $L = 3\ \mu$m. This may be surprising at first sight, but follows because the concentration in the case of drive-in develops from higher values at smaller times, as can be seen from the curves for $L = 1\ \mu$m. Later, for $L = 10\ \mu$m, the *smaller* surface concentration of the drive-in distribution leads to *reduced* penetration depth. To get deep-reaching diffusion profiles with simultaneously high surface density, a diffusion with constant surface density is advantageous, since otherwise a very high deposition dose S would be necessary.

The integral impurity number per area increases now with time. Integrating (4.11) over x one obtains:

$$S = \int_0^\infty N(x,t)dx = N_s \cdot 2\sqrt{Dt}/\sqrt{\pi} \tag{4.14}$$

since $\int_0^\infty erfc(x)dx = 1/\sqrt{\pi}$. If a short diffusion with constant surface concentration is used for pre-deposition, Eq. (4.14) yields the deposition dose to be used in (4.8) for drive-in.

The diffusion case with constant surface concentration is used besides others for comparison of theory and measurements. It is found, that the experimental profiles of the group III and V dopants are in agreement with Eq. (4.12), as long as the surface concentration is below about 1×10^{19} cm^{-3}.

(c) Drive-in diffusion with a deposition layer of limited width: In this case the diffusion proceeds not only vertically to the surface but near the edges of the window also sideways or laterally under the mask (Fig. 4.7). The lateral diffusion is generally not undesirable but on the contrary widely used in device concepts. A well-known fact is that the lateral penetration depth under the oxide, y_{pd}, is always smaller than the vertical penetration depth x_{pd} in the central region of the window. In most cases, the relation y_{pd}/x_{pd} is in the range $\approx 0.6 - 0.9$. The two-dimensional impurity distribution is described often in a simplified manner, using for the region below the window, $-d \leq y \leq d$, $x \geq 0$, the unchanged one-dimensional profile obtained without mask, whereas the lateral decay is modeled by a cylindrical distribution under the oxide. An exact mathematical analysis of diffusion through a window

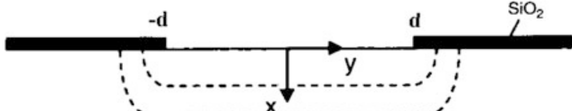


Fig. 4.7 Vertical and lateral diffusion through a window in a masking layer

limited on one side by a mask has been given both for the case of constant surface concentration as well as constant amount of impurities by Kennedy and O'Brien [Ken65]. In the following, the latter case, which is the more important and simultaneously simpler, is examined for a strip-like window.

As in case (a), the initial impurity layer is assumed to be infinitesimally thin and located at the surface $x = 0$, now however only in the stripe $-d \leq y \leq d$, $-\infty < z < \infty$. The integral of the $\delta(x)$-function representing the initial distribution is given by the doping dose S per area. Outside the window the concentration is everywhere zero at $t = 0$. Since no impurities are introduced later and losses by out-diffusion are excluded here, one has for all $t \geq 0$:

$$\int_{x=0}^{\infty} \int_{y=-\infty}^{\infty} N(x', y', t \geq 0) dx' dy' = 2d \cdot S = \text{const} \tag{4.15}$$

The solution of the differential Eq. (4.5) which satisfies these conditions is given by:

$$N(x, y, t) = \frac{S}{\sqrt{\pi} L} \cdot e^{-\left(\frac{x}{L}\right)^2} \cdot \left\{ erf\left(\frac{y+d}{L}\right) + erf\left(\frac{d-y}{L}\right) \right\}, \tag{4.16}$$

where as before $L = 2\sqrt{Dt}$ and the coordinate system of Fig. 4.7 is used. *erf* denotes the normal error-function defined as

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-x'^2} dx' = 1 - erfc(x) \tag{4.17}$$

The distribution (4.16) differs from (4.10) by the factor

$$F(y) = \frac{1}{2} \cdot \left(erf\left(\frac{y+d}{L}\right) + erf\left(\frac{d-y}{L}\right) \right) \tag{4.18}$$

which describes the lateral variation of N. $F(y)$ satisfies the one-dimensional diffusion equation in y as follows from the above case (b); similarly, the remaining factor in (4.16) was found under (a) to be a solution of (4.6). From this it follows in a simple manner that the product solves the two-dimensional Eq. (4.5). The integral of $F(y)$ from $-d$ to d is by the same amount, namely $L \int_0^{2d/L} erfc(u) du$, smaller than $2d$ as is contributed by the integral over the outer regions, so also the condition (4.15) is satisfied.

According to Eq. (4.16), $N(x, y)$ obeys at all points y the same Gaussian x-dependency (in relative units), including the region under the oxide. Likewise the y-dependency expressed by $F(y)$ is the same for all distances x from the surface. In Fig. 4.8, $F(y)$ is plotted for several diffusion lengths L assuming a window width

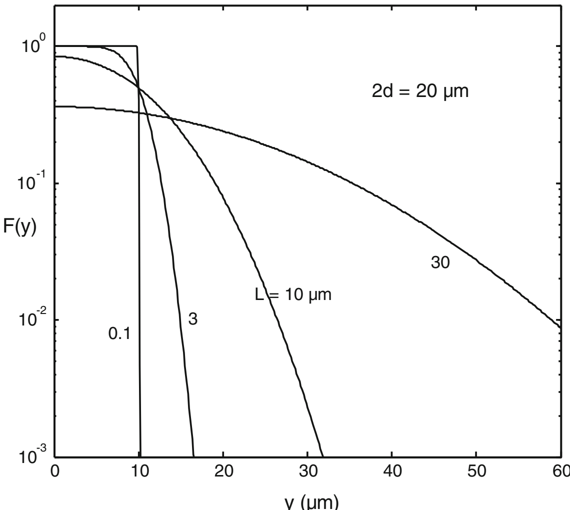


Fig. 4.8 Lateral distribution function $F(y)$ for various diffusion lengths L. Window width $2d = w = 20$ μm

$2d = w = 20$ μm. For $L = 0.1$ μm, $F(y)$ is box-shaped, meaning that there is virtually still no lateral diffusion. With increasing L this changes. As long as $w/L > 4$, the value of $F(y)$ in the center of the window at $y = 0$ is equal to 1 whereas up to the edges it decreases to ½. The above mentioned assumption of y–independent concentration in the window region hence fails considerably. For $y > d$, the concentration decreases as $0.5 \cdot erfc((y - d)/L)$ for $w/L > 2$, similar for $y < -d$. For long diffusion times, cases $L = 10$ and 30 μm in Fig. 4.8, also the concentration in the central region is reduced by the lateral diffusion. The reduction of the impurity amount in the window region by the lateral diffusion is a point to be considered selecting the diffusion conditions.

Curves of constant concentration N of the function (4.16) are shown for an example in Fig. 4.9. The window width is in a range where the mask at one side has no influence on the impurity distribution at the other side for the used diffusion length. The lateral decrease of the concentration within the window from the maximum value $N_0 = 2 \cdot S/(\sqrt{\pi} \cdot L) = 3.76 \times 10^{18}$ cm^{-3} is conspicuous. As can be seen, furthermore, the lateral penetration depth y_{pd}, defined as the distance from the edge of the window to the point of a given concentration $N < N_0/2$ at the surface is smaller than the vertical penetration depth x_{pd}, i.e. the vertical distance to this concentration deep in the window. The penetration depths coincide with the lateral and vertical junction depths $y_j(N_B)$, $x_j(N_B)$, if a background doping of opposite polarity with density $N_B = N$ is present.

Equation (4.16) involves an interrelation between x_{pd} and y_{pd}: With the origin of the y-coordinate at the window edge as in Fig. 4.9, the concentration at $x = 0$, $y = y_{pd}(N)$ is per definition equal to the concentration at $x = x_{pd}(N)$, $y < -2L$. Hence

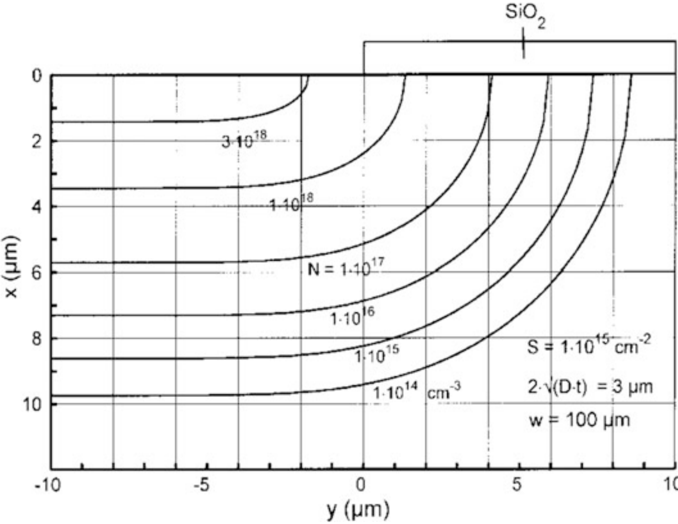


Fig. 4.9 Curves of constant concentration for a drive-in diffusion near the diffusion mask, according to Eq. (4.16). Deposition dose 1 × 10^{15} cm^{-2}, window width 100 µm, L = 3 µm

$$\frac{N}{N_0} = \frac{1}{2} erfc(y_{pd}/L) = \exp\left(-(x_{pd}/L)^2\right) \tag{4.19}$$

where $w/L > 4$ is assumed so that the window width drops out. Solving for x_{pd}/L one obtains for the ratio of penetration depths:

$$\frac{y_{pd}}{x_{pd}} = \frac{y_{pd}}{L} \Big/ \sqrt{\ln\left(\frac{2}{erfc\left(y_{pd}/L\right)}\right)}, \tag{4.20}$$

According to (4.20), the ratio y_{pd}/x_{pd} increases with increasing y_{pd}/L and hence with decreasing N/N_0. In Fig. 4.10, y_{pd}/x_{pd} is shown as function of N/N_0 using Eq. (4.19) to correlate y_{pd} with N. The variation of y_{pd}/x_{pd} is in accordance with the above-mentioned experimental observations.

In MOSFETs and IGBTs, the n-channel is formed over the rim of the p-base stemming from lateral diffusion. Equation (4.16) provides a relative simple tool to determine the diffusion conditions required for the appropriate channel length and doping concentration under the gate. In cases where the neglected loss of dopant by out-diffusion or possibly by oxide growth is significant or if the variation of the diffusivity with N is considerable, more complex numerical process simulation is needed.

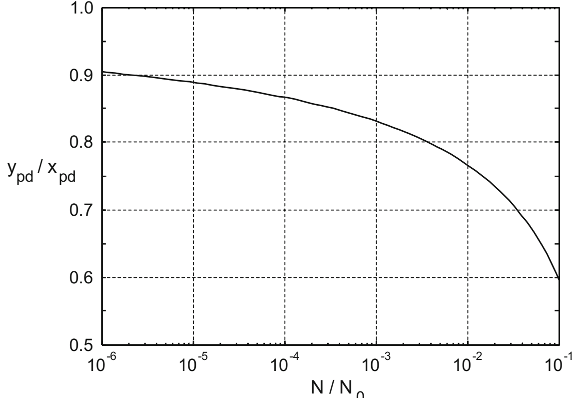


Fig. 4.10 Ratio of lateral to vertical penetration depth as function of the defining concentration N divided by the concentration N_0 in the middle of the window

4.4.2 *Diffusion Constants and Solubility of Dopants*

The diffusion of impurities is thermally activated, hence the temperature dependence of the diffusion constant is given by an Arrhenius relationship

$$D = D_0 \cdot e^{-E_A/kT} \tag{4.21}$$

where E_A is the activation energy and T the absolute temperature. Both E_A and the pre-factor D_0 can be considered as constant.

For the commonly used dopants in silicon the parameters in (4.21) are compiled in Table 4.1. The values are representative and apply to typical diffusion processes for power devices. Diffusion parameters in the literature differ partly considerably from one another, depending on the conditions of measurements. The activation energies fall in the range 3.3–4.2 eV. A plot of the diffusivities versus temperature calculated with these parameters is shown in Fig. 4.11. For a temperature rise of 100 °C, the diffusion constants increase approximately by an order of magnitude.

Aluminum is the fastest diffusing dopant, its diffusion constant is about two orders of magnitude higher than that of the slowest doping elements, As and Sb, and about a factor 20 higher than that of P and B, which are most commonly used for diffusion. P and B as well as also As and Sb have nearly equal diffusivity.

Table 4.1 Parameters of diffusion constants in the Arrhenius relationship (4.21)

Element	D_0 [cm^2/s]	E_A [eV]	Reference
B	0.76	3.46	[Sze88]
Al	4.73	3.35	[Kra02]
Ga	3.6	3.5	[Ful56]
P	3.85	3.66	[Sze88]
As	8.85	3.971	[Pic04]
Sb	40.9	4.158	[Pic04]

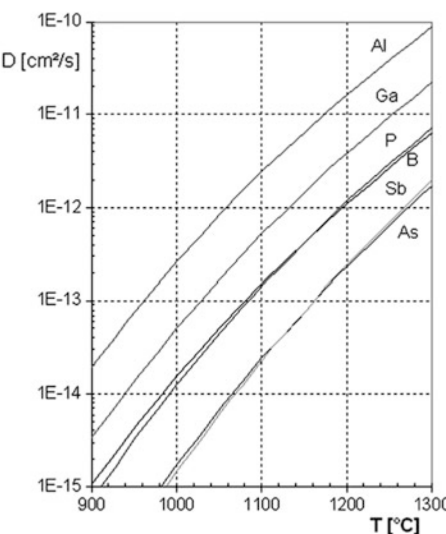


Fig. 4.11 Diffusion constants of dopants in silicon as a function of temperature

The usability of substances for diffusion (and generally doping) can be judged only if also the solubility in the semiconductor is taken into account. Figure 4.12 shows the solubility of impurities in silicon in dependence of temperature. Besides the normal dopants, also oxygen, Au and Pt are included. All elements are only little solvable. Related to the density of Si atoms, 5.0×10^{22} atoms/cm^3, even the highest solvable impurities, phosphorous and arsenic, have a solubility smaller than 1%. The solubilities of P and As given in the figure [Bor87] are a factor 3–4 smaller than often cited previous values [Tru60]. The maximum carrier concentration obtainable with a doping element approaches to a high share the new solubility limits, meaning that nearly all the doping atoms are electrically active. The solubility of electrically inactive dopants on interstitial sites or in complexes of more than one atom is only small. Apart from a region near the melting point of Si the solubilities decrease with decreasing temperature. This can lead to precipitation if the system is slowly cooled down.

Whereas Al has the highest diffusivity, its solubility is the smallest of group III and V elements. Al is used for creating deep pn-junctions especially in thyristors in which pn-junctions with a depth up to 100 µm are used. Its solubility, however, is not sufficient for a surface concentration necessary for good ohmic contacts. Therefore, when used for deep diffusions, a second and for some thyristors even a third p-type diffusion step is carried out: The Al diffusion is followed by a Ga diffusion, and since also Ga has, next to a good diffusivity, only a moderate solubility, some manufacturers use a third diffusion step with boron. In Fig. 8.3, a diffusion profile of a thyristor manufactured in this way is shown. Acceptor profiles with a penetration depth in the range of 20 µm are made preferably with boron. Besides its high solubility it can be masked by SiO_2.

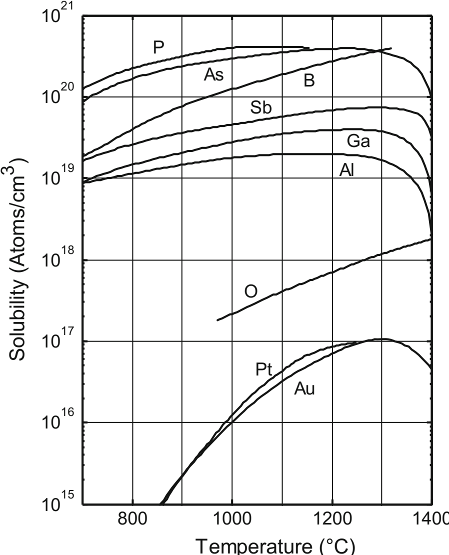


Fig. 4.12 Solubility of some impurity elements in silicon. For P, As, B, Sb the data are from [Bor87], for Ga, Al and O from [Tru60], for Au and Pt from [Bul66] and [Lis75], respectively

Donor profiles for power devices are almost solely generated with phosphorous. P is the only n-doping element in silicon, which has a sufficient solubility and acceptable diffusion constant. Even with phosphorous, deep n-profiles, needed for example for bipolar transistors, require a high diffusion temperature and a long diffusion time. For the 120 µm deep profile of the collector layer of a npn-bipolar transistor (Fig. 7.3), a diffusion time of about 140 h at a temperature of 1260 °C is necessary.

Because of its small diffusivity, arsenic is not used for diffusion in manufacturing of power devices, which contrasts to VLSI technology, where As is widely used just because of its small diffusivity. Because of its high *solubility*, arsenic is used in power devices as dopant of substrate wafers for epitaxy. With As, a very low specific resistivity of less than 5 mΩ cm can be reached while only values near 15 mΩ cm are possible with Sb, the alternative n-type dopant of substrates. Since in many power devices the resistivity of the substrate determines the series resistance, a very low resistivity is important for low conduction losses. If a higher resistivity can be tolerated, substrates are doped with antimony, because unwanted doping of the epitaxial layer by transfer from the substrate (autodoping) can be minimized with Sb.

As mentioned, the given diffusion constants refer to typical diffusion conditions and not too high impurity concentrations. Typically, the diffusion is carried out in an inert ambient. Under oxidizing conditions, the diffusion of B and P is accelerated, and higher diffusion constants are measured [Kar95]. At high doping concentrations the diffusivity increases with N, as is indicated by a flattening of the impurity profile in the surface region, see Fig. 4.13 further below.

4.4.3 High Concentration Effects, Diffusion Mechanisms

At high doping concentrations, there is firstly the **field effect** which leads to an increase of the effective diffusion constant with concentration. Since dopants are ionized, the built-in field associated with a diffused impurity distribution (see Sect. 3.1.2) causes a field component of the impurity particle current. Whereas the field current of *carriers* produced by the built-in field compensates the diffusion current, the opposite charge of the doping ions implicates that the field current of impurities has the same sign as the diffusion current. Analogous to (3.32) the field in a donor region is $E = -kT/q \cdot d\,\mathrm{In}\,n/dx$. Using Eq. (2.20) and the Einstein relationship (2.44) the field current of impurities is obtained as

$$J_E = -D_0 \cdot N \cdot \frac{d\,\ln n}{dx} = -D_0 \cdot N \frac{dN/dx}{\sqrt{N^2 + 4n_i^2}} \tag{4.21}$$

where D_0 is the low concentration diffusion constant. Because of the proportionality to *dN/dx*, the field current is taken together with the diffusion current expressing the total impurity current in the form (4.2), $J = -D \cdot dN/dx$, with the effective diffusion constant

$$D = \left(1 + \frac{N/2n_i}{\sqrt{1 + (N/2n_i)^2}}\right) \cdot D_0 \tag{4.22}$$

With increasing N, D increases around the intrinsic concentration at the diffusion temperature by a factor 2. Especially in the region $0.2 \le N/n_i \le 4.1$, where D varies from 1.1 to 1.9 times D_0, the diffusion equation in the general form (4.4) should be used. From Eqs. (2.6), (2.8), (2.9) together with (2.25), (2.27), the intrinsic concentration at 1200 °C, for example, is obtained to be 4.3×10^{19} cm^{-3} for a doping concentration $N = 1 \times 10^{19}$ cm^{-3}.

However the *observed* increase of D with N is for the dopants in silicon stronger than explainable by the field effect. The cause of this lies in the **diffusion mechanism** on atomic scale. The doping impurities are of the substitutional type, they substitute silicon atoms on lattice sites. Such impurities can diffuse only with the help of crystal defects such as vacancies and interstitials, since a movement directly by exchange with adjacent silicon atoms is energetically very improbable. There are two mechanisms which are made responsible for the observed diffusion constants: the vacancy-assisted and the interstitialcy mechanism using silicon interstitials. Whereas previously the vacancy assistance was assumed to determine the diffusion of dopants in silicon, there is now evidence that this mechanism is used only by Sb and partly As, the slowest diffusing doping atoms, which are larger in size than Si [Ura99]. In the vacancy-mediated diffusion the impurity jumps step by step from one substitutional site to a neighboring one, when there is a vacancy. After each step the probability is high that the impurity jumps back to the vacancy located now

at the old place of the impurity. Since also the vacancy concentration is rather small in silicon [Kar95], the vacancy-assisted diffusion is slow and requires a high temperature.

The elements with higher diffusion constants, P, B, Ga and Al, diffuse by the interstitialcy mechanism [Ura99]. This diffusion mechanism of substitutional impurities is initiated by an interstitial silicon atom (self-interstitial) displacing the impurity from its lattice site and occupying the place itself (kick-out mechanism). The impurity migrates then interstitially to a vacancy or kicks out a silicon atom to take its final or intermediate substitutional site. Although the fraction of interstitial impurities is small, the much faster interstitial diffusion enables this mechanism to produce the higher diffusion constants.

The increase of the diffusivity with N beyond the field effect now is assumed to be due to different charge states of the involved defects, for example as I^-, I^0, I^+ for self-interstitials. The relative concentrations of these states depend on the Fermi level: According to Eq. (2.2) the ratio c^-/c^0 of negative to neutral point defects densities is proportional to $\exp(E_F/kT)$ in thermal equilibrium. Since with increasing impurity density above the intrinsic concentration the Fermi level moves towards the respective band edge (see Fig. 2.10), this results in an increase of the concentration of negatively charged point defects in the case of donor doping and of positive point defects in the case of acceptors, each in relation to the density c^0. Assuming c^0 to be independent of x, the total concentration of involved defects and hence the diffusivity itself increases with doping density [Tsa83]. The assumption of homogeneity of concentration c^0 is justified by the much faster diffusion of vacancies and self-interstitials compared with doping atoms. This explains also the observed constancy of the diffusivity below n_i. Impurities with a *larger* diffusion constant than the involved crystal defects show completely different diffusion profiles, as will be seen for gold and platinum in Sect. 4.9.

How the diffusion profile is changed by the increase of D with N is shown for an example in Fig. 4.13. As reported for phosphorous, the diffusivity is assumed to be proportional to N^2 at high concentrations [Fai81] and on the whole can be expressed as

$$D = D_0 \cdot \left(1 + (N/N_{tr})^2\right)$$

The transition concentration $N_{tr} = 3 \times 10^{19}$ cm^{-3} is chosen to reflect the reported increase of D by an order of magnitude in the range up to $N = 1 \times 10^{20}$ cm^{-3} [Fai81, Sze02]. The value equals approximately the intrinsic concentration at 1100 °C, if bandgap narrowing at $N = N_{tr}$ is taken into account using (2.25), (2.27). Due to the strongly enhanced diffusivity the slope of the profile (solid curve) is reduced in the surface region compared with the *erfc*-distribution for $D = D_0$ (dashed curve). The penetration depth is enhanced. The profile depends only on the single variable $x/\sqrt{t}$, as is the case always for constant surface concentration if the diffusivity depends only on N. Using this variable the one-dimensional form of the

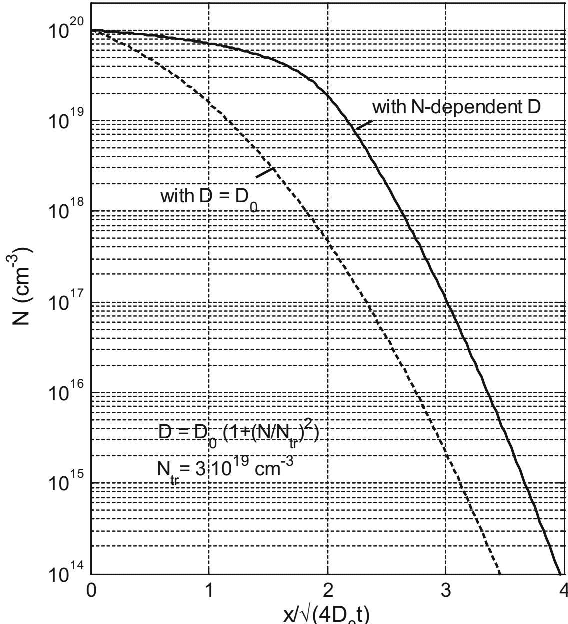


Fig. 4.13 Diffusion profile for a diffusivity increasing quadratically with doping concentration (solid line) compared with the *erfc*(x/L)-distribution for $D = D_0$ (dashed). The $D(N)$ data refer to phosphorous, see the text

diffusion Eq. (4.4) transforms into an ordinary differential equation [Cra56], which is used to calculate the diffusion profile.

In contrast to silicon, the diffusion process cannot be used to introduce dopants in SiC, because the diffusion constants in SiC are too small. Apart from epitaxial processes, which are less versatile, the only possible method to produce required device structures is ion implantation.

4.5 Ion Implantation

In the ion implantation process atoms of the dopants are ionized and accelerated in an electric field to form a beam of mono-energetic ions. The focused beam is scanned vertically and horizontally and the particles are ejected into the wafer homogenously distributed across the wafer surface. The ions will slow down and stop in the semiconductor. The kinetic energy of the impacting ions is lost to the wafer by two types of collisions; with the cores of the lattice atom – elastic nuclear collisions – and by retardation in the electron shells of the lattice atoms – electronic deceleration. The nuclear collisions will also scatter the ions randomly, giving rise to a distribution of implanted ions where the mean penetration depth is mainly determined by the initial energy of the ions.

The dose of the implanted ions and thereby the quantity of the dopants can be controlled very exactly and it is also possible to mask different areas of the wafer to prevent ions from reaching the semiconductor at these locations. Power semiconductors with challenging technology are often doped using ion implantation. Typically, all semiconductor devices with MOS-structures at the surface, as power MOSFET and IGBT, use ion implantation to form the p- and n-type regions.

The generated profile of the dopants can be described in a simplified way with a Gauss function

$$N(x) = \frac{S}{\sqrt{2\pi} \cdot \Delta R_{pr}} \cdot e^{-\frac{(x - R_{pr})^2}{2 \cdot \Delta R_{pr}^2}} \tag{4.23}$$

where R_{pr}, the projected range, corresponds to the mean depth of the implanted ions, i.e. the peak of the profile in this simplified description. ΔR_{pr}, the projected range straggling, is the statistical variation around this mean value, similar to the standard deviation. The integral amount of doping atoms S corresponds to the implanted dose. A doping profile calculated with Eq. (4.23) is drawn in Fig. 4.14.

The projected range depends primarily on the implantation energy, this relation is shown for boron in Fig. 4.15. Ions that are heavier than boron will have a smaller R_{pr} for the same energy.

With increasing energy, the peak height of the doping profile will decrease due to the increasing straggling. This is shown for boron in Fig. 4.16, where a linear scale is used for the concentration. Most often a log-scale is used, since the concentration varies over many orders of magnitude. Tables for the projected range and projected range straggling for the different ions are found in the literature, e.g. in [Rys86]. The ion distributions can also be simulated using for instance the SRIM software, which is a widely used program available on www.SRIM.org [Zie06].

The description discussed up to now has assumed a solid target with unorganized distribution of atoms, i.e. an amorphous target. But in a single crystalline

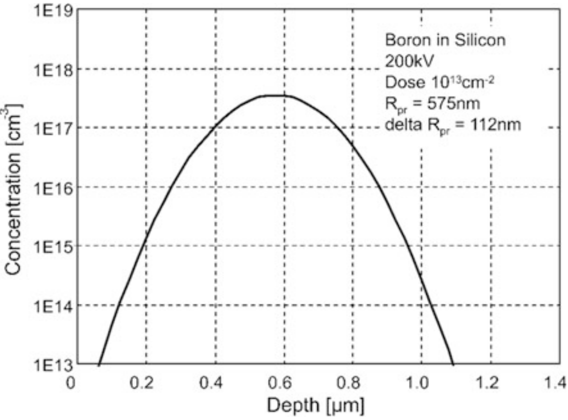


Fig. 4.14 Simplified doping profile of an ion implantation of boron in silicon

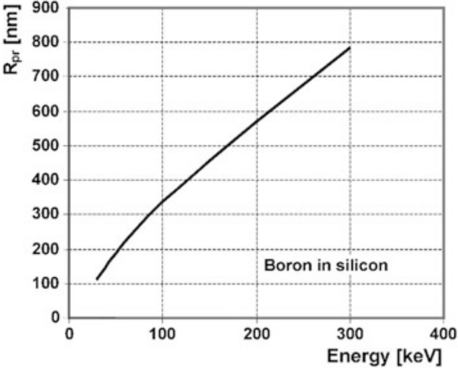


Fig. 4.15 Dependency of projected range R_{pr} on energy for implantation of boron in silicon

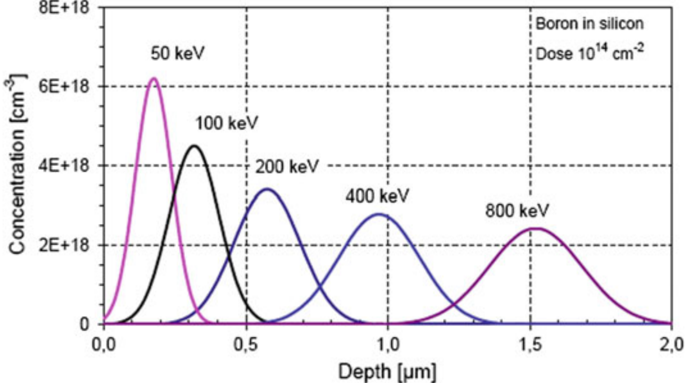


Fig. 4.16 Calculated doping profiles according (4.23) after implantation of boron in silicon with different energies. R_{pr} and ΔR_{pr} are calculated with SRIM [Zie06]

target the atoms are regularly ordered. The elementary cell of the silicon lattice has been already shown in Fig. 2.1. In Fig. 4.17 a drawing of the arrangement of atoms at a view along the [110] direction into a silicon crystal is shown.

If the direction of the penetrating ion beam is aligned with a major crystal orientation, it can penetrate deeper. In the so-called channels, nuclear collisions are very rare and many of the ions are not deflected from their original path. The retardation occurs only due to inelastic impacts with the electron shells of the lattice atoms, i.e. due to the electronic deceleration. The penetration depth in the channels can therefore be 10-times higher than the projected range in an amorphous solid object.

Normally one tries to avoid the channeling, and to counteract this effect the wafer is tilted, so that the implantation occurs in a direction deviating from the major crystal axis. In Fig. 4.18 the dependence of the channeling effect on the tilt angle is shown. Even at a tilt angle of 8° a partial channeling is still visible.

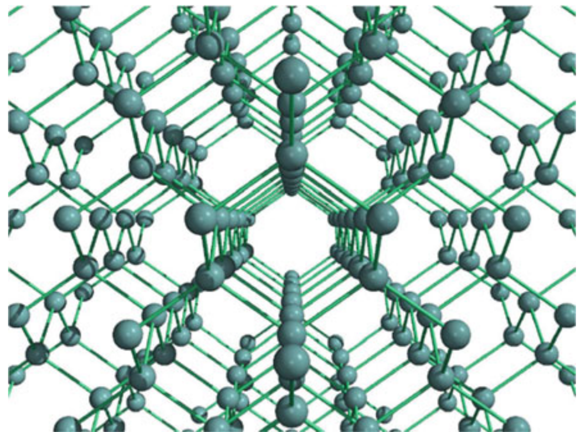


Fig. 4.17 Atomic structure of silicon: viewed in the [110] direction through a silicon crystal of 3 × 3 × 3 unit cells. Figure from [Pic04] © 2004 Springer

A further increase of the tilt angle would cause channeling into other channels. Usually semiconductor wafers are tilted by 7° during ion implantation.

Another way to reduce the channeling is by covering the crystal with an amorphous layer, for instance SiO_2. In this amorphous layer the ions are scattered and the ions will no longer travel in parallel paths. Oxides with a thickness of just 10– 20 nm reduce effectively the channeling effect.

An increased target temperature results in increased amplitude of the lattice oscillations and this will also increase the number of ions that are scattered at the surface. Therefore, the channeling effect is reduced with increased target temperature.

However, none of these countermeasures is sufficient to avoid the channeling effect completely [Rys86]. The best way to reduce the channeling is to use a previously amorphized target. If the silicon wafer is first implanted with Si ions at a high dose, an amorphous Si layer of sufficient thickness is created at the surface, which will effectively prevent the channeling.

When dopants are introduced by ion implantation, one serious complication is the generation of lattice defects by the elastic collisions. These defects consist of atoms knocked out of their lattice positions into interstitial sites, with vacant lattice sites left behind. Combinations of these defects are also known, for instance the Si divacancy, formed by two neighboring vacancies. The maximum amount of ion beam induced lattice defects is located just before the ion comes to rest, e.g. at an implantation of boron in Si at $0.8 \cdot R_{pr}$. The profile of lattice defect is reaching up to the semiconductor surface. The distribution of lattice defects in comparison to the distribution of implanted atoms is shown in Fig. 4.19 on example of an implantation of heavy As ions into silicon.

If the implantation dose of is sufficiently high, an amorphous layer can be generated due to the large number of lattice defects. For every ion there is a critical dose for silicon amorphization, which depends on temperature. The higher the mass of the ion, the lower is the critical dose. However, there is a balance between defect build-up by the incoming ions and the recombination of Si interstitials and

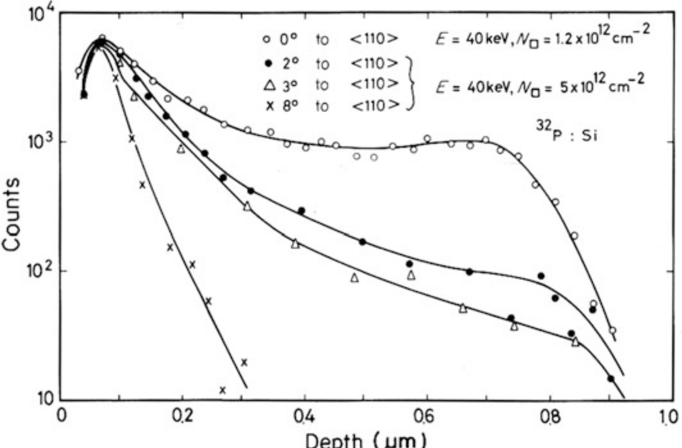


Fig. 4.18 Dependency of the channeling effect on the tilt angle between the ion beam direction and the wafer surface normal for a phosphorus implantation in silicon [Dea68]. Copyright 1968 National Research Council of Canada

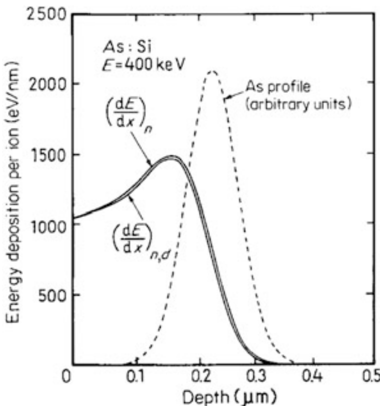


Fig. 4.19 Energy deposition for an implantation of arsenic in silicon in comparison to the distribution of implanted atoms. The profile of lattice defects follows the profile of energy deposition. Figure taken from [Rys86] Original copyright 1971 Phys. Society Japan JPSJ Vol. 31, pp. 1695–1711

vacancies, and this balance depends very much on the target temperature and the flux of incoming ions. For boron implantation at room temperature for instance, no amorphous layer can be generated.

The implantation process is always followed by a thermal annealing process and there are two reasons for it:

– The lattice must be restored to its crystalline form and remaining point defects should recombine.
– The implanted donor- or acceptor ions must be positioned at substitutional sites where they become electrically active, i.e. contribute with an electron to the conduction band or a hole to the valence band.

The annealing of lattice defects starts already at room temperature, but defect complexes of higher order anneal only at a higher temperature. Mostly used in Si processing is a fast annealing process – rapid thermal annealing, RTA, to keep the effect of diffusion of the implanted ions small during annealing. In an RTA process the wafer is heated in a short time to a high temperature by using very intensive light radiators. The temperature is typically above 1000 °C and the annealing time is only some seconds or half a minute. A fast cooling step follows. With this process an effective electrical activation of the doping atoms and an effective annealing of lattice defects is possible without a significant increase of the penetration depth of the doping profile. Even amorphous layers can be restored to monocrystalline layers. Careful optimization of the implantation- and annealing processes is necessary to achieve the desired device structures in the field of micro- and nanoelectronics. For power devices, higher penetration depths than can be reached by conventional implanters are often necessary. In these cases, a diffusion step follows the ion implantation step. However, the process of ion implantation is frequently used because of the possibility of an exact control of the dose and thereby a very exact adjustment of the resulting profile. Modern implanters can produce doping profiles with a deviation of less than 1% across 8-inch wafers.

4.6 Oxidation and Masking

One of the great advantages of silicon compared with other semiconductors is the ability to form an oxide of high quality. Silicon dioxide, SiO_2, is widely used for thin insulating layers (e.g. gate insulation in MOSFETs), and for protecting or masking certain areas of a wafer during processing. SiO_2 features a disordered amorphous structure. For the oxidation of silicon two processes are used.

Dry oxidation:

$$\mathrm{Si} + \mathrm{O_2} \rightarrow \mathrm{SiO_2}$$

This process gives a low growth rate of the oxide, but the quality is very good. It is used for fabrication of thin oxide layers, which are needed as scattering layer for ion implantation to prevent channeling, and also for the creation of gate oxide in field controlled devices, so-called metal-oxide-semiconductor (MOS) structures.

Wet oxidation:

$$\mathrm{Si} + 2\mathrm{H_2O} \rightarrow \mathrm{SiO_2} + 2\mathrm{H_2}$$

The growth rate of the oxide is higher with this process. It is used to fabricate oxides for masking purposes and oxides as passivation layers. After the wet oxidation step, a dry oxidation step to improve the surface quality of the oxide layer follows in many applications.

The thickness of the oxide layer, d_{ox}, can be calculated with [Ben99]

$$d_{ox} = d_0 + A \cdot t \quad \text{for thin oxide}$$
$$d_{ox} = B \cdot \sqrt{t} \quad \text{for thick oxide.}$$

The constants A and B are temperature dependent. For instance, an oxide layer of the thickness of 1.2 μm for masking purpose is created with a wet oxidation at 1120 °C and a time duration t = 3 h. The process is finished with a dry oxidation in the same reactor for 1 h at the same temperature, which improves the oxide quality.

The diffusion constants of the dopants B, P, As and Sb in SiO_2 are several decades lower than their diffusion constants in crystalline Si. This is used in the manufacturing process. The oxide layer is structured with a photolithographic process. The photoresist is exposed to ultraviolet light through a photomask with the intended pattern. After developing and hard baking of the photoresist, the SiO_2 layer is etched in buffered hydrofluoric acid NH_4F/HF, and finally the photoresist is removed ("stripped"). After removal of the photoresist the wafer has a pattern of SiO_2 layers as shown in Fig. 4.20a.

After cleaning, this pattern can now be used as a mask for a diffusion process. A phosphorus diffusion is typically executed from the vapour phase in a diffusion furnace. With the parameters temperature and time the penetration depth x_j is adjusted. However, the time can only be so long as the phosphorus does not penetrate the masking oxide layer with the thickness d_{ox}. With an oxide thickness of 1.2 μm a penetration depth x_j in the range of 10 μm can be achieved. During diffusion a layer of phosphor silicate glass is growing at the surface of the silicon and of the oxide, it acts as source of diffusion atoms. This layer is subsequently removed.

During the vertical diffusion process a simultaneous lateral diffusion under the oxide takes place. An exact analytical solution was given in Sect. 4.4, see Eqs. (4.15)–(4.18) and Fig. 4.9.

In the mathematical treatment, a loss of doping by diffusion out of the surface during the drive-in diffusion has been neglected. This out-diffusion is low for boron, however significant for Al and Ga. Additionally, during drive-in diffusions often an

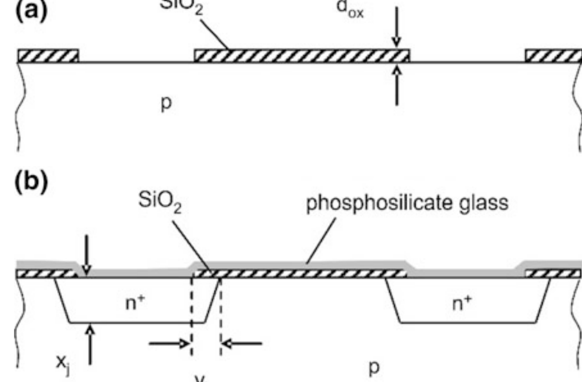


Fig. 4.20 Masking, example of fabrication of n^+-structures in a p-layer. **a** p-layer, SiO_2 structured with a photolithographic process. **b** Phosphorus diffusion with a penetration depth x_j

oxide layer grows at the surface and consumes some of the semiconductor. More detailed calculations of this process can be done with process simulation tools.

Masking with SiO_2 is not possible for the dopants Ga and Al, since for these impurities the diffusion constants in SiO_2 are too high.

Ion implantation is also often masked with SiO_2. The penetration depth of the ions into the oxide layer is in the same range as in Si because of the similar density. Therefore, the thickness of the oxide must be chosen accordingly. Other coating layers are also used for laterally masking during ion implantation, e.g. Si_3N_4 and even photoresist layers can be used, as long as the temperature during ion implantation is kept low.

4.7 Edge Terminations

The one-dimensional investigation of the blocking behavior in Chap. 3 is valid only as long as the semiconductor body is unlimited in lateral direction. In reality, however, the device structures have a finite dimension and an edge termination must be applied to lower the electric field at the edges. Edge termination structures may be divided in two main groups:

1. Edge structures with *beveled termination structures*. By beveling a defined angle can be adjusted between a lateral pn-junction and the surface, and thereby the edge is relieved from high electric fields. An overview is given in [Ger79].
2. Edge structures with a planar semiconductor surface are denoted as *planar termination structures*. An overview is given in [Fal94].

Beveled termination structures

The beveled edge contour is produced by mechanical grinding. The angle α is defined in relation to the junction from the high doped side to the low doped side. A beveled edge with negative bevel angle is shown in Fig. 4.21. The effect can be explained in a simplified way as follows. The equipotential lines of the space charge must intersect with the surface orthogonally, if there are no surface charges. This forces the space charge region to widen at the edge, and thereby the electric field strength is lowered at the surface.

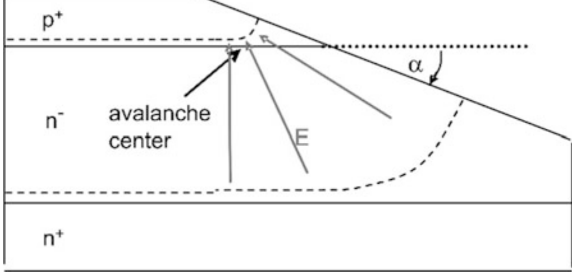


Fig. 4.21 Edge termination with negative bevel angle

In a structure with negative bevel angle, a compression of the field lines occurs close to the junction on the p-side near the beveled surface. Therefore, the electric field is increased at this location. To counteract this, an edge structure with negative bevel angle is always fabricated with a very shallow angle, usually between 2° und 4°. For this case, approximately 90% of the volume blocking capability can be achieved. The avalanche breakdown always starts at the edge close below the semiconductor surface at the location which is marked in Fig. 4.21 as avalanche center.

With a beveled edge termination using a positive bevel angle, as shown in Fig. 4.22, the distance between the equipotential lines is also increased at the surface. Especially close to the pn-junction, where the electric field is high, the field lines are widened at the edge. Therefore no avalanche center occurs and 100% of the volume breakdown voltage can be achieved with this termination structure. The angle α can be chosen in a wide range between 30° und 80°.

The etched structure shown in Fig. 4.23 is also a junction termination structure with positive bevel angle. The semiconductor wafer is etched starting from the n^+-side. Also for this structure, the breakdown will take place in the volume and the breakdown voltage is not lowered compared to the one-dimensional calculation.

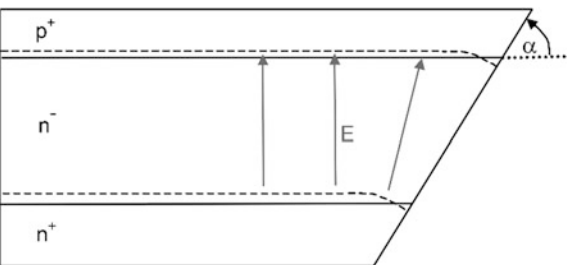


Fig. 4.22 Edge termination with positive bevel angle

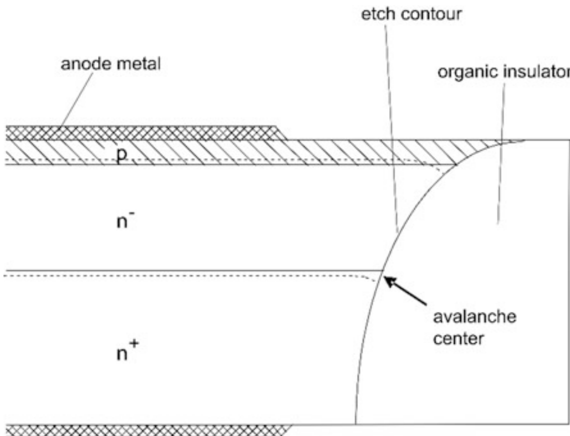


Fig. 4.23 Etched edge termination structure with positive bevel angle at the pn-junction

However, if the space charge penetrates the n^+-layer, an avalanche center may occur at the nn^+-junction at the location marked in Fig. 4.23.

If an avalanche center at the nn^+-junction is avoided, the structure according Fig. 4.23 proves to be very insensitive against surface charges. For the long time stability in this case a passivation layer of silicone gel is sufficient. The sharp corner at the anode side is mechanically very susceptible and such devices can easily be damaged. Therefore, this edge structure is not suited for modern devices with shallow penetration depth of the p-layer.

Planar junction termination structures

Planar structures are mechanically much more insensitive. The structure with floating potential rings, as shown in Fig. 4.24, can be fabricated in a single mask step together with the p-type anode layer. The potential rings lead to a widening of the space charge at the semiconductor top surface. The potential ring structure was first suggested by Kao and Wolley [Kao67].

The maxima of the electric field are marked in Fig. 4.24, they can be lowered by selecting the optimal distance between the potential rings. With numerical simulation using the Poisson equation in a two dimensional grid, the optimal arrangement of potential rings can be calculated, as shown by Brieger and Gerlach [Bri83]. Nevertheless field maxima cannot be completely avoided and the avalanche breakdown will occur in the region of the junction termination. Around 85–95% of the volume breakdown voltage can be achieved. A big advantage of this structure is that no additional photolithographic step is necessary in production; it accrues simultaneously with the fabrication of the p-layer, which is used as anode layer for a diode or p-base for a transistor. Therefore, this structure is the most frequently used edge termination structure. A disadvantage is the large space requirement.

With a very lowly doped p^--zone, the so-called Junction Termination Extension (JTE) structure, it is also possible to reach the volume breakdown voltage using planar structures. The Variation of Lateral Doping (VLD) structure shown in Fig. 4.25 is one of the possible varieties of the JTE structure, it was first suggested by Stengl und Gösele [Ste85]. A p^--zone, in which the doping is decreasing outwards to the device edge, is connected to the p anode layer.

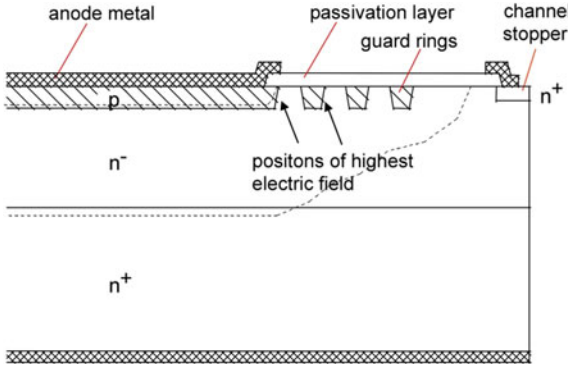
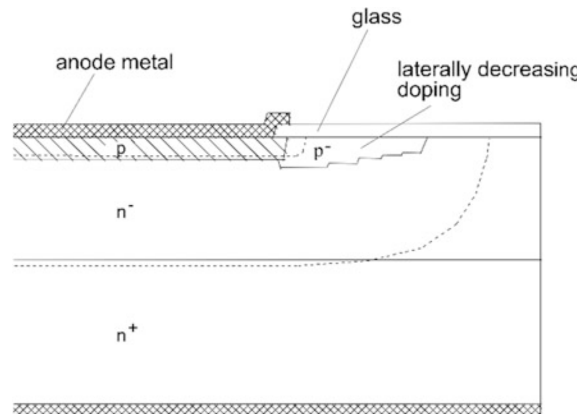


Fig. 4.24 Planar junction termination with floating potential rings



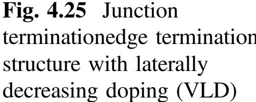


Fig. 4.25 Junction terminationedge termination structure with laterally decreasing doping (VLD)

The structure is manufactured in such a way that the mask for the p$^-$-layer has stripe shaped openings, with a pitch in a form that the share of open area decreases from the anode region towards the edge. The deposition of the p-dopant is executed in these stripes, and during the following drive-in diffusion the p$^-$-zones are joined by lateral diffusion. This process results in a profile with decreasing doping concentration and decreasing depth towards the edge as shown in Fig. 4.25. With an optimal design the breakdown takes place in the volume. Structures fabricated with implantation of Al reach 100% of the volume blocking voltage [Scu89]. Because of the lower solubility of Al, low-doped regions are easier to realize with Al.

Compared to the structure with floating potential rings, the VLD structure features a smaller space requirement and it is also much more insensitive against surface charge states [Scu89]. Because of the small tolerance window of the doping in the edge region, an ion implantation is necessary for controlling the deposited amount of B or Al. Regarding other parameters, e.g. the penetration depth, the VLD structure is less sensitive.

With field plate structures, the metallization of the p-layer is extended above the insulating passivation layer at the edge of the device. Figure 4.26a shows the effect of a one-step field plate. Also with this structure the space charge is elongated at the edge. A one-step field plate is rarely sufficient to reach blocking voltages close to the volume breakdown voltage. For the fabrication of field-controlled devices, several insulating layers are necessary in the cell structure, and using a combination of these layers several steps at the edge can be realized as shown in Fig. 4.26b. The determination of the position of the specific steps is done by numeric simulation solving the Poisson equation, similar to the calculation of field rings, mentioned earlier. Field plate structures are often used in MOSFETs and IGBTs.

It is also possible to combine guard rings and field plates in order to spare some potential rings and thus to reduce the area requirement.

For economic reasons, it is necessary to keep the edge termination area as small as possible, since this area cannot be used for current conduction. On the other hand, additional photolithographic steps should also be avoided. In several ways the

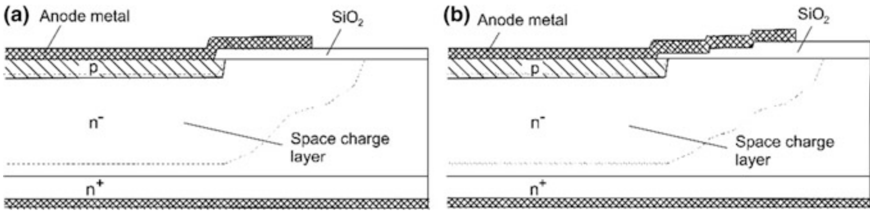


Fig. 4.26 Field plate junction termination. **a** One-step field plate. **b** Multi step field plate

edge is the weak point of a power semiconductor device and the development of protecting junction terminations is one of the most important tasks in the development of stable and rugged power devices.

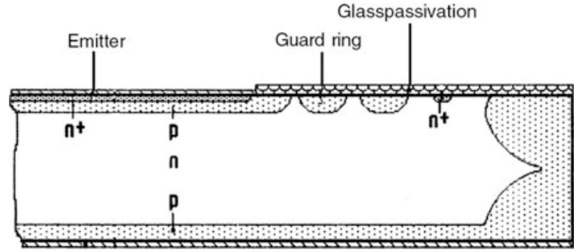
Junction termination for bidirectional blocking devices

Bidirectional blocking devices like the thyristor have two blocking pn-junctions, where the low doped layer n^{-}-layer is commonly used for both directions. Thyristors are fabricated for example with two structures of negative bevel angle like in Fig. 4.21, or with one structure of negative and one with a positive bevel angle as shown Fig. 4.22. Also other structures like etched mesa grooves from both sides are applied.

An attractive solution for a bidirectional blocking device is the structure with deep edge diffusion as shown in Fig. 4.27. The deep p-diffusions on the right side are carried out from both sides as one of the first process steps. In a final step the wafer is cut in the regions of the deep p-diffusions. The upper pn-junction has a planar junction with guard-rings and a channel stopper as in Fig. 4.24. The lower pn-junction is connected to the top side of the wafer via the deep p-layers. The deep p-diffusion diffuses additionally to the side and thus a p^{-}-structure as used in JTE-structures is achieved. No further structures are necessary.

The advantage of the edge diffusion structure is that the wafer is compatible with processes of modern semiconductor technology, in which photolithographic processes can usually be applied only to one side of a wafer. Bidirectional blocking IGBTs are using a similar junction termination, see Sect. 10.7.

Fig. 4.27 Thyristor with edge diffusion. Figure from IXYS Semiconductors AG



4.8 Passivation

The surface of a semiconductor is very sensitive to high electric fields. In addition to the edge terminations just described, it is also necessary to treat the surface in some way to obtain a well-defined surface and to terminate the free bonds of the silicon atoms at the surface. This treatment is called surface passivation and involves cleaning processes and subsequent deposition of an insulating material or a material with high resistivity.

For conventional devices with beveled edge an organic passivation layer is often used which consists of silicon rubber or polyimide. For junction terminations like the beveled structure with positive angle (see Figs. 4.22 and 4.23) the passivation layer is not critical, since no field peaks occur at the interface.

In junction terminations such as the planar structure with potential rings, field peaks occur at the surface and the passivation becomes crucial. The resulting blocking capability of the device is very sensitive to charges in the passivation layer and the charge state of the passivating material must be taken into account in the calculation of the breakdown voltage.

Often SiO_2 is used as passivation layer. After diffusion processes, an oxidized semiconductor surface is attained often and no additional process for passivation is necessary. This oxide layer must have a very high purity and this requirement becomes more critical for lower background doping levels N_D, since at low N_D already a very small density of surface charges is sufficient to generate an inversion layer at the surface. Such inversion layers lead to increased leakage currents and long term instability of the device. Instead of SiO_2 also different glass layers are used, which consist of SiO_2 and additional elements.

Also semi-insulating layers are possible to use as a combined passivation and edge termination. By adjusting the electric conductivity of a semi-insulating layer, a continuous decrease of the potential at the surface can be achieved.

One commonly used evaluation method and selection criterion for the quality of a passivation layer is the hot reverse test (see Sect. 11.6), where devices are subjected to the maximal allowed temperature and the maximum allowed voltage, applied as DC voltage, for 1000 h. If there are mobile ions or other mobile charges in the passivation material, they will move, driven by the high electric field, and accumulate at unfavorable locations where an inversion layer may be created. In this case a significant leakage current increase will be observed and the device will be rejected.

Most challenging is the passivation of very high voltage devices between 5 and 10 kV, because of the very low doping which is mandatory for such devices. For these very demanding conditions a passivation layer of amorphous hydrated carbon (a-C:H) is used in some applications [Bar99]. The mechanical and chemical properties of a-C:H features a diamond-like characteristic, although the bandgap is smaller than for diamond, it is only in the range of 1–1.6 eV. One positive property of a-C:H is that mirror charges can be induced in the bandgap, which have the capability to compensate disturbing charges. They can even reduce field peaks at

the surface. Layers of a-C:H are very stable under the condition of hermitically sealed housings.

For SiC devices, ten times higher fields are allowed in the bulk, as compared to silicon. This imposes difficult challenges for SiC passivation layers and new solutions are required both for surface processing and for the insulating layers.

4.9 Recombination Centers

Fundamental device characteristics depend on the charge carrier lifetime. The carrier lifetime in silicon, as well as all other indirect bandgap semiconductors, is adjusted by recombination centers. Recombination centers always decrease the carrier lifetime and therefore increase the voltage drop at forward conduction of a device, which leads to higher on-state losses. On the other hand, recombination centers decrease many parameters important for the switching behavior, for instance the switching time and the reverse recovery charge. This lowers the losses associated with the turn-off. The following part of this chapter deals with the particularities of different recombination centers, for instance their temperature dependence and how they can affect device characteristics. For power devices it is very important to know the differences between different recombination center technologies and how they can be used to optimize the trade-off between on-state and switching losses.

Gold and platinum as recombination centers

Gold is the earliest used recombination center in silicon [Far65]. Platinum is used as a recombination center since middle of 1970s as an alternative to gold [Mil76]. Both recombination centers, their energy levels (Fig. 2.19) and physical characteristics have been discussed in Sect. 2.7.

Gold as well as platinum are diffused into silicon, and both have similar characteristics in their diffusion mechanism. In silicon they can occupy interstitial and substitutional positions, however, the solubility of the heavy metal is higher at the substitutional sites. On the other hand, gold and platinum on interstitial sites are much more mobile and diffuse rapidly. The diffusion from substitutional heavy metals can in fact be neglected, but there is always a relatively small barrier for interstitial to substitutional sites which results in a very fast diffusion. For instance, at a diffusion temperature of 850 °C, a substantial part of the heavy metals can be found at the opposite side of the semiconductor wafer already after a diffusion time of only 10 min. This fast diffusion leads to an U-shaped concentration versus depth, or a bathtub-shaped profile with increased density of heavy metal atoms close towards the wafer surfaces, as shown in Fig. 4.28. This profile also means that the density is increased close to the pn- and nn^+-junctions in devices.

Figure 4.28 does not contain highly doped zones, which are produced in power devices before the gold diffusion. Highly boron-doped p-layers will have more

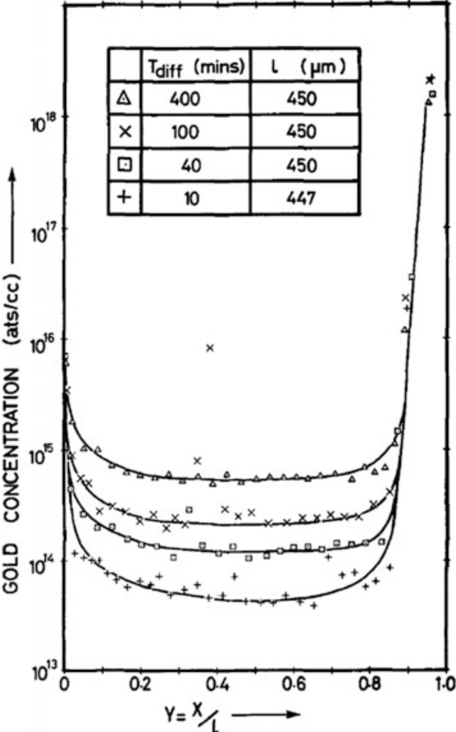


Fig. 4.28 Concentration profiles of gold diffused into FZ silicon slices at 900 °C. Gold deposition before diffusion was executed on the right-hand side. The very high concentrations on this side are not valid. Figure from [Hun73] © 1973 The Electrochemical Society

stress in the lattice and will exhibit an increased gold concentration. Highly phosphorous-doped n^+-layers will collect gold atoms (gettering), a gold diffusion through such layers is not possible.

It is difficult to control the diffusion profile of gold and platinum. Because of the interaction of the diffusion mechanism with other defects in the crystal, it is also difficult to obtain a good reproducibility of the diffused heavy metal profiles. Over many years of application of gold and platinum diffusion to control the charge carrier lifetimes, a wide scattering in the device characteristics even within the same batch, as well as a bad yield had to be accepted. Much due to the problems to control the gold and platinum diffusion, a large difference is fond between the real parameters and the maximal allowed values in data sheets of many fast diodes of older generations.

While the diffusion mechanism is very similar for gold and platinum, the characteristics of gold and platinum diffused devices are very different.

Platinum features a decrease of the lifetime with the injection level, see Fig. 2.21, this holds not for gold. Using gold, it is possible to achieve a relative good trade-off between forward voltage drop and reverse recovery charge extracted at turn-off. In this respect, gold would be a suitable recombination center. However, one of the Au levels is located almost exactly in the middle of the bandgap, see Fig. 2.19. This has the consequence that it acts also very effectively as generation center during the

blocking state and this generation creates a high reverse leakage current at elevated temperature, for further details see Fig. 3.13. The leakage current of devices, where gold has been diffused to control the recombination rate, are at 150 °C a factor of 50 higher than for platinum diffused devices.

The leakage current for different recombination center technologies are compared in Fig. 4.29. For devices with a high density of gold recombination centers as necessary for instance for fast freewheeling diodes and for a rated voltage over 1000 V, at a temperature of 150 °C such high reverse blocking losses occur that they will, in fact, result in thermal instability. Gold must therefore be excluded as recombination center for fast freewheeling diodes in this voltage range. As a thermal limit in Fig. 4.24, it is assumed that the temperature increase by the leakage current is allowed to be at maximum $\Delta T = 15$ K at a DC voltage of 2/3 of the specified blocking voltage, and a thermal resistance $R_{thjc} \times A = 31$ Kmm^{2}W^{-1} is assumed which is, for example, a typical value for a power module.

For gold-diffused devices at a voltage above 1000 V, the maximum allowed junction temperature is limited, typically to 125 °C. For a DC voltage load it must be further reduced and limited to only 100 °C.

In a neutral n-region the gold atoms are mostly negatively charged because the Fermi level lies above the acceptor level of gold. If the density of the gold atoms is in the range of the background doping – this is the case for very fast power diodes – then a compensation effect occurs and the device behaves according to the reduced background doping [Mil76, Nov89]. This affects also the turn-on behavior. The voltage peak V_{FRM}, occurring at turn-on of a device, is a function of the resistance of the low-doped layer – see Chap. 5, Eq. (5.66). For gold diffused devices, V_{FRM} can amount to the multiple of the voltage peak of diodes completely without, or with other recombination centers.

For platinum, the trade-off between forward voltage drop and reverse recovery charge is significantly less convenient than that for gold. On the other hand, for platinum there is no energy level close to the middle of the bandgap, and the small

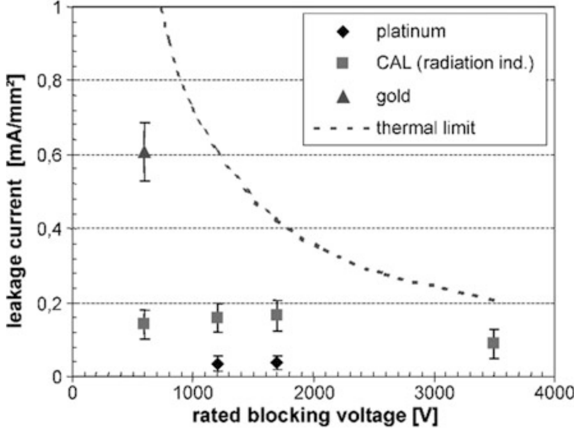


Fig. 4.29 Leakage current of diodes fabricated with different recombination centers at $T = 150$ °C. X-axis: Rated voltage of the respective device. The density of recombination centers is chosen so that requirements for freewheeling diodes for IGBTs with the specified voltage are fulfilled. Figure from [Lut97] © 2007 EPE

leakage current of a platinum diffused device can hardly be distinguished from a device without recombination centers. Therefore higher junction temperatures can be realized with platinum, for instance 150 or 175 °C and in the future probably 200 °C.

For platinum diffused fast diodes the reverse recovery charge is strongly increasing with temperature. The effect of the recombination center is therefore decreasing with increasing temperature. As a consequence platinum diffused diodes, in which the p-emitter is strongly doped and the emitter recombination is low, show a negative temperature dependency of the forward voltage. The temperature dependency of capture coefficients of different recombination centers is summarized in [Sie01].

Radiation induced recombination centers

The characteristics of the most important centers generated by radiation have been discussed in Sect. 2.7. This technology has a much better reproducibility, and it is easy to control the lifetime by adjusting the radiation dose. The process of irradiating silicon power devices with high energy electrons to induce lattice defects, which act as recombination centers, has been in use since the 1970s. Electron and also and γ-radiation creates a homogeneous distribution of defects by knocking out lattice atoms in random collisions with low probability. This technique has many advantages over diffusion technology, but the resulting constant distribution of recombination centers is often a disadvantage for the switching characteristics of devices. More recently, implantation of protons or helium ions has been introduced for charge carrier lifetime control [Lut97, Won87]. These particles also collide with the host atoms and knock them out of their positions, but the collisions occur with large probability towards the end of the ion track. As a result, low mass ion implantation creates a localized region of recombination centers, whose position can be adjusted by the energy of the incoming particles. This process has become widely used for the adjustment of charge carrier lifetime (see also Sect. 5.7, CAL diode). Irradiation processes feature high accuracy and high reproducibility and, furthermore, the irradiation techniques offer the possibility to perform the carrier lifetime control at the end of the fabrication process after the devices have been metalized and passivated.

The vacancies and interstitials that are generated by the collisions diffuse and form stable complexes with each other and with impurity atoms which are present even in high purity silicon, such as carbon, oxygen and phosphorus. The irradiation process is followed by a mild annealing step to eliminate the thermally unstable centers and to ensure a long-term stability of the device characteristics. The annealing process is also needed to reduce the effects of further processing steps, such as soldering etc., that can alter the defect properties.

Devices for packaging in press pack technology are annealed in the range of 220 °C and they will not be exposed to a higher process temperature after the irradiation. Devices which are exposed to a soldering process during packaging and other possible thermal treatments should be annealed at temperature in the range

340–350 °C to ensure that no further defect annealing occurs in subsequent soldering processes, whose temperatures can be quite high.

A lot of work has been done to determine the characteristics of the radiation induced centers [Ble96, Hal96, Scu02, Sue94], a summarizing work of recent results is given in [Sie06] and [Haz07]. The most important centers are shown in Fig. 2.22. Figure 4.30 shows additionally the centers which are still present after annealing at 220 °C and which impact the carrier lifetime or other device characteristics [Sie02, Sie06]. The notations in the base line E(90 K) denote the signal of the center found in Deep Level Transient Spectroscopy (DLTS). This notation is often used, since the determination of the specific atomic structure is difficult and the literature on it is not always consistent.

The OV-center or A-center E(90K) is a vacancy-oxygen complex. This center starts to anneal at temperatures above 350 °C and vanishes almost completely at an annealing temperature of 400 °C [Won85]. As described in Sect. 2.7, the high-injection lifetime τ_{HL} and the resulting forward characteristics as well as switching properties are mainly controlled by the OV-center.

The K-center H(195K) was found as a hole trap at +0.35 eV above the valance band. Different assignments on its origin are fond in the literature, often a COVV-complex involving a carbon impurity, an oxygen atom and two vacancies is assumed. Recent work using the Cathode Luminescence method identified the relevant center as C_iO_i, an association of an interstitial carbon atom and an interstitial oxygen atom [Niw08]. The K-center starts to anneal at a temperature in the range of 370–400 °C and vanishes after annealing at temperatures over 450 °C [Won85]. It has an energy level close to the middle of the bandgap, but the contribution to recombination processes is rather small due to low probability to capture charge carriers, i.e. small capture cross sections. Under high-injection condition, which is given in a forward-biased power diode for example, this center is positively charged. When the diode is turned-off sufficiently fast, the defects remain positively charged for a certain time, typically some 100 ns or 1 µs, due to the relatively low electron-capture rate. The positively charged K-centers will act as donors, and they will increase the effective doping concentration. This reduces temporarily the breakdown voltage at the pn-junction according to Eq. (3.84), resulting in avalanche breakdown at voltages far below the static breakdown

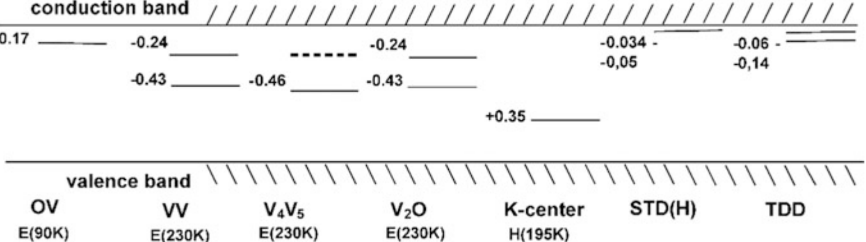


Fig. 4.30 Energy positions in the bandgap for the most prominent irradiation-induced defects in silicon

voltage [Lut98]. These detrimental effects are further described in Sect. 13.3. It is thus important to avoid high densities of K-centers, and the maximum irradiation dose is therefore limited for some applications.

The divacancy, VV E(230K), has as many as three different charge states, that give rise to three separate levels in the bandgap. Most important for charge carrier dynamics is the level at 0.43 eV below the conduction band edge. It acts as a recombination center, but since this energy level is so close to the middle of the bandgap, it also acts as a generation center and will determine the generation lifetime and leakage current. However, its effect as generation center is fortunately weaker than gold.

Divacancies anneal out at temperatures of about 300 °C [Bro82, Won85], but when using proton- or $^4He^{2+}$-irradiation, this level is still found even after annealing temperatures in excess of 350 °C. This can be explained by a transformation of the defect to another defect with nearly the same bandgap position, and the remaining centers have been assigned to the singly- and doubly-charged states of the V_2O defect [Mon02]. Another possible origin of this E(230K) peak is given in [Gul77], which suggests that it arises from a V_4 or V_5 complex. Because of their low concentration, the influence of the V_2O defects on τ_{HL} is rather small, but this center is responsible for the fact that devices implanted with helium ions exhibit a higher leakage current than platinum diffused devices (see Fig. 4.29). The leakage current of helium-implanted devices is about 20% of the leakage current of comparable gold diffused devices and poses no problem for the thermal stability at junction temperatures up to 150 °C.

The E(230K) defect also shows a significant compensation effect on the doping in n-type silicon, where the Fermi-level is above the trap level W_C −0.43 or −0.46 eV. After annealing at a temperature of 350 °C, one can find a decrease of the effective doping due to the charged acceptor-states of E(230K) in the region affected by the helium-implantation. This effect can be used to increase or to adjust the blocking voltage of a device after the main manufacturing steps [Sie06].

Annealing temperatures significantly above 350 °C may result in the formation of so called thermal double-donors TDD. The maximum TDD concentration was found after annealing at $T \approx 450$ °C. The TDDs increase the doping concentration in n-type silicon but compensate the doping in p-type silicon.

While radiation induced centers are used widely, their atomic structure and the charge carrier capture and emission processes are still not fully understood. An overview with a detailed treatment of the characteristics of the specific centers can be found in [Sie06]. Many important details, for instance the temperature dependency of the center characteristics, are still an object of active research.

Radiation enhanced diffusion of Pt and Pd

As described, the diffusion mechanism as well as the built-in mechanism of Au and Pt into the Si lattice interacts with crystal defects. Therefore the final profile can be influenced if there is defined crystal damage at a defined penetration depth as it is created by radiation with particles such as H^+ and He^{++}-ions. After a He^{++}-

irradiation, Pt can be diffused at a temperature much lower than usual for Pt-diffusions into the position of maximal radiation damage caused by the He^{++} implantation [Vob02]. If the temperature is high enough to anneal the radiation-induced defects, they are replaced by a local profile of Pt atoms. Since not all of the radiation induced defects have optimal properties, the final device characteristics can be further improved by the radiation enhanced diffusion. A level close to the middle of the bandgap, which leads to increased leakage current, is avoided if Pt or Pd is used.

The energy levels of Pd are at positions in the bandgap close to that of Pt, so that the electronic structure of a substitutional Pd atom seems to be quite similar to a substitutional Pt atom. A wider usable temperature range was found for the radiation enhanced diffusion of Pd [Vob07]. Depending on the temperature, the formation of an acceptor is observed and a buried p^--layer can be formed at the penetration depth of the primary He-implantation [Vob09]. This p^--layer increases the static breakdown voltage and reduces peaks of the electric fields and dynamic avalanche at fast switching events of diodes. Diodes with increased ruggedness have been presented using this technique [Vob09]. Further work is of interest to understand the details of the created centers.

4.10 Radiation-Induced Doping

After proton irradiation, additional effects arise since the protons (hydrogen ions) may also participate in the defect complexes [Gor74, Ohm72]. The hydrogen-related shallow thermal donors STD(H) are an example for this. These centers are found after annealing of a proton implanted sample above 200 °C [Won85]. Their maximum density is found in the region of the projected range R_{pr} of the protons, a clear indication that the defects involve the implanted hydrogen. The STD(H) is a stable donor with an energy level close to the conduction band. If proton implantation is used to reduce the carrier lifetime, the dose must be limited to keep the density of these hydrogen related centers below the background doping. Otherwise, deep buried n-layers are created. However, this is also used for intentionally doping to adjust the threshold voltage of protection devices [Sie06].

The relationship between proton dose and created donors is shown in Fig. 4.31 taken from [Klu11].

For FZ-silicon, the following equation expresses not too high doses by [Klu11]

$$N_D[\text{cm}^{-3}] = (12.6 \pm 0.8)\text{cm}^{-1} \times \Phi \ [\text{cm}^{-2}] \tag{4.24}$$

At about 10^{16} cm^{-3}, the amount of created donors reaches an upper limit. In [Klu11] it is shown that not only hydrogen but also oxygen is involved in the created center and the maximum density is limited by the oxygen content.

If proton implantation is executed in conjunction with an adjacent subsequent annealing step at temperatures between 300 and 500 °C, “buffers” which lead to a

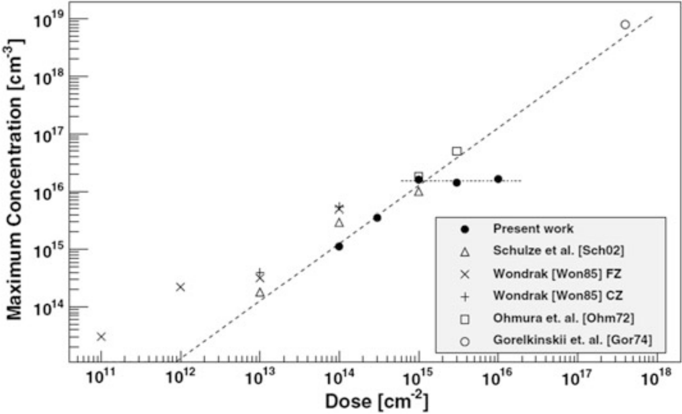
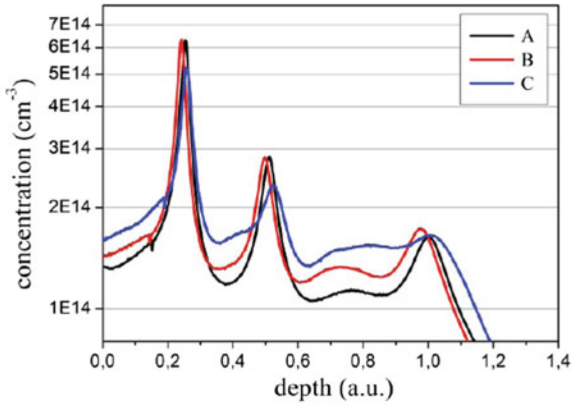


Fig. 4.31 Maximum peak charge carrier concentration for 2 meV protons. The dashed line shows the assumed linear relationship. The upper limit found for high doses is indicated by the dotted line. Fig. from [Klu11]

trapezoidal field in the device (PT-dimensioning, Chap. 5) can be created. This technology is applied for some IGBTs of the latest generations.

In [Scu16] it is shown that, at presence of H, the radiation-induced C_iO_i centers transform into donor-like electrically active C_iO_i-H complexes at annealing above 350 °C. The resulting profile depends on the carbon content, see Fig. 4.32.



Impurity (cm^{-3})	Substrate		
	A	*B*	*C*
oxygen	1.9×10^{17}	2.4×10^{17}	2.1×10^{17}
carbon	5.4×10^{14}	7.5×10^{14}	2.4×10^{15}

Fig. 4.32 Profile of radiation-induced donors for a three-step proton implantation in magnetic CZ silicon, depending on carbon and oxygen content of the substrate. © 2016 IEEE. Reprinted, with permission, from [Scu16]

The profile in Fig. 4.32 was achieved with a three-step irradiation of magnetic CZ-Silicon. The peaks at the maximum penetration of hydrogen are visible, and a slowly increased doping towards the surface, as intended, is created. Immediately after irradiation, hydrogen-associated donors are created only in a narrow region around the projected range R_{pr} in accordance with the distribution of the hydrogen atoms. Lattice defects, however, are present up to the surface. Due to the charged acceptor-states of E(230K) in the region affected by lattice defects, a compensated or even nearly p-type region can occur. During subsequent annealing, a diffusion of hydrogen towards the surface takes place. The intensity of intended donors depends on the carbon content, see Fig. 4.31. Therefore, all process steps have to be developed carefully respecting the initial carbon and oxygen content of the wafer.

Radiation-induced donor doping needs, compared to diffusion processes, only a low temperature at annealing. Therefore, this method is applicable for thin-wafer technology even for a wafer diameter of 30 cm. Based on magnetic Czochralski-wafers, this leads to a very effective production process of IGBTs and freewheeling diodes.

4.11 Some Aspects on Technology of GaN Devices

As reported in the introduction, the semiconductor GaN is with respect to bandgap with 3.4 eV and the critical field for avalanche similar to 4H-SiC, also the mobility of electrons is similar. It can be doped to n- and p-type conductivity, but it is mostly not intentionally doped (see below). Contrary to SiC, large area single crystals of GaN are not available, although significant progress in the growth of single crystal GaN has been made in recent years. GaN power devices as far as developed till now are based on a high quality crystal layer of GaN grown on different substrates. The devices consist actually of a hetero structure with AlGaN, which is an essential constituent deposited on the GaN as a thin film. Because the technology of GaN power devices differs widely from devices in Si and SiC, some aspects of the technology are described in the present section.

GaN crystallizes apart from a metastable cubic zinkblende polytype in the stable hexagonal wurtzite lattice, which is used for devices. The substrates must match with this lattice to some degree, to enable the hetero-epitaxy without creating a huge density of crystal defects not admitted for high-performance devices. Besides the lattice, the coefficient of thermal expansion (CTE) should not much differ from that of GaN in order to limit mechanical stresses caused by cooling down from the temperature of epitaxy. In Fig. 4.33 the CTE and the relevant lattice parameter are plotted for GaN itself and the substrate materials AlN, 4H-SiC, silicon and sapphire (monocrystalline Al_2O_3).

Like GaN, AlN crystallizes in the wurtzite lattice, and also the other substrates have a hexagonal symmetry, except silicon, for which this holds however within the (111) plane, the surface used for the epitaxial growth. As is seen, AlN and 4H-SiC come nearer to GaN than Si and sapphire. Nevertheless, from cost reasons Si is the mainly used substrate for power devices, while optoelectronic devices, e.g. light

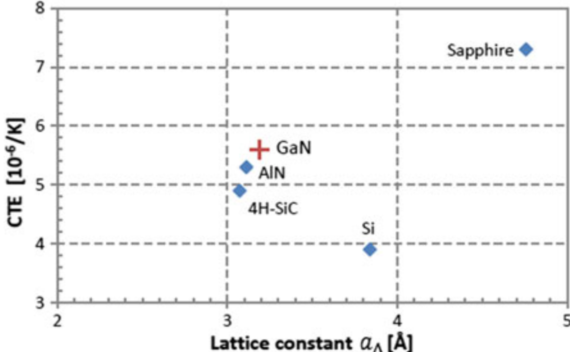


Fig. 4.33 Coefficient of thermal expansion and lattice constant a_Δ for GaN and several substrate materials. a_Δ is the distance between nearest lattice points within the c-plane or within the (111) plane for Si. In the latter case $a_\Delta = a_\square/\sqrt{2}$ where $a_\square$ is the side length of the cubic unit cell. The CTEs are mean values over the range of 20–800 °C and for the hexagonal crystals hold for the c-plane [Rod05, Yim74, LiB86, Swe83]

emitting diodes (LEDs), are primarily produced using sapphire as substrate. Some microwave devices are produced with GaN on SiC because of the better thermal conductivity of SiC. AlN is used as a thin interface layer (see below).

The use of Si as substrate implicates that cooling down after epitaxy generates high thermal strain. The GaN layer tends to constrict itself stronger than the substrate, and beyond a critical point it cracks. To reduce the strain the temperature of epitaxy is chosen as low as possible, typically in the range 740 to 800 °C. It turned out that high quality GaN layers could be grown on silicon only by using an intermediate buffer layer of $Al_xGa_{1-x}N$, in which Ga ions are partly replaced by Al [Krü98]. The layers are deposited from a gaseous composition of Triethylgallium and Trimethylalluminium and pure GaN or AlN (Metal-Organic Chemical Vapor Deposition method, MOCVD). During the deposition of the buffer layer the wafer gets a convex bowing, which is chosen in a manner that it is just compensated by the strain produced during cooling. By these measures of 'stress engineering' the final wafer bow must be reduced below about 50 µm in height, this is required for the following manufacturing steps, especially the photolithographic alignment [Ger15].

A typical graded structure of a GaN-on-Si wafer is shown in Fig. 4.34. On Si, a thin layer of AlN is deposited to improve the nucleation for the following low-temperature epitaxy. Then the buffer layer of $Al_xGa_{1-x}N$ is deposited, where x decreases from 0.75 to 0.25. The bandgap of this material varies between the bandgap of AlN with 6.2 eV and that of GaN as determined by the parameter x. Besides the purpose of strain compensation this layer has the function to isolate the drain of the device from the substrate, which is preferably on source potential. Hence the buffer layer is intentionally not doped. If necessary, the resistivity is enhanced by compensation with deep impurities. Also its thickness has to comply with the requirement of blocking. For a 600 V switching transistor, a buffer

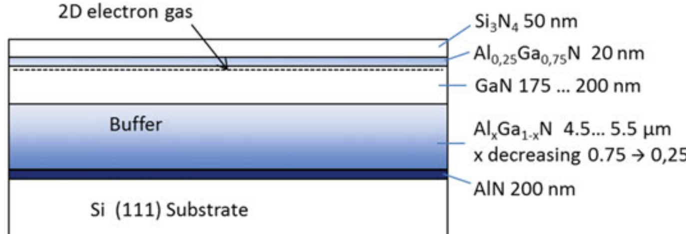


Fig. 4.34 GaN-on-Si wafer as used for lateral high-electron mobility transistors and Schottky-diodes. Drawing according to information from EpiGaN

thickness of minimum 4.5 µm is required. Often a further buffer layer not shown here is used consisting of several thin layers of GaN and AlN, forming a so-called super-lattice buffer [Ued05, Ued17].

The GaN layer following the buffer layer is the active channel layer used for switching. Its thickness is for example 175–200 nm. Thickness and n-doping must be low enough that it can be depleted with a gate voltage. An essential part of the channel is a highly conducting electron layer, termed as two-dimensional electron gas (2DEG), which is induced by the high electric polarization of GaN, as discussed below. The next following layer consisting of $Al_xGa_{1-x}N$ with typically $x = 0.25$ is called barrier layer. Its thickness is in the 20 nm range.

GaN and to a still larger extent AlGaN exhibit a high electric polarization. This is visualized by Fig. 4.35, which shows the wurtzite lattice structure of GaN. The crystal consists of stacked bilayers, one plane of which is closely packed with Ga and the other with N atoms. Nitrogen is an element with high electronegativity: Instead of spending an electron for a covalent bond it has a strong tendency to attract electrons from Ga to complete its outer electron shell. Hence the Ga-N bonds are to an essential degree of ionic or polar character. The result is that the surface of the crystal consisting of Ga atoms (Ga-face, upper surface in Fig. 4.35) has a positive charge and the other surface consisting of N atoms (N-face) a negative, so

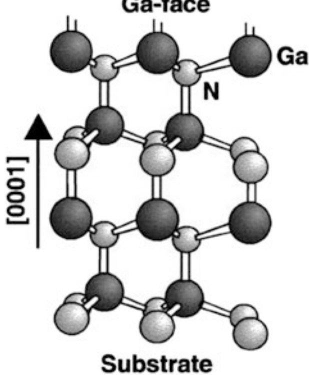


Fig. 4.35 Schematic drawing of the crystal structure of wurtzite GaN, Ga-face. Reproduced from [Amb99], with the permission of AIP Publishing

the crystal is strongly polarized. The polarization is equal to the sheet surface charges and directed by definition from the negative to the positive charge. The polarization of AlN is even nearly 3 times stronger than that of GaN, and this transfers to $Al_xGa_{1-x}N$ in proportion to the share x of Al. In addition to the 'spontaneous' polarization without strain there is a considerable piezoelectric polarization part caused by the strain of the layers resulting from the mismatch of the lattice constants and the coefficient of thermal expansion.

A precondition for the creation of the 2DEG in the GaN is that the needed carriers are generated sufficient quickly in the volume or flow to from the contacts at the sides. In contrast to GaN the barrier and buffer layer of AlGaN can be assumed essentially as isolators without available carriers. If now the GaN layer is grown on the buffer with the Ga-face at the top, also the AlGaN barrier layer will grow on it with the same orientation, i.e. with the N-face at the interface and the AlGa-face at the top. The fixed charge at the interface is made up by the positive charge of the Ga-face of the GaN layer and the higher negative charge density of the N-face of the AlGaN barrier film, the net interface charge is hence negative. In addition to the interface charge also the positive charge of the AlGa-face at the top of the barrier layer has a strong influence. This is illustrated in Fig. 4.36, which shows the surface and interface charges and the resulting field distribution in the barrier and the adjacent part of the channel layer. The charge of the AlGa-face neutralizes the charge of the N-face of the AlGaN-layer. Hence only the positive charge of the Ga-face of GaN remains in effect to attract electrons in the GaN layer. The jump of the field at the surfaces is q/ε times the sheet charge density there (The dielectric constant has practically the same value $\varepsilon = 8.9$ in both layers). The field in the AlGaN film is very high. The field in the GaN is a measure for the number of electrons existing behind a given point, see the next passage. The main result is that the charge of the Ga-face of the GaN layer, which is equal to the net charge at the surface of AlGaN and the interface, determines the sheet charge density of the

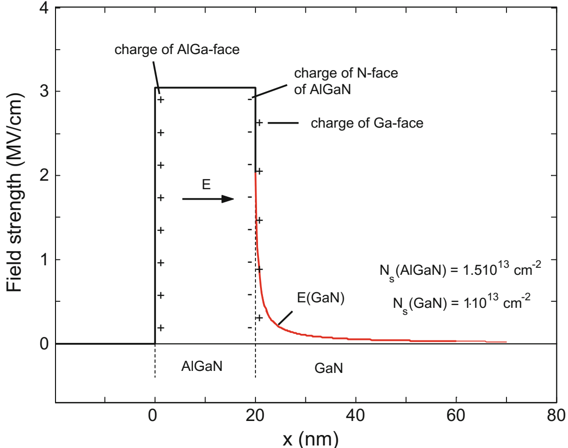


Fig. 4.36 Surface and interface charges and field distribution in the AlGaN barrier layer and the adjacent part of the GaN channel layer. The surface charge densities of AlGaN are assumed to be $N_s = \pm 1.5 \times 10^{13}$, that of the GaN layer 1×10^{13} cm^{-2} (see text)

induced 2DEG. This will hold also, if the field in the AlGaN-layer is not maintained because carriers are still present.

The 2DEG is known from MOSFETs as accumulation layer. Its extension in the depth can be calculated analytically, if the doping concentration is negligible. By double integration of the Poisson equation using Boltzmann statistics one obtains the following dependency (x coordinate perpendicular to the interface):

$$n = \frac{n(0)}{(1 + x/L_D)^2}, \tag{4.25}$$

where L_D is the Debye length at half of the surface concentration $n(0)$:

$$L_D = \sqrt{\frac{2\,\mathrm{kT}\varepsilon}{q^2 n(0)}}. \tag{4.26}$$

At $n(0) = 1 \times 10^{20}$ cm^{-3} Eq. (4.26) yields for GaN with $\varepsilon = 8.9 \cdot \varepsilon_0$: $L_D = 0.504$ nm. The sheet concentration is obtained from (4.25) as $N_s = n(0) \times L_D = 5.04 \times 10^{12}$ cm^{-2}. Although 2/3 of the electrons are contained in the range $x \leq 2L_D = 1.01$ nm, the concentration at $x = 30$ nm is still 2.7×10^{16} cm^{-3}. The extension is comparable to the thickness of the inversion channel of a silicon MOSFET. The example agrees nearly with the case of Fig. 4.36. The field in GaN was calculated there using (4.25) from

$$E = -\frac{\mathrm{kT}}{q}\frac{d\ \ln n}{dx} = \frac{E(0)}{1 + x/L_D}.$$

The polarization of GaN amounts up to 2×10^{13} elementary charges per cm^2, an equal number of mobile electrons is induced in GaN. Compared with a silicon MOS-device this is very high: For an oxide thickness $d_{ox} = 50$ nm, a threshold voltage $V_T = 5$ V and an applied gate voltage $V_G = 15$ V, one obtains from Eqs. (9.1) and (9.2) later in Chap. 9.4 a sheet electron density of 4.3×10^{12} cm^{-2} in the channel.

The mobility of the electrons in the 2DEG is found to be higher than the channel mobility in silicon MOSFETs. Values ranging from 1000 to 2000 cm^2/(Vs) are reported at room temperature for a sheet carrier density of 1×10^{13} cm^{-2} [Amb99]. The spread is explained by a varying surface roughness. The mobility $\mu_{n(Ch)}$ in Si IGBTs or MOSFETs is under typical conditions in the range 300–500 cm^{2}V^{-1}s^{-1}. Electron mobilities in bulk GaN are found to be up to 990 cm^{2}V^{-1}s^{-1} [Qua08]. The carriers in the 2DEG (as well as in a MOSFET channel) are spatially separated from dopants, hence scattering at impurities is less important than in the neutral bulk. For SiC MOSFETs the best values reported are about 70 cm^{2}V^{-1}s^{-1}, compare Chap. 9. Due to both the high sheet concentration and the high electron mobility, the 2DEG in GaN is very effective in current conduction.

If necessary, n-type doping of GaN can be established by silicon as dopant. The activation energy ΔE_G of silicon in GaN is 5–9 meV allowing effective ionization [Qua08]. For p-type doping Mg is mostly used, the ionization energy is reported to be 173 meV [Qua08]. Unintentionally doped GaN is typically of n-type with electron concentration $n = N_D - N_A$ of the order 10^{16} cm^{-3} [Tan10]. The n-type conductivity is attributed to oxygen incorporation [Van04]. Also nitrogen vacancies are shallow donors, whereas Ga vacancies are triple charged acceptors [Van04]. The little controllable doping of GaN is still a problem.

Meanwhile, 8-inch GaN-on-Si wafers can be produced. The GaN-on-Si wafers can be used for device fabrication in a silicon CMOS production line in CMOS foundries, using the standard tool set for creating structures. To avoid cleanroom contamination by free-floating GaN particles, a Si_3N_4 passivation layer which is about 50–100 nm thick is of advantage. It promises a smooth and contamination-free surface [Beh15]. Because of the crystalline mismatch and different CTE, GaN-on-Si wafers are more fragile compared to silicon wafers. A GaN-on-Si wafer uses a thicker substrate as it is the case for Si in CMOS-lines [Beh15]. The functioning of GaNdevices is discussed in Sect. 9.12.

References

[Amb99] Ambacher, O., et al.: Two-dimensional electron gases induced by spontaneous and piezoelectric polarization charges in N- and Ga-face AlGaN/GaN heterostructures. J. Appl. Phys. **85**(6), 3222–3233 (1999)

[Amm92] von Ammon, W.: Neutron transmutation doped silicon – technological and economic aspects. Nucl. Instrum. Methods Phys. Res. **B63**, 95–100 (1992)

[Bar99] Barthelmeß, R., Beuermann, M., Winter, N.: New diodes with pressure contact for hard-switched high power converters. Proceedings of the EPE '99, Lausanne (1999)

[Beh15] Behet, M.: GaN—Promise to Reality: The Next Generation of Power Electronics is Taking Shape. Display + web Publication, http://www.displayplus.net/news/articleView.html?idxno=64606. Download 6 Jan 2017

[Ben99] Benda, V., Govar, J., Grant, D.A.: Power Semiconductor Devices. Wiley, New York (1999)

[Ble96] Bleichner, H., Jonsson, P., Keskitalo, N., Nordlander, E.: Temperature and injection dependence of the Shockley–Read–Hall lifetime in electron irradiated n-type silicon. J. Appl. Phys. **79**, 9142 (1996)

[Boe11] Böving, H., Laska, T., Pugatschow, A., Jakobi, W.: Ultrathin 400V FS IGBT for HEV applications. In: Proceedings International Symposium on Power Semiconductor Devices and ICs ISPSD 2011, pp. 64–67 (2011)

[Bor87] Borisenko, V.E., Yudin, S.G.: Steady-state solubility of substitutional impurities in silicon. Phys. Status Solidi (a) **101**, 123–127 (1987)

[Bri83] Brieger, K.P., Gerlach, W., Pelka, J.: Blocking capability of planar devices with field limiting rings. Sol. State Electron. **26**, 739 (1983)

[Bro82] Brotherton, S.D., Bradley, P.: Defect production and lifetime control in electron and γ-irradiated silicon. J. Appl. Phys. **53**(8), 5720–5732 (1982)

[Bul66] Bullis, W.M.: Properties of gold in silicon. Solid-State Electron. **9**, 143–168 (1966)

[Cho11] Chowdhury, I., Chandrasekhar, M., Klein, P.B., Caldwell, J.D., Tangali, S.: High growth rate 4H-SiC epitaxial growth using dichlorosilane in a hot-wall CVD reactor. J. Cryst. Growth **316**(1), 60–66 (2011)

[Cla11] Clark, P.: IMEC plans 450 mm wafer fab module for 2015. EETimes.com, October 11 (2011)

[Cra56] Crank, J.: The Mathematics of Diffusion, p. 148. Clarendon Press, Oxford (1956)

[Dea68] Dearnaley, G, Freeman, J.H., Gard, G.A., Wilkins, M.A.: Implantation Profiles of ^{32}P Channelled into Silicon Crystals. Can. J. Phys. **46**, 587ff (1968)

[ECP16] Friedrichs, P. et al.: ECPE position paper on next generation power electronics based on wide bandgap devices—challenges and opportunities for Europe http://www.ecpe.org/roadmaps-strategy-papers/strategy-papers/ (2016)

[Fai81] Fair, R.B.: Concentration profiles of diffused dopants. In: Wang, F.F.Y., (ed) Impurity doping processes in silicon, North Holland (1981)

[Fal94] Falck, E.: Untersuchung der Sperrfähigkeit von Halbleiter-Bauelementen mittels numerischer Simulation, Dissertation, Berlin (1994)

[Far65] Farfield, J.M., Gokhale, B.V.: Gold as recombination center in silicon. Solid State Electron. **8**, 685–691 (1965)

[Ful56] Fuller, C.S., Ditzenberger, J.A.: Diffusion of donor and acceptor elements in silicon. J. Appl. Phys. **27**, 544–553 (1956)

[Gal02] Galindo, V., Gerbeth, G., von Ammon, W., Tomzig, E., Virbulis, J.: Crystal growth melt flow control by means of magnetic fields. Energy Convers. Manage. **43**(3), 309–316 (2002)

[Ger79] Gerlach, W.: Thyristoren. Springer, Berlin (1979)

[Ger15] Germain, M., Derluyn, J., Leys, M., Degroote, S.: The material challenge: heteroepitaxial growth of GaN-on-Si, tutorial slides ESSCIC/ESSDERC Conference (2015)

[Gor74] Gorelkinskii, Y.V., Sigie, V.O., Takibaev, Z.S.: EPR of conduction electrons produced in silicon by hydrogen ion implantation. Phys. Status Solidi (a). **22**, K55–57 (1974)

[Gul77] Guldberg, J.: Electron trap annealing in neutron transmutation doped silicon. Appl. Phys. Lett. **31**(9), 578 (1977)

[Haa76] Haas, E.W., Schnoller, M.S.: Phosphorus doping of silicon by means of neutron irradiation with protons. IEEE Trans. Electron Devices. **23**(8), 803–805 (1976)

[Hal96] Hallén, A., Keskitalo, N., Masszi, F., Nágl, V.: Lifetime in proton irradiated silicon. J. Appl. Phys. **79**, 3906 (1996)

[Haz07] Hazdra, P., Komarnitskyy, V.: Local lifetime control in silicon power diode by ion irradiation with protons: introduction and stability of shallow donors. IET J. Circuits Devices Syst. **1**(5), 321–326 (2007)

[Hil15] Hilt, O., Bahat-Treidela, E., Knauer, A., Brunner, F., Zhytnytska, R., Würfl, J.: High-voltage normally OFF GaN power transistors on SiC and Si substrates. MRS Bull. **40**(5), 418–424 (2015)

[Hun73] Huntley, F.A., Willoughby, A.F.W.: The effect of dislocation density on the diffusion of gold in thin silicon slices. J. Electrochem. Soc. **120**(3), 414–422 (1973)

[Jan76] Janus, H.M., Malmros, O.: Application of thermal neutron irradiation with protons for large scale production of homogeneous phosphorous doping of floatzone silicon. IEEE Trans. ED **21**, 797–805 (1976)

[Kao67] Kao, Y.C., Wolley, E.D.: High voltage planar pn-junctions. IEEE Trans El. Dev. **55**, 1409 (1967)

[Kar95] El-Kareh, B.: Fundamentals of Semiconductor Processing Technology. Kluwer Academic Publishers, Boston (1995)

[Ken65] Kennedy, D.P., O'Brien, R.R.: Analysis of the impurity atom distribution near the diffusion mask for a planar p-n junction. IBM J. Res. Dev. **9**, 179–186 (1965)

[Klu11] Klug, J.N., Lutz, J., Meijer, J.B.: n-type doping of silicon by proton implantation. In: Proceedings of the 2011 14th European Conference on Power Electronics and Applications EPE (2011)

[Kra02] Krause, O., Pichler, P., Ryssel, H.: Determination of aluminum diffusion parameters in silicon. J. Appl. Phys. **91**(9) (2002)

[Krü98] Krüger, J., Kim, Y., Subramanya, S., Weber, E.R.: Towards the development of defect-free GaN substrates: defect control in hetero-epitaxially grown GaN by new buffer layer design. Final Report 1997–1998 for MICRO Project 97-202, Berkley. http://www2.lbl.gov/tech-transfer/publications/1461pub.pdf

[Lar51] Lark-Horovitz, K.: Nuclear-bombarded semi-conductors. In: Semiconductor Materials, Proceedings of a Conference at University of Reading. Butterworths, London, 1951, pp. 47–69 (1951)

[Lap08] LaPedus, M.: Industry Agrees on First 450-mm Wafer Standard. EETimes 22 Oct 2008

[Las97] Laska, T., Matschitsch, M., Scholz, W.: Ultra thin-wafer technology for a new 600V-NPT-IGBT. In: Proceedings IEEE International Symposium on Power Semiconductor Devices and IC's, ISPSD '97, pp. 361–364 (1997)

[LiB86] Li, Z., Bradt, R.C.: Thermal expansion of the hexagonal (4H) polytype of SiC. J. Appl. Phys. **60**(2) (1986)

[Lis75] Lisiak, K.P., Milnes, A.G.: Energy levels and concentrations for platinum in silicon. Solid-State Electron. **18**, 533–540 (1975)

[Lut97] Lutz, J.: Axial recombination centre technology for freewheeling diodes. In: Proceedings of the 7th EPE, Trondheim, 1.502 (1997)

[Lut98] Lutz, J., Südkamp, W., Gerlach, W.: IMPATT oscillations in fast recovery diodes due to temporarily charged radiation induced deep levels. Solid-State Electron. **42**(6), 931–938 (1998)

[Mil76] Miller, M.D.: Differences between platinum- and gold-doped silicon power devices. IEEE Trans. Electron. Dev. **23**(12) (1976)

[Mon02] Monakhov, E.V., Avset, B.S., Hallen, A., Svensson, B.G.: Formation of a double acceptor center during divacancy annealing in low-doped high-purity oxygenated Si. Phys. Rev. B **65**, 233207 (2002)

[Mue93] von Münch, W.: Einführung in die Halbleitertechnologie. B.G. Teubner, Stuttgart, Germany (1993)

[Niw08] Niwa, F., Misumi, T., Yamazaki, S., Sugiyama, T., Kanata, T., Nishiwaki, K.: A study of correlation between CiOi defects and dynamic avalanche phenomenon of PiN diode using he ion irradiation. In: Proceedings of the PESC, Rhodos (2008)

[Nov89] Novak, W.D., Schlangenotto, H., Füllmann, M.: Improved Switching Behaviour of Fast Power Diodes. PCIM Europe (1989)

[Ohm72] Ohmura, Y., Zohta, Y., Kanazawa, M.: Electrical properties of n-type Si layers doped with proton bombardment induced shallow donors. Solid State Commun. **11**(1), 263–266 (1972)

[Pic04] Pichler, P.: Intrinsic Point Defects, Impurities, and Their Diffusion in Silicon. Springer Wien, New York (2004)

[Qua08] Quay, R.: Gallium Nitride Electronics. Springer, Berlin Heidelberg (2008)

[Rod05] Roder, C., Einfeldt, S., Figge, S., Hommel, D.: Temperature dependence of thermal expansion of GaN. Phys. Rev. B **72**, 085218 (2005)

[Rys86] Ryssel, H., Ruge, I.: Ion Implantation. Wiley, New York (1986)

[Scl74] Schnöller, M.S.: Breakdown behaviour of rectifiers and thyristors made from striation-free silicon. IEEE Trans. Electron Devices ED **21**, 313–314 (1974)

[Sco82] Schlangenotto, H., Silber, D., Zeyfang, R.: Halbleiter-Leistungsbauelemente - Untersuchungen zur Physik und Technologie. Wiss. Ber. AEG-Telefunken **55**(1–2) (1982)

[Scu89] Schulze, H.J., Kuhnert, R.: Realization of high voltage planar junction termination for power devices. Solid State Electron. **32**(S), 175 (1989)

[Scu02] Schulze, H.J., Niedernostheide, F.J., Schmitt, M., Kellner-Werdehausen, U., Wachutka, G.: Influence of irradiation-induced defects on the electrical performance of power devices. In: ECS Proceedings, 2002-20, pp. 320–335 (2002)

[Scu16] Schulze, H.J., Öfner, H., Niedernostheide, F.J., Laven, J.G., Felsl, H.P., Voss, S., Schwagmann, A., Jelinek, M., Ganagona, N., Susiti, A., Wübben, T., Schustereder, W., Breymesser, A., Stadtmüller, M., Schulz, A., Kurz, T., Lükermann, F.: Use of 300 mm magnetic Czochralski wafers for the fabrication of IGBTs. In: Proceedings of the 28st ISPSD, Prague, pp. 355–359 (2016)

[Sie01] Siemieniec, R., Netzel, M., Südkamp, W., Lutz, J.: Temperature dependent properties of different lifetime killing technologies on example of fast diodes. IETA2001, Cairo (2001)

[Sie02] Siemieniec, R., Südkamp, W., Lutz, J.: Determination of parameters of radiation induced traps in silicon. Solid-State Electron. **46**, 891–901 (2002)

[Sie06] Siemieniec, R., Niedernostheide, F.J., Schulze, H.J., Südkamp, W., Kellner-Werdehausen, U., Lutz, J.: Irradiation-induced deep levels in silicon for power device tailoring. J. Electrochem. Soc. **153**(2), G108–G118 (2006)

[SIL06] Siltronic, A.G.: Float Zone Silicon at Siltronic. www.siltronic.com/int/media/publication/…/Leaflet_Floatzone_en.pdf (2006)

[Ste85] Stengl, R., Gösele, U.: Variation of lateral doping – a new concept to avoid high voltage breakdown of planar junctions. In: IEEE IEDM 85, pp. 154 ff (1985)

[Sue94] Südkamp, W.: DLTS-Untersuchung an tiefen Störstellen zur Einstellung der Trägerlebensdauer in Si-Leistungsbauelementen, Dissertation, Technical University of Berlin (1994)

[Swe83] Swenson, C.A.: Recommended values for the thermal expansivity of silicon from 0 to 1000 K. J. Phys. Chem. Ref. Data. **12**, 179–182 (1983)

[Sze88] Sze, S.M.: VLSI Technology. McGrawHill, New York (1988)

[Sze02] Sze, S.M.: Semiconductor Devices, Physics and Technology, 2nd edn. Wiley, New York (2002)

[Tan59] Tannenbaum, M.: Uniform n-type silicon, U.S. patent 3076732, filled 12 Dec 1959

[Tan61] Tannenbaum, M., Mills, A.D.: Preparation of uniform resistivity n-type silicon by nuclear transmutation. J. Electrochem. Soc. **108**, 171–176 (1961)

[Tan10] Tang, H., Fang, Z.Q., Rolfe, S., Bardwell, J.A., Raymond, S.: Growth kinetics and electronic properties of unintentionally doped semi-insulating GaN on SiC and high-resistivity GaN on sapphire grown by ammonia molecular-beam epitaxy. J. Appl. Phys. **107**, 103701 (2010)

[Tru60] Trumbore, F.A.: Solid solubilities of impurity elements in germanium and silicon. Bell Syst. Tech. J. **39**, 205–233 (1960)

[Tsa83] Tsai, J.C.C.: Diffusion. In: Sze, S.M (eds.) VLSI Technology, McGraw-Hill Book Company, pp. 169–218 (1983)

[Ued05] Ueda, D., et al.: AlGaN/GaN Devices for Future Power Switching Systems. IEEE International Electron Devices Meeting, IEDM Technical Digest, pp. 377–380 (2005)

[Ued17] Ueda, D.: Properties and Advantages of Gallium Nitride. In Meneghini, M., Meneghesso, G., Zanoni, E. (eds.) Power GaN Devices - Materials, Applications and Reliability, Springer International Publishing, Switzerland (2017)

[Ura99] Ural, A., Griffin, P.B., Plummer, J.D.: Fractional contributions of microscopic diffusion mechanisms for common dopants and self-diffusion in silicon. J. Appl. Phys. **85**, 6440 ff (1999)

[Van04] Van de Walle, C.G., Neugebauer, J.: First-principles calculations for defects and impurities: Applications to III-nitrides. J. Appl. Phys. **95**, 3851 (2004)

[Vob02] Vobecký, J., Hazdra, P.: High-power P-i-N diode with the local lifetimecontrol based on the proximity gettering of platinum. IEEE Electron Device Lett. **23**(7), 392–394 (2002)

[Vob07] Vobecký, J., Hazdra, P.: Radiation-enhanced diffusion of palladium for a local lifetime control in power devices. IEEE Trans. Electron Devices **54**(6), 1521–1526 (2007)

[Vob09] Vobecký, J., Záhlava, V., Hemmann, K., Arnold, M., Rahimo, M.: The radiation enhanced diffusion (RED) diode - realization of a large area $p^+p^-n^-n^+$ structure with high SOA. In: Proceedings of the 21st ISPSD, Barcelona, pp. 144–147 (2009)

[Won85] Wondrak, W.: Erzeugung von Strahlenschäden in Silizium durch hochenergetische Elektronen und Protonen, Dissertation, Frankfurt (1985)

[Won87] Wondrak, W., Boos, A.: Helium implantation for lifetime control in silicon power devices. In: Proceedings of ESSDERC 87, Bologna, pp. 649–652 (1987)

[Yim74] Yim, W.M., Paff, R.J.: Thermal expansion of AlN, sapphire, and silicon. J. Appl. Phys. **45**(3) (1974)

[Zie06] Ziegler, J.F., Biersack, J.P.: The stopping and range of ions in matter. [Online]. http://www.srim.org/SRIM/SRIMINTRO.htm. Accessed 1 Mar 2006

Chapter 5
pin Diodes

Most power diodes are pin-diodes, i.e. they possess a middle region with a much lower doping concentration than the outer p and n layers enclosing it. Compared with unipolar devices (see Chap. 6), pin-diodes have the advantage that the on-resistance is strongly reduced by high-level injection in the base region, which is known as conductivity modulation. Hence pin-diodes can be used up to very high blocking voltages. The base region is not intrinsic, as suggested by the name. The intrinsic case – doping in the range of $<10^{10}$ cm^{-3} – would not only be difficult to attain by technology, it would be even disadvantageous for the turn-off behavior and other properties. Power diodes usually have a $p^+n^-n^+$-structure, hence the so-called "i"-layer is actually an n^--layer. Since it is several orders of magnitude lower than the doping of the outer layers, the name pin-diode has become the usual denotation in almost every case.

From the viewpoint of application, power diodes can be distinguished in two main types:

Rectifier diodes for grid frequency of 50 Hz or 60 Hz: the switching losses play a subordinate role, and there is a high carrier lifetime in the middle layer.

Fast recovery diodes that work as freewheeling diodes for a switching device or that are in the output rectifier after a high-frequency transformer. They have to be generally capable of switching frequencies of up to 20 kHz and in switch-mode power supplies of 50–100 kHz and more. In fast diodes manufactured from silicon, the charge carrier lifetime in the middle low-doped layer has to be reduced to a defined low value.

5.1 Structure of the pin Diode

With respect to technology and doping profile, pin-diodes can be classified into two types. For pin diodes using epitaxial technology (epitaxial diodes, Fig. 5.1a), first, an n^--layer is deposited by epitaxy on a highly doped n^+-substrate. Then, the

© Springer International Publishing AG 2018

J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8_5

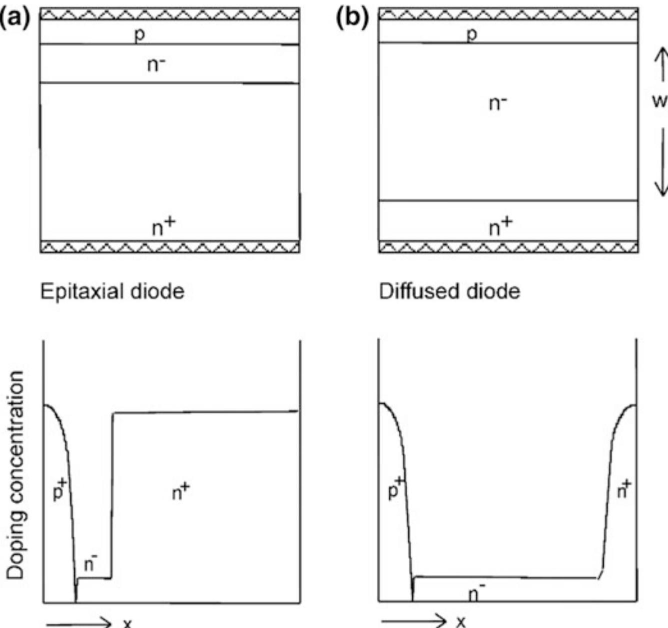


Fig. 5.1 Structure of pin power diodes. **a** Epitaxial diode. **b** Diffused diode

p-layer is diffused. With this process a very small base width w_B down to some micrometers can be created, whereby the silicon wafer is thick enough by the substrate to allow production with low wafer breaking and at high yield. By implementing recombination centers – in most cases by gold or platinum diffusion – very fast diodes can be realized. Since w_B is kept very small, the voltage drop across the middle layer is low. Epitaxial (epi-) diodes are mainly applied for blocking voltages between 100 and 600 V, however, some manufacturers also produce 1200 V with epi-diodes.

Because the costs of the epitaxy process are notable, diodes for higher blocking voltages – usually 1200 V and above – are fabricated by diffusion. For a diffused pin-diode (Fig. 5.1b), one starts with a low doped wafer in which the p^+-layer and the n^+-layer are created by diffusion. The thickness of the wafer now is determined by the thickness w_B of the middle n^--layer and the depths of the diffusion profiles. The required w_B is small for lower voltages. With deep n^+ and p^+-layers the wafer thickness can be increased again, but deep p-layers have disadvantages regarding the reverse-recovery behavior. The processing of such thin wafers is challenging. Infineon has introduced a technology for handling very thin wafers, down to a thickness of < 60 μm in the manufacturing process. With this technology, also freewheeling diodes for 600 V with shallow p and n^+ boarder layers can be fabricated as diffused diodes.

5.2 I–V Characteristic of the pin Diode

The I-V characteristic of a fast 300-V pin-diode measured at 25 °C as well as some definitions for parameters of the I–V characteristic are shown in Fig. 5.2. In the figure different scales in forward and reverse direction are applied. In forward bias, the characteristic associates a defined current I_F to the voltage drop V_F. This has to be distinguished from the maximally allowed voltage drop V_{Fmax} specified in manufacturers' data sheets. V_{Fmax} is the maximum forward voltage drop that can occur at a diode of this type under specified conditions. In most cases this value is significantly higher than the value measured for an individual sample due to tolerances of parameters during production, e.g. of the base width w_B. In fast diodes, the carrier lifetime strongly effects V_F. For fast gold- or platinum-diffused diodes of older generations, relatively high variations in the carrier lifetime are typical due to the difficulties to control these technologies, see Chap. 4. Some manufacturers also specify typical values, but this is not guaranteed for an individual sample.

In reverse bias, V_{BD} is the physical breakdown voltage of a given sample. I_R denotes the leakage current measured at a defined reverse voltage. This has to be distinguished from V_{RRM}, the maximal reverse voltage that is specified in the data sheet, as well as from I_{RM}, the specified maximal leakage current at V_{RRM}. Since the manufacturer takes into account data scattering and may add additional margins, at a single diode I_R cab be significantly lower and V_{BD} will be typically higher; the manufacture extends a warranty, however, only for V_{RRM} and I_{RM}, respectively.

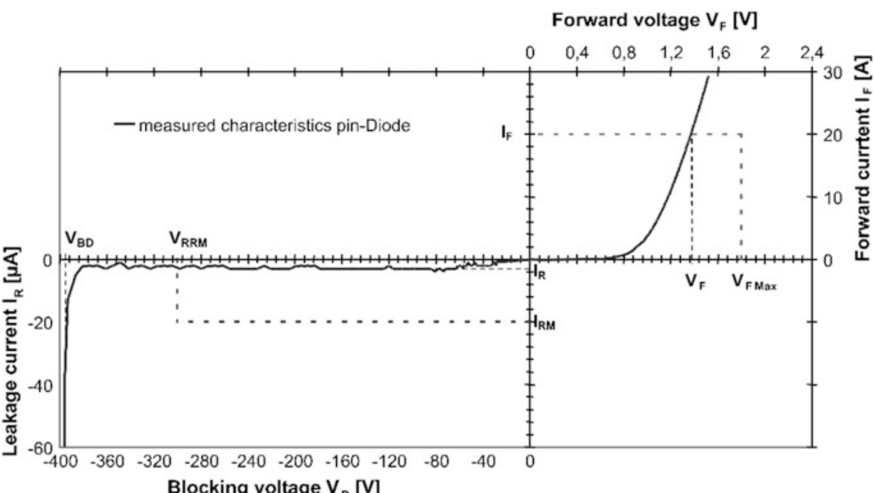


Fig. 5.2 I–V-characteristic of a fast pin-diode with some definitions of parameters

The I–V-characteristic of a diode is strongly temperature-dependent. With increasing temperature

- the leakage current I_R increases, I_R can be several orders of magnitude higher at the typical maximally allowed operation temperature of 150 °C than that at room temperature. Both, the diffusion component and generation component of the leakage current also increase; see Eq. (3.59) and Fig. 3.13
- the blocking voltage V_{BD} somewhat increases in accordance with the increase in breakdown voltage of avalanche breakdown; see Eq. (3.79) with the temperature dependent parameters in Eq. (3.90)
- the built-in voltage V_{bi} decreases, because according to Eq. (3.6), the determining parameter for the temperature dependency is the strongly temperature-dependent n_i^2. This corresponds to a decreased threshold voltage V_s. Also in the derivation of a threshold voltage from Eqs. (3.51) and (3.52), n_i^2 dominates the result.

5.3 Design and Blocking Voltage of the pin Diode

A dominating parameter for all characteristics of the diode is the width w_B of the low doped base-region. First of all, the base width together with the base doping concentration determines the blocking voltage. As illustrated in Fig. 5.3, different cases of the field shape during blocking near breakdown can be distinguished.

If w_B and N_D are chosen such that the space-charge region does not reach the n^+-layer (triangular field shape), the design is called a *Non-Punch-Through (NPT) dimensioning* [Bal87]. If w_B and N_D is chosen such that the space-charge penetrates the n^+-layer, then the field shape is trapezoidal or rectangular (Fig. 5.3b, c), and the structure is said to be of the *Punch-Through (PT)* type. The term 'punch-through' is used here in a different meaning as for thyristors. Whereas in the latter devices the space-charge region reaches up to a layer of opposite type of conductivity leading to breakdown, in diodes with PT-design the space charge layer

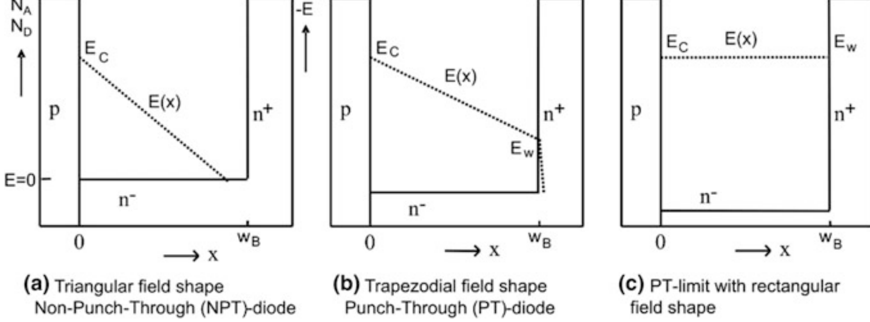


Fig. 5.3 Field distributions at breakdown in pin-($p^+n^-n^+$)-diodes with different design

is stopped by a highly doped region of the same type. The notation of diode design by NPT and PT is widely used. As shown by the area below the $|\mathsf{E}(x)|$-line, the PT-design yields a higher breakdown voltage for a given base width w_B, so it is enabling a smaller base width for a given V_{BD} than in the NPT-case. Fast pin power diodes of old generations have been dimensioned as NPT-diodes, because for a PT-design an acceptable reverse recovery behavior is more difficult to achieve. Nowadays, optimized fast diodes (see Sect. 5.7.4) use a PT-design to achieve low forward voltage drop and low stored charge. IGBTs are designed with trapezoidal field shape using a 'field stop layer' at the end of the n-base. This results in a significant reduction of the on-sate voltage drop. Also MOSFETs and Schottky diodes use mostly the PT-concept. In the present section we investigate how the base width necessary for a given blocking voltage in different cases of PT-design (Fig. 5.3b, c) compares with NPT-dimensioning. The p^+n- and nn^+-junctions are assumed to be abrupt. The power approach (3.75), $\alpha_{eff} = B|\mathsf{E}|^n$, is used with exponent $n = 7$, the constant B has then the value 2.11×10^{-35} cm^6/V^7 according to (3.83).

For NPT design, the relationships between breakdown voltage, base doping concentration and extension of the space charge region at breakdown have been given already in Sect. 3.3.2. Equation (3.85) says that width and doping density of the base region must satisfy for NPT the condition

$$w_B \geq w = \left(\frac{8}{B}\right)^{1/8} \cdot \left(\frac{\varepsilon}{q \cdot N_D}\right)^{7/8}$$

As condition, which w_B must satisfy for a given breakdown voltage V_{BD}, the inversion of (3.80) yields

$$w_B \geq w = 2^{2/3} \cdot B^{1/6} \cdot V_{BD}^{7/6} \tag{5.1}$$

pin-diodes with NPT-design in silicon must satisfy at 300 K these conditions. The NPT-case is assumed to include the equal sign, i.e. the case where the space-charge region just touches the n^+ layer without being stopped by it. If not stated otherwise, we are just dealing in what follows with the minimum base width $w_B = w$ satisfying the above conditions.

We consider now diodes with PT-dimensioning. Base width and doping density are then smaller than according to (5.1) and the preceding equation. The space-charge region penetrates the n^+-layer, where the field drops rapidly to zero as is depicted in Fig. 5.3b. To calculate the blocking voltage one has to determine first the critical field strength E_c, which can be different from that in the NPT case. The electric field over the base region at breakdown is given by

$$\mathsf{E}(x) = -\mathsf{E}_c + \frac{q \cdot N_D}{\varepsilon} \cdot x \tag{5.2}$$

Inserting this into $\alpha_{eff} = B \cdot |\mathsf{E}(\mathsf{x})|^7$ the breakdown condition (3.71) takes the form

$$\int_0^{w_B} B \cdot \left(\mathsf{E}_c - \frac{q \cdot N_D}{\varepsilon} \cdot x\right)^7 dx = 1 \tag{5.3}$$

The integration yields

$$\mathsf{E}_c^8 - \left(\mathsf{E}_c - \frac{q \cdot N_D}{\varepsilon} w_B\right)^8 = \frac{8 \cdot q \cdot N_D}{\varepsilon \cdot B} \tag{5.4}$$

Hence one can write:

$$\mathsf{E}_c = \left(\frac{8qN_D}{B\varepsilon} + \mathsf{E}_w^8\right)^{1/8} \tag{5.5}$$

with

$$\mathsf{E}_w = \mathsf{E}_c - \frac{qN_D}{\varepsilon} \cdot w_B \tag{5.6}$$

E_w represents the absolute value of the field strength at the nn^+ junction (see Fig. 5.3b). The implicit Eq. (5.5) for E_c can be solved by iteration, which except for very low N_D and $\mathsf{E}_w \approx \mathsf{E}_c$ converges rapidly. With E_c determined, the breakdown voltage is given by

$$V_{BD} = \frac{\mathsf{E}_c + \mathsf{E}_w}{2} = \left(\mathsf{E}_c - \frac{qN_D}{2\varepsilon} w_B\right) \cdot w_B \tag{5.7}$$

The breakdown voltage calculated in this way is plotted in Fig. 5.4 for a given base width (85 μm) as function of N_D (solid line). With decreasing N_D, the breakdown voltage increases monotonously. This holds also for other base widths.

Often one uses a design of moderate or weak punch-through with $\mathsf{E}_w \leq \mathsf{E}_c/2$. Then the term for E_w^8 in (5.5) is negligible and one obtains for the critical field.

$$\mathsf{E}_c = \left(\frac{8 \cdot q \cdot N_D}{B \cdot \varepsilon}\right)^{1/8} \tag{5.8}$$

This equation is identical with (3.78) for $n = 7$. From (5.7) and (5.8) the breakdown voltage is obtained as an explicit function of doping density and base width:

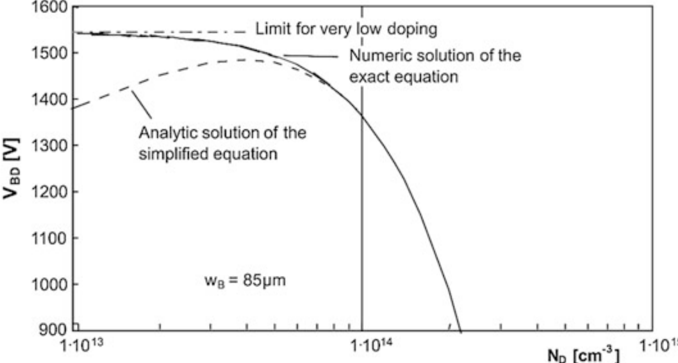


Fig. 5.4 Blocking voltage of a p⁺nn⁺ diode with given base width as a function of doping concentration of the base region

$$V_{BD} = \left(\frac{8qN_D}{\varepsilon B}\right)^{\frac{1}{8}} w_B - \frac{1}{2}\frac{qN_D}{\varepsilon} w_B^2 \tag{5.9}$$

This dependency on N_D for w_B = 85 µm is plotted in Fig. 5.4 too (dashed curve). At high N_D down to 8×10^{13} cm^{-3}, where $E_w/E_c = 0.51$, the curve coincides with the exact curve. Below this point the approximate solution shows a maximum, but becomes more and more incorrect. In some textbooks it is concluded from this approximation that the breakdown voltage has in fact a maximum at an optimal doping density and from this point *decreases* with decreasing doping. As is shown by the figure, this is an error. Neglecting the term E_w^8 in (5.5) is only allowed as long as doping is not too low.

The limit for very low doping can be derived immediately from the ionization integral. In this case, the field shape is rectangular, $E_w = E_c$, and the condition for avalanche breakdown (3.71) with the power approximation of ionization rates with n = 7 is

$$B \cdot \int_0^{w_B} E_c^7 dx = 1 \tag{5.10}$$

Hence it follows $E_c = 1/(Bw_B)^{1/7}$ and for $V_{BD} = E_c \cdot w_B$

$$V_{BD} = \left(\frac{w_B^6}{B}\right)^{\frac{1}{7}} \tag{5.11}$$

This limiting value for the example w_B = 85 µm is also depicted in Fig. 5.4. The blocking voltage approaches this limit very quickly already for a doping in the range of 2×10^{13} cm^{-3}.

For the base width as function of V_{BD}, (5.11) leads to

$$w_B(PT, lim) = B^{\frac{1}{6}} \cdot V_{BD}^{\frac{7}{6}} \quad (5.12)$$

This lowest attainable base width is by a factor $2^{2/3} = 1.59$ smaller than the minimum w_B for NPT-design given by (5.1):

$$w_B(PT, lim) = 2^{-\frac{2}{3}} w_B(NPT) \cong 0.63 \cdot w_B(NPT) \quad (5.13)$$

Although $w_B(PT,lim)$ is not half of $w_B(NPT)$, as would follow for equal critical field strength in both cases, the reduction by the PT-design is very significant.

Figure 5.5 shows the minimum width of the middle region for PT-design (5.12) and for NPT-design (5.1) as a function of breakdown voltage. The difference between w_B in the two cases results in a significant difference in forward voltage drop. For fast diodes with a rated voltage of 1200 V and with the necessary low carrier lifetime, the difference in V_F amounts up to 0.8 V. For higher voltages the reduction of forward voltage by the PT dimensioning is even larger. This is significant for the conduction losses. Consequently, a PT-dimensioning should be applied if possible. Nonetheless, this design also poses challenges, in particular regarding reverse-recovery behavior, as will be shown later.

A very high field at the nn$^+$-junction has on the other hand some disadvantages from a technological viewpoint. One drawback is that a field $E_w \approx E_c$ requires much more effort for edge termination at the surface. Therefore, only a moderate PT-dimensioning, often with $E_w \leq {}^1/_2\ E_c$ or smaller, is preferred. If the ratio E_w/E_c is kept independent of the breakdown voltage, the dependency $w_B \sim V_{BD}^{7/6}$ is obtained for each value of E_w/E_c, where however the proportionality factor varies with the field ratio. For $E_w/E_c = 1/2$, the calculation yields

$$w_B(PT) = 0.70 \cdot w_B(NPT) \quad (5.14)$$

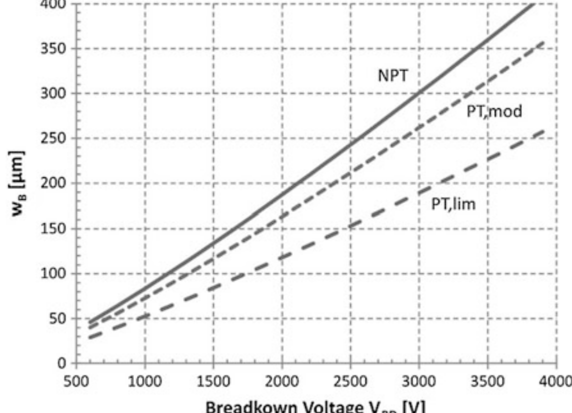


Fig. 5.5 Width w_B of the base region for dimensioning with triangular (NPT) and with rectangular field shape (PT,limit). The additional curve (PT,mod) refers to a moderate PT used for fast pin-diodes

In practice, tolerances in background doping and the adjusted base width implicate that the theoretical breakdown voltage is often not fully attained. Also the junction termination structure at the surface does not lead to 100% of the volume breakdown voltage in many cases. Hence the theoretical breakdown voltage must exceed the wanted maximum blocking voltage by a certain percentage. Taken from experience, an orientation value for a moderate PT-dimensioning can be given by:

$$w_B = \chi \cdot V_{BD}^{\frac{7}{6}} \quad \text{with} \quad \chi = 2.3 \times 10^{-6}\,\text{cmV}^{-\frac{7}{6}} \tag{5.15}$$

with a doping concentration chosen suitably somewhat below the NPT-value. The dependency (5.15) is depicted in Fig. 5.5 as the curve indicated with 'PT,mod'.

The described considerations on dimensioning are not only of relevance for diodes but also for other power devices whose base region contains a higher doped buffer layer to limit the extension of the space charge region. Particularly Schottky-diodes, MOSFETs and modern IGBTs use a PT-design. A discussion specifically on unipolar device design is given in Sect. 6.4.

A power device has to function over the temperature range from about 230 K to more than 400 K. The temperature dependence of V_{BD} is influenced by the PT-design as well, because it is determined besides the T-dependence of the critical field also by the manner how E_c enters the breakdown voltage. For NPT-design $V_{BD}(T)$ is described by (3.78) together with the temperature dependent parameters $n(T)$ and $B(T)$ as given by (3.87). For PT-design, we confine ourselves to the case $\mathsf{E}_w/\mathsf{E}_c \leq 1/2$. To get V_{BD} in dependence of temperature, Eq. (5.9) has to be used in a form which contains n and B as T–dependent variables. This is obtained by substituting E_c in (5.9) by the expression (3.78) which yields:

$$V_{BD} = \left(\frac{(n+1)qN_D}{\varepsilon B}\right)^{\frac{1}{n+1}} w_B - \frac{1}{2}\frac{qN_D}{\varepsilon} w_B^2 \tag{5.16}$$

$n(T)$ and $B(T)$ has to be inserted from (3.90). Since the critical field enters (5.16) only linearly, while for NPT-design V_{BD} is proportional to E_c^2 according to (3.79), Eq. (5.16) predicts a smaller variation with temperature than for NPT-layout. The cause is that the base width for PT-design is constant, whereas the width of the space-charge region in the NPT-case increases with T proportional to E_c.

Since the blocking capability must be given at the lowest operation temperature, the decrease of V_{BD} from room temperature downwards is of particular interest. Often the nn$^+$-junction is not abrupt, but the doping increases gradually. Then the measured temperature dependency is found to be between the predictions of (5.16) and (3.79).

Equation (5.16) can be used also for devices made of other semiconductors, if the PT-design with $E_w \lesssim {}^1/_2\, E_c$ is used and the parameters n and B are known. Thus the breakdown voltage of diodes in 4H–SiC-diodes can be calculated using the data for n und B given in Sect. 3.3.3.

5.4 Forward Conduction Behavior

5.4.1 Carrier Distribution

At forward bias, the lowly doped base zone of a pin-diode is flooded with carriers injected from the highly doped outer regions. The density of free carriers is increased by a few orders of magnitude above the background doping, and the conductivity in the low-doped zone is strongly enhanced or 'modulated'. Because in the neutrality condition $n = p + N_D^+$, the doping concentration N_D^+ is negligible, the hole and electron concentration in the base are approximately equal:

$$n(x) \approx p(x).$$

Figure 5.6 shows the carrier distribution $n = p$ in the middle region calculated for a 1200 V diode at a current density of 160 A/cm^2. Here, the emitter efficiency of highly doped regions has been assumed to be 1. n and p exceed the background doping by more than two orders of magnitude. The broken line is obtained if there would be equal mobilities for electrons and holes. Since the electrons in silicon are much more mobile than the holes, an asymmetric distribution, which is increased at the pn-junction, results.

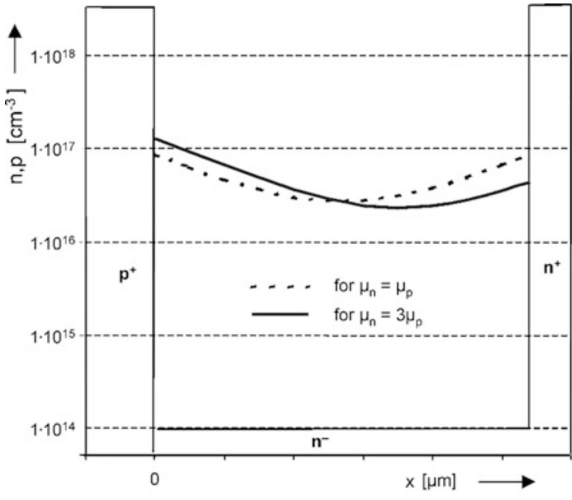


Fig. 5.6 Distribution of charge carriers in the base-region at forward conduction. Example of a diode with $\tau_{HL} = 0.48$ µs, $w_B = 108$ µm. The p$^+$- and n$^+$-layers are assumed to be ideal emitters

To calculate the carrier distribution, one starts with the transport Eqs. (2.43a) and (2.43b) which for $n = p$ take the form:

$$j_p = q \cdot \mu_p \cdot p \cdot \mathsf{E} - q \cdot D_p \cdot \frac{dp}{dx} \tag{5.17}$$

$$j_n = q \cdot \mu_n \cdot p \cdot \mathsf{E} + q \cdot D_n \cdot \frac{dp}{dx} \tag{5.18}$$

Since in the stationary case, which we consider, the total current density $j = j_n + j_p$ is independent of x according to (2.106), the electric field strength E is suitably expressed by j. Adding (5.17) and (5.18) one obtains

$$j = j_n + j_p = q \cdot (\mu_n + \mu_p) \cdot p \cdot \mathsf{E} + q \cdot (D_n - D_p) \cdot \frac{dp}{dx} \tag{5.19}$$

Hence, the electric field is

$$\mathsf{E} = \frac{\frac{j}{q} - \left(D_n - D_p\right) \cdot \frac{dp}{dx}}{\left(\mu_n + \mu_p\right) \cdot p} \tag{5.20}$$

The first term proportional to j is the resistive or ohmic field. The second field term determined by the concentration gradient and the difference in the diffusion constants is denoted as "Dember-field". The current caused by this field and the diffusion current in (5.19) add up to zero. Inserting Eq. (5.20) into (5.17) and (5.18), one obtains

$$j_p = \frac{\mu_p}{\left(\mu_n + \mu_p\right)} \cdot j - q \cdot D_A \cdot \frac{dp}{dx} \tag{5.21}$$

$$j_n = \frac{\mu_n}{\left(\mu_n + \mu_p\right)} \cdot j + q \cdot D_A \cdot \frac{dp}{dx} \tag{5.22}$$

where D_A is the following combination of the single diffusion constants:

$$D_A = \frac{2 \cdot D_n \cdot D_p}{D_n + D_p} \tag{5.23}$$

For this relationship, the Einstein relation $D_{n,p} = (kT/q) \cdot \mu_{n,p}$ has been used. D_A is called 'ambipolar diffusion constant', and the terms $\mp q D_A \cdot dp/dx$ are the ambipolar diffusion current densities. These include the current resulting from the Dember field, which makes the ambipolar diffusion current of electrons and holes oppositely equal. To obtain a differential equation for the carrier concentration, one

has to insert (5.20) or (5.21) into the relevant continuity equation. According to (2.103), the continuity equation for holes in the stationary case can be written

$$\frac{dj_p}{dx} = -q \cdot R = -q \frac{p}{\tau_{HL}} \tag{5.24}$$

where the recombination rate R is expressed according to Eqs. (2.49), (2.51) by p and the high level carrier lifetime τ_{HL}, neglecting the equilibrium carrier density. Inserting j_p from (5.21) the following differential equation for the carrier concentration is obtained:

$$D_A \cdot \frac{d^2 p}{dx^2} = \frac{p}{\tau_{HL}} \tag{5.25}$$

where the mobility ratios and ambipolar diffusion constant are assumed to be constant. In what follows, also the lifetime τ_{HL} is assumed to be constant, which agrees with Eq. (2.74) of the SRH model. Then the differential Eq. (5.25) has the solutions $p(x) \propto \exp(\pm x/L_A)$, where L_A is the ambipolar diffusion length defined by

$$L_A = \sqrt{D_A \cdot \tau_{HL}} \tag{5.26}$$

In addition to the differential equation, the wanted carrier distribution has to satisfy conditions at the boundaries of the base region which are given by the emitter efficiency of the p^+ and n^+ region. In the present section, we assume that these regions are ideal emitters (with emitter efficiency 1), meaning that $j_n = 0$ at the p^+n junction and $j_p = 0$ at the nn^+-junction. Using these current densities in (5.21) and (5.22) one obtains the boundary conditions:

$$q\,D_p \frac{dp}{dx}(0) = -j/2, \qquad q\,D_n \frac{dp}{dx}(w_B) = j/2 \tag{5.26a}$$

To the same extent as $D_p < D_n$, the absolute value of the concentration gradient at $x = 0$ (p^+-side of the base region) is higher than at $x = w_B$ (n^+-side). Considering the simplified case $D_n = D_p = D_A = D$, the solution of (5.25) satisfying the conditions (5.26a) is

$$p(x') = \frac{jL_A}{2\,q\,D\,\sinh(w_B/(2 \cdot L_A))} \cosh\left(\frac{x'}{L_A}\right)$$

Here the coordinate x' has its origin in the middle of the base region: $x' = x - w_B/2$. The symmetrical carrier distribution in Fig. 5.6 (dashed line) was calculated using this equation together with the parameters given in the legend. The obtained cosh-distribution is well-known from the chainline. The smaller the ambipolar diffusion length L_A, the more pronounced is the sagging of the carrier distribution.

In the real case of different mobilities of electrons and holes, the carrier distribution can be written as

$$n(x') = p(x') = \frac{j \cdot \tau_{HL}}{2 \cdot q \cdot L_A} \left(\frac{\cosh \frac{x'}{L_A}}{\sinh \frac{w_B}{2 \cdot L_A}} - \frac{\mu_n - \mu_p}{\mu_n + \mu_p} \cdot \frac{\sinh \frac{x'}{L_A}}{\cosh \frac{w_B}{2 \cdot L_A}} \right) \tag{5.27}$$

It can be verified that this solution of (5.25) satisfies the boundary conditions (5.26a). The carrier distribution calculated from (5.27) for the given parameters is plotted in Fig. 5.6 as solid line. The concentration is at the pn-junction more than a factor 2 higher than that at the nn$^+$-junction (which later will prove to be a disadvantage for the reverse recovery behavior). The minimum has shifted towards the nn$^+$-junction. The term with $sinh(x'/L_A)$ in (5.27) expresses this asymmetry. Since for silicon $\mu_n \approx 3\mu_p$, the factor before this term consisting of the mobilities is approximately 0.5. As is noted, integration of (5.27) yields for the mean carrier concentration across the base:

$$\bar{n} = \bar{p} = \frac{1}{w_B} \int_{-w_B/2}^{w_b/2} p \cdot dx' = \frac{j \cdot \tau_{HL}}{q \cdot w_B}$$

For real p$^+$ and n$^+$-regions with emitter efficiencies < 1, the boundary conditions change and the factors in (5.27) are no longer valid. The carrier distribution in the base is written in this case suitably in the form:

$$p(x) = \frac{1}{\sinh(w_B/L_A)} \left(p_R \sinh(\frac{x}{L_A}) + p_L \sinh(\frac{w_B - x}{L_A}) \right) \tag{5.28}$$

p_L, p_R are the concentrations at the left and right edges of the base region and have to be determined from the general boundary conditions (see Sect. 5.4.5).

5.4.2 Junction Voltages

In a p$^+$nn$^+$-structure one has a space-charge region at the p$^+$n-junction and another at the nn$^+$-doping step, whereby both are connected with a built-in voltage V_{bi}(p$^+$n) and V_{bi}(nn$^+$), respectively. If a forward voltage V_F is applied, a part of it is used at the junctions to reduce the potential steps there and to raise the injected carrier densities in the base region similarly as for a single pn-junction. Additionally, the forward voltage provides for an ohmic voltage drop V_{drift} over the weakly doped base region needed for the current transport. Hence, if the junction parts are called V_j(p$^+$n), V_j(nn$^+$), one obtains:

$$V_F = V_j(p^+n) + V_{drift} + V_j(nn^+) \tag{5.29}$$

The internal voltage steps at the junctions

$$\Delta V(p^+n) = V_{bi}(p^+n) - V_j(p^+n),$$

$$\Delta V(nn^+) = V_{bi}(nn^+) - V_j(nn^+)$$

are related to the carrier concentrations at the neutral boundaries of the space-charge regions via Boltzmann factors. For the hole concentration p_L near the p^+n-junction in the neutral base and the electron density n_R at the n^+-side of the base (see Fig. 5.6), one obtains:

$$\frac{p_L}{p^+} = \exp\left(-\frac{q \cdot \Delta V(p^+n)}{kT}\right), \quad \frac{n_R}{n^+} = \exp\left(-\frac{q \cdot \Delta V(nn^+)}{kT}\right) \tag{5.30}$$

where the carrier densities p^+, n^+ of the highly doped regions are given by the doping densities N_A, N_D. Dividing the relationships (5.30) by the corresponding thermal equilibrium equations (see Chap. 3)

$$\frac{p_{n0}}{p^+} = \exp\left(-\frac{q \cdot V_{bi}(p^+n)}{kT}\right), \qquad \frac{N_D}{n^+} = \exp\left(-\frac{q \cdot V_{bi}(nn^+)}{kT}\right),$$

one obtains for the external parts to the voltage drop at the junctions:

$$\begin{aligned} V_j(p^+n) &= \frac{kT}{q}\ln\frac{p_L}{p_{n0}} = \frac{kT}{q}\ln\frac{p_L \cdot N_D}{n_i^2} \\ V_j(nn^+) &= \frac{kT}{q}\ln\frac{n_R}{N_D} \end{aligned} \tag{5.31}$$

$V_j(p^+n)$ and $V_j(nn^+)$ differ significantly from one another depending on the doping density N_D of the base region. By leveling out the strong difference of built-in voltages, they make the internal voltage steps $\Delta V(p^+n)$ and $\Delta V(nn^+)$ more similar. The sum of both, the total external junction voltage V_j, does not depend on N_D:

$$V_j \equiv V_j(p^+n) + V_j(nn^+) = \frac{kT}{q}\ln\frac{p_L \cdot n_R}{n_i^2} \tag{5.32}$$

These equations hold independently of the injection level, but they will be used below for high-injection conditions where $p_L = n_L$ and $n_R = p_R$.

For the example of Fig. 5.6 with $p^+ = 2 \times 10^{18}$ cm^{-3}, $N_D = 7 \times 10^{13}$ cm^{-3} as doping of the middle layer, and $n^+ = 1 \times 10^{19}$ cm^{-3}, one obtains $V_{bi}(p^+n) = 0.721$ V,

$V_{bi}(\mathrm{nn}^+) = 0.307$ V. With p_L and p_R from the asymmetrical carrier distribution in Fig. 5.6, one attains $V_j(\mathrm{p}^+\mathrm{n}) = 0.654$ V, $V_j(\mathrm{nn}^+) = 0.161$ V, $V_j = 0.815$ V. For the total internal voltage steps, the results are $\Delta V(\mathrm{p}^+\mathrm{n}) = 0.067$ V, $\Delta V(\mathrm{nn}^+) = 0.136$ V.

5.4.3 Voltage Drop Across the Middle Region

Now the voltage V_{drift} that drops across the middle region has to be calculated. V_{drift} results by integration of the electric field which is given in Eq. (5.20). It contains in the nominator as second term the Dember field, which is proportional to the gradient of the carrier density. This term leads to the Dember voltage V_{Dem}. Using the Einstein relation (2.44), it follows from (5.20) that

$$V_{Dem} = \frac{kT}{q} \cdot \frac{\mu_n - \mu_p}{\mu_n + \mu_p} \cdot \ln \frac{p_L}{p_R} \tag{5.33}$$

For a symmetric carrier distribution this voltage term vanishes since $p_L = p_R$. But also for the actual asymmetrical distributions V_{Dem} is very small. For the example in Fig. 5.6, $V_{Dem} = 14.3$ mV results, which can be neglected.

From the term in Eq. (5.20) which is proportional to the current density, the voltage over the middle region is obtained as:

$$V_{drift} = \frac{j}{q \cdot (\mu_n + \mu_p)} \int_0^{w_B} \frac{1}{p(x)} dx \tag{5.34}$$

For a homogeneous distribution $p(x) = const$ the integral would be equal to w_B/p. For a not too strongly inhomogeneous distribution, this holds approximately if p is replaced by the average value $\bar{p}$. This leads to

$$V_{drift} = \frac{j \cdot w_B}{q \cdot (\mu_n + \mu_p) \cdot \bar{p}} \tag{5.35}$$

As can be verified by carrying out the integration in (5.34) with the exact carrier distribution (5.27) (see below), this is a good approach for $w_B/L_A \leq 3$, but can be used as a rough approximation up to $w_B/L_A = 4$. For $w_B/L_A = 3$ the error made with (5.35) is only 7%, although the inhomogenity is already considerable (cosh (1.5) = 2.35). The density of carriers represents a stored charge

$$Q_F = q \cdot A \cdot w_B \cdot \bar{p} \tag{5.36}$$

and, with this, it follows that

$$V_{drift} = \frac{I_F \cdot w_B^2}{(\mu_n + \mu_p) \cdot Q_F} \tag{5.37}$$

where $I_F = A \cdot j$ denotes the forward current (A area of the device).

5.4.4 Voltage Drop in the Hall Approximation

For the further calculation, we assume now as in Sect. 5.4.1, that recombination in the p^+ and n^+ region is negligible (emitter efficiency 1). This is suggested because the excess charge in the base is orders of magnitude higher than in the end regions (see Fig. 5.6). In the theory of diode characteristic this case is called Hall approximation [Hal52]. Since the minority carrier currents in the end regions are neglected, integration of the continuity Eq. (5.24) yields for the current density:

$$j = \frac{q \cdot w_B \cdot \bar{p}}{\tau_{HL}} \tag{5.38}$$

Multiplying (5.38) with $A \cdot \tau_{HL}$, one obtains for the stored charge:

$$Q_F = I_F \cdot \tau_{HL} \tag{5.38a}$$

Hence from (5.37) it follows that

$$V_{drift} = \frac{w_B^2}{(\mu_n + \mu_p) \cdot \tau_{HL}} \quad \text{for} \quad \mathrm{w_B/L_A} < 3 \tag{5.39}$$

Expressing τ_{HL} by the ambipolar diffusion length L_A defined in (5.26) together with (5.23), Eq. (5.39) can be written in the form

$$V_{drift} = f(b) \cdot \frac{kT}{q} \left(\frac{w_B}{L_A} \right)^2 \quad \text{with} \quad f(b) = \frac{2\,b}{(1+b)^2} \tag{5.40}$$

where $b \equiv \mu_n/\mu_p$. Equation (5.40) can be found in many textbooks on device physics. The factor $f(b)$ is always $\leq$ ½. For silicon, the mobility ratio is $b \approx 3$ which yields $f(b) \approx 3/8$. Like (5.37) Eqs. (5.39), (5.40) apply up to about $w_B/L_A = 4$. Even for $w_B = 4\,L_A$, V_{drift} amounts only to 0.16 V according to (5.40).

For a more pronounced sagging of the carrier distribution, the voltage drop is mainly produced by the region of lowest concentration, hence V_{drift} is underestimated by (5.39) and (5.40). The accurate formula for V_{drift} is obtained by carrying

out the integral in (5.34) with the exact distribution (5.27). With $d = w_B/2$ the result can be written as

$$V_{drift} = 4f(b) \cdot \frac{kT}{q} \cdot \sinh\left(\frac{d}{L_A}\right) \cdot \cosh\left(\frac{\Delta}{L_A}\right) \cdot \mathrm{arctg}\left(\frac{\sinh(d/L_A)}{\cosh(\Delta/L_A)}\right) \tag{5.41}$$

where Δ denotes the distance of the minimum of $p(x)$ from the middle of the base. The asymmetry enhances the voltage drop a little. In the limit $d/L_A \gg 1$, Eq. (5.41) turns into

$$V_{drift} = \pi f(b) \cdot \frac{kT}{q} \cdot \cosh\left(\frac{\Delta}{L_A}\right) \exp\left(\frac{d}{L_A}\right) \quad \text{for } \mathrm{w_B} \gg 2\mathrm{L_A} \tag{5.41a}$$

As mentioned, the equations of this section hold for the case of negligible recombination in the highly doped outer layers. However, the right hand side of (5.39) and (5.41) can be used essentially also in the general case as a factor in V_{drift}.

According to Eqs. (5.39) to (5.41), the voltage drop across the middle region does not depend (explicitly) on the current. The increase of current is neutralized by a proportional increase of the carrier concentration and so the ratio $j/\bar{p}$ in (5.35) is constant. Actually, the decrease of μ_n and μ_p causes an increase of V_{drift} with current. According to (5.38), the concentration $\bar{p}$ reaches very high values even at normal forward current densities. At such high concentrations, also the *lifetime* is significantly reduced because of Auger recombination. These effects together with the slight increase of the junction voltage according to Eq. (5.32) lead to an appreciable increase of the forward voltage with current. In the range where the Auger recombination predominates strongly and hence the lifetime in the base is very inhomogeneous, however, the analytical approach above becomes insufficient.

However, the application range of the equations is anyway restricted to much smaller concentrations and current densities. As mentioned, they are based on the condition of negligible recombination in the emitter regions, and this becomes significant at a decade smaller $\bar{p}$ than the Auger recombination. Typically the Hall approximation applies up to a current density of about 5 – 30 A/cm^2, depending on the lifetime and width of the base region. Above this limit the experimental forward voltage (see Fig. 5.2) is considerably higher and increases essentially stronger with current than calculated. This is supported by the observation that the stored charge increases essentially slower with current in the relevant range than according to the linear dependency (5.38a). In the next section, the theory will be expanded taking into account the emitter recombination, which resolves these discrepancies.

That the conductivity of the base region is enhanced very strongly by the high injection remains valid. However, due to this conductivity modulation, even high-voltage devices with their wide, weakly doped base region can be operated with high forward current densities without causing a very high forward voltage.

5.4.5 *Emitter-Recombination, Effective Carrier Lifetime and Forward Characteristic*

The influence of emitter recombination on the injection efficiency of pn-junctions has been described in detail in Sect. (3.4). Using (3.96) for the junctions of a forward biased pin-diode, a minority hole current $j_p(n^+) = q{\cdot}h_n{\cdot}p_R^2$ flows into the n$^+$-region, and an electron current $j_n(p^+) = q{\cdot}h_p{\cdot}p_L^2$ is flowing into the p$^+$-region, where h_p and h_n are the constants of the respective emitter region given by (3.107, 3.109). In the case of Fig. 5.7, these currents amount to $j_n(\mathrm{p}^+) \approx 32$ A/cm^2 and $j_p(\mathrm{n}^+) \approx 8$ A/cm^2, if h_p and h_n have a typical value of 2×10^{-14} cm^4/s (see Fig. 3.21). Hence, the current due to recombination in the end regions amounts to 25% of the base recombination current in this case where the lifetime in the base is very small (0.48 µs). For a higher τ_{HL}, higher current density or intentionally reduced emitter efficiency, the emitter recombination current can predominate.

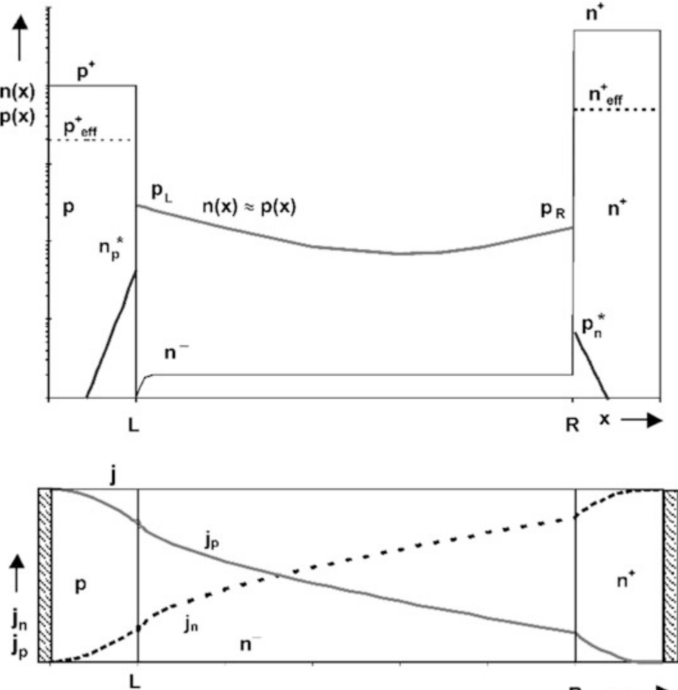


Fig. 5.7 pin-Diode with consideration of recombination in the border regions

To calculate the influence of emitter recombination on the forward characteristics of a pin-diode, we introduce an effective carrier lifetime τ_{eff} by [Sco69]:

$$\frac{\int_{-\infty}^{\infty} \Delta p dx}{\tau_{eff}} = \int_{-\infty}^{\infty} \frac{\Delta p}{\tau_p} dx \tag{5.42}$$

By this definition, τ_{eff} is a mean carrier lifetime of the structure including the emitter recombination. The integration extends from a point deep in the p^+-region ($x = -\infty$) over the base to a point deep in the n^+-layer ($x = \infty$). In the base where the injection level is high, the excess hole concentration Δp is equal to $p = n$, and in the p^+-region the recombination rate $\Delta p/\tau_p$ can be equated to the minority recombination rate $\Delta n/\tau_n(n^+) \approx n/\tau_p$. To realize the importance of the effective lifetime for device characteristics, we use the continuity Eq. (2.103) which in one-dimensional form can be written:

$$-\frac{\partial j_p}{\partial x} = q \cdot \frac{\Delta p}{\tau_p} + q \cdot \frac{\partial \Delta p}{\partial t} \tag{5.43}$$

Since the hole current deep in the p^+-region equals the total current ($j_p(-\infty) = j$), and deep in the n^+-region is zero ($j_p(\infty) = 0$), the integration of (5.43) yields

$$j = q \cdot \int_{-\infty}^{\infty} \frac{\Delta p}{\tau_p} dx + q \cdot \frac{d}{dt} \int_{-\infty}^{\infty} \Delta p \cdot dx; \tag{5.44}$$

Inserting (5.42) and multiplying with the area, one obtains

$$I = \frac{Q}{\tau_{eff}} + \frac{dQ}{dt} \tag{5.45}$$

where I denotes the current and Q the stored charge of excess carriers: $Q \equiv q \cdot A \cdot \int \Delta p \cdot dx$. (5.45) is a generally valid equation of charge dynamics. For a stationary forward current I_F, it takes the form:

$$Q_F = I_F \cdot \tau_{eff} \tag{5.46}$$

where Q_F is the stored charge for this special case. According to (5.46), the effective lifetime can be directly determined by measuring Q_F for a given forward current I_F. Compared with Eq. (5.38a), the lifetime τ_{HL} in the base is replaced in (5.46) by the effective lifetime of the structure. When the current is interrupted, τ_{eff} represents the decay time of the stored charge, since for $I_F = 0$ Eq. (5.45) yields $dQ/dt = -Q/\tau_{eff}$.

Immediately important for the I–V characteristic is that within the widely applicable approximation (5.35), (5.37), the effective lifetime determines the voltage drop across the middle region. By insertion of (5.46) in (5.37) one obtains:

$$V_{drift} = \frac{w_B^2}{(\mu_n + \mu_p) \cdot \tau_{eff}} \tag{5.47}$$

This equation is the generalization of (5.39) for the real case that the emitter efficiency of the junctions is below unity. Equation (5.47) is applicable if $w_B < 4 \cdot \sqrt{D_A \cdot \tau_{HL}}$ (see the discussion to Eq. (5.35)).

We evaluate the effective lifetime now in dependence on device parameters and on the stored charge or the mean concentration $\bar{p}$ in the base region. By splitting up the integration interval on the right hand side of (5.42) into the three neutral regions with constant lifetime one obtains (see Fig. 5.7).

$$\frac{1}{\tau_{eff}} \int_{-\infty}^{\infty} \Delta p \, dx = \frac{1}{\tau_n} \int_{-\infty}^{x_p} n \, dx \; + \; \frac{1}{\tau_{HL}} \int_{L}^{R} p \, dx \; + \; \frac{1}{\tau_p} \int_{x_n}^{\infty} p \, dx \tag{5.48}$$

The equilibrium minority carrier concentrations and likewise the contributions of the space charge layers from x_p to L and from R to x_n are neglected on the right hand side. Since the integrals are proportional to the respective stored charges, (5.48) can be written

$$\frac{Q}{\tau_{eff}} = \frac{Q_n(p^+)}{\tau_n(p^+)} + \frac{Q_B}{\tau_{HL}} + \frac{Q_p(n^+)}{\tau_p(n^+)} \tag{5.49}$$

where Q_B denotes the stored charge in the base, $Q_n(p^+)$, $Q_p(n^+)$ the stored charge of minority carriers in the p^+ and n^+ region, respectively, and $Q = Q_B + Q_n(p^+) + Q_p(n^+)$ the whole stored charge. Because of the low injection in the end regions and the relative small minority carrier diffusion length, the stored charges $Q_n(p^+)$, $Q_p(n^+)$ are small compared with the stored charge $Q_B = q \cdot w_B \cdot \bar{p}$, if the base width is not too small and the injection level not extremely high. Equation (5.49) shows that in spite of the small stored charges $Q_n(p^+)$, $Q_p(n^+)$, the *recombination* in the end regions can be significant if the lifetimes $\tau_n(p^+)$, $\tau_p(n^+)$ are correspondingly smaller than τ_{HL}. The latter can be caused by Auger recombination, a high density of recombination centers in the outer layers (see the discussion to Eq. (3.108)) or by a design of the end regions leading to high surface recombination.

Insertion of the exponential minority carrier distribution in the first and third integral on the right side of (5.48) yields the connection with the emitter parameters

introduced in Sect. 3.4. For the n+-layer, we obtain with (3.43) and (3.104), (3.105) and neglecting the equilibrium density p_{n0}:

$$\frac{1}{\tau_p}\int_{x_n}^{\infty} p dx = \frac{L_p}{\tau_p}(n^+)\cdot p_n^* = \frac{L_p}{\tau_p}\cdot\frac{p_R^2}{n^+}e^{\Delta E_g/kT} = h_n\cdot p_R^2 \tag{5.50}$$

The bandgap narrowing ΔE_g results in an enhancement of the minority carrier concentration p_n^* and hence of the emitter parameter h_n. The analogous equation holds for the recombination integral over the p+-region (first term on the right hand side of (5.48)):

$$\frac{1}{\tau_n}\int_{-\infty}^{L} n\, dx = \frac{L_n(p)}{\tau_n(p)}\cdot n_p^* = \frac{L_n}{\tau_n}\cdot\frac{p_L^2}{p^+}e^{\Delta E_g/kT} = h_p\cdot p_L^2 \tag{5.51}$$

If the integral on the left side in (5.48) is approximated by $w_B\cdot\bar{p}$, neglecting the stored minority carrier charges in the end regions, one obtains from (5.48), (5.50) and (5.51)

$$\frac{1}{\tau_{eff}} = \frac{1}{\tau_{HL}} + h_p\cdot\frac{p_L^2}{w_B\cdot\bar{p}} + h_n\cdot\frac{p_R^2}{w_B\cdot\bar{p}} \tag{5.52}$$

To correlate the mean concentration $\bar{p}$ with the concentrations p_L and p_R at the boundaries, the carrier distribution is used in the form (5.28). One obtains

$$\bar{p} = \frac{1}{w_B}\int_{L}^{R} p\, dx = \frac{L_A}{w_B}\cdot\tanh\left(\frac{d}{L_A}\right)\cdot(p_L + p_R) \tag{5.53}$$

For simpler writing we use again the letter d for $w_B/2$. Using (5.53), Eq. (5.52) can be written

$$\frac{1}{\tau_{eff}} = \frac{1}{\tau_{HL}} + \frac{H}{d}\cdot\left(\frac{d/L_A)}{\tanh(d/(L_A))}\right)^2\cdot\bar{p} \tag{5.54}$$

where

$$H = 2\cdot\frac{\eta^2 h_p + h_n}{(\eta+1)^2} \tag{5.55}$$

with $\eta = p_L/p_R$. The quantity H is to a sufficient approximation often independent of $\bar{p}$. For a symmetrical distribution ($\eta = 1$), H reduces to $(h_n + h_p)/2$. With a typical H-value of 2×10^{-14} cm^4/s, one obtains from (5.54) for a base width $w_B = 200$ µm at $\bar{n} = \bar{p} = 1\times10^{17}$ cm^{-3}:

$$\frac{1}{\tau_{eff}} = \frac{1}{\tau_{HL}} + \frac{1}{5.0\ \mu\text{s}} \left(\frac{d/L_A}{\tanh(d/L_A)} \right)^2 \quad \text{at} \quad \bar{p} = 1 \times 10^{17}\ \text{cm}^{-3}$$

Even if the carrier lifetime τ_{HL} is very high, the effective carrier lifetime remains below 5 µs in this case, the value given by the recombination in the boarder regions for $w_B/L_A \rightarrow 0$.

Measurements of the effective lifetime τ_{eff} and the high-level lifetime τ_{HL} in the base are plotted in Fig. 5.8 as functions of the mean concentration $\bar{n} = \bar{p}$ in the base. $\tau_{eff} = Q_F/I_F$ was determined by measuring the stored charge Q_F, which simultaneously delivers the concentration $\bar{n}$. The base lifetime τ_{HL} was determined from the carrier distribution measured via the recombination radiation profile. While τ_{HL} is found to be nearly independent of the carrier concentration, τ_{eff} decreases nearly by an order of magnitude in the shown range. This is in agreement with the above theory; particularly the variation of τ_{eff} with $\bar{n}$ can be described by (5.52) and (5.54). A point not in agreement with these equations is that τ_{eff} does not tend to the base lifetime τ_{HL} for small $\bar{n}$, but remains considerably smaller. We will come back to this effect at the end of this paragraph.

The boundary concentrations p_L, p_R depend strongly on the parameters h_p, h_n. In some modern fast power diodes, the injection efficiency of the p-emitter is made small while the n^+-region keeps the usual form with a small parameter h_n. This leads to an inversion of the carrier distribution compared with Fig. 5.6 and by means of this to an improved reverse recovery behavior. The influence of the

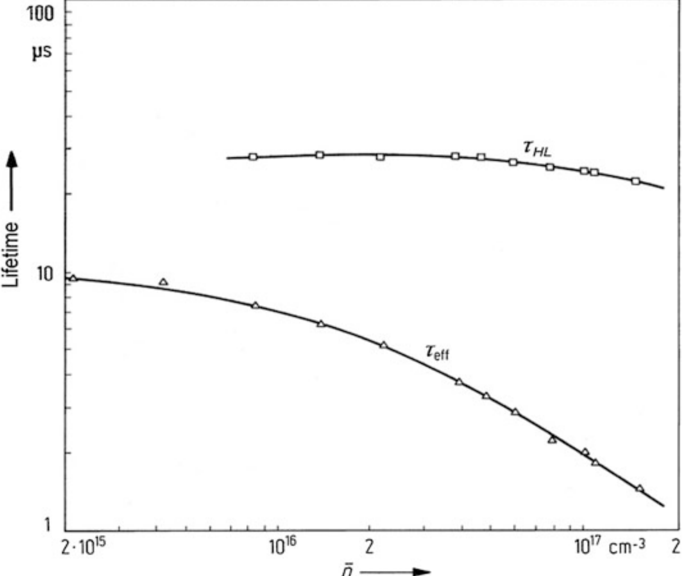


Fig. 5.8 Effective lifetime τ_{eff} of a forward biased pin-diode and lifetime τ_{HL} in the base region as functions of the mean carrier concentration in the base. See text. From [Sco82]

h-values on the carrier distribution follows from the continuity of the electron current at the p$^+$n-junction and the hole current at the nn$^+$-junction. Using (5.21), (5.22) together with (5.28) the boundary conditions can be written:

$$\frac{D_n}{D_n + D_p} j - q \frac{D_A}{L_A} \left(\frac{p_L}{\tanh(w_B/L_A)} - \frac{p_R}{\sinh(w_B/L_A)} \right) = q\, h_p p_L^2 \tag{5.56a}$$

$$\frac{D_p}{D_n + D_p} j + q \frac{D_A}{L_A} \left(\frac{p_L}{\sinh(w_B/L_A)} - \frac{p_R}{\tanh(w_B/L_A)} \right) = q\, h_n p_R^2 \tag{5.56b}$$

Without solving these equations explicitly for p_L and p_R, one can see, that p_L becomes smaller if h_p is enhanced. In the limit of very large emitter recombination (right hand side) compared with the recombination in the base, the ratio $\eta = p_L/p_R$ tends to $(D_n h_n/(D_p h_p))^{1/2}$, as follows by division of (5.56a) and (5.56b). A calculation of η as a function of $\bar{p}$ and the device parameters h_p, h_n, w_B and L_A is given in [Sco69]. The quantity H, which according to (5.54) determines the effective lifetime and hence the voltage drop V_{drift}, varies less than h_p because η decreases with increasing h_p. A measured carrier profile for a 1200-V diode using the mentioned design principle is shown later in Fig. 5.33 in Sect. 5.7.4. In this diode the p-emitter layer had a doping $N_A = 5 \times 10^{16}$ cm^{-3} and a thickness w_p = 2 µm, which according to Eq. (3.106) results in $h_p = 2.6 \times 10^{-12}$ cm^4/s. The n$^+$-region had a doping concentration of about 10^{19} cm^{-3} and an estimated h_n-value of 2×10^{-14} cm^4/s. As shown by the figure, the boundary concentration p_L is about a factor 4 smaller than p_R. However, the recombination term $h_p p_L^2$ in (5.52) is still seven times higher than $h_n p_R^2$.

From (5.46) and (5.54), the current density can be expressed as a function of the concentration $\bar{p}$:

$$\begin{aligned} j &= \frac{Q_F/A}{\tau_{eff}} = \frac{q \cdot w_B \cdot \bar{p}}{\tau_{eff}} \\ &= q \cdot \bar{p} \cdot \left(\frac{w_B}{\tau_{HL}} + 2H \cdot \left(\frac{d/L_A}{\tanh(d/L_A)} \right)^2 \cdot \bar{p} \right) \end{aligned} \tag{5.57}$$

The voltage drop across the base follows by insertion of (5.54) into (5.47)

$$V_{drift} = \frac{w_B}{\mu_n + \mu_p} \cdot \left(\frac{w_B}{\tau_{HL}} + 2H \cdot \left(\frac{d/L_A}{\tanh(d/L_A)} \right)^2 \cdot \bar{p} \right) \tag{5.58}$$

The junction voltage (5.32) can be written using (5.53):

$$V_j = 2\, \frac{kT}{q} \ln \left(\frac{2\sqrt{\eta}}{1+\eta} \cdot \frac{d/L_A}{\tanh(d/L_A)} \cdot \frac{\bar{p}}{n_i} \right) \tag{5.59}$$

where as before $\eta = p_L/p_R$. With (5.57) to (5.59) the current-voltage characteristic is given in a parameter representation $j(\bar{p})$, $V_F(\bar{p}) = V_j(\bar{p}) + V_{drift}(\bar{p})$. This form of the characteristic has the advantage, that also the mobilities and ambipolar diffusion constant as well as the lifetime in L_A can be inserted as functions of $\bar{p}$ to consider carrier-carrier scattering and, if necessary, Auger recombination.

Forward characteristics calculated with these equations for different values of h_p, h_n are shown in Fig. 5.9 (solid curves). The dimensions of the base region are suitable for a 3 kV-diode. Carrier-carrier scattering is taken into account as described in Chap. 2. The neglect in the step from (5.34) to (5.35) is abandoned using the inhomogeneity factor given by the later Eq. (5.63). In the case $h_n = h_p = 2 \times 10^{-14}$ cm^4/s a realistic characteristic for diodes with these dimensions is obtained. For $h_n = h_p = 0$ (Hall case), the concentration $\bar{p}$ runs so high at high current densities that Auger recombination in the base becomes very significant. Within the above model, Auger recombination is considered replacing $1/\tau_{HL}$ by $1/\tau_{HL} + (c_{A,n} + c_{A,p}) \cdot \bar{p}^2$ [see Eq. (2.59)]. To test the accuracy of the model, also characteristics calculated numerically are plotted in Fig. 5.9 (dotted lines). They were obtained by determining the carrier distribution for each current density from the differential Eq. (5.25) using a lifetime including Auger recombination and variable mobilities in (5.21); the mobility sum in (5.34) was pulled under the integral. As is seen, the analytical approach is fairly accurate. Above the current range of the figure and in the Hall case, the analytical model is no longer suited. However, the Hall case is anyway hypothetical.

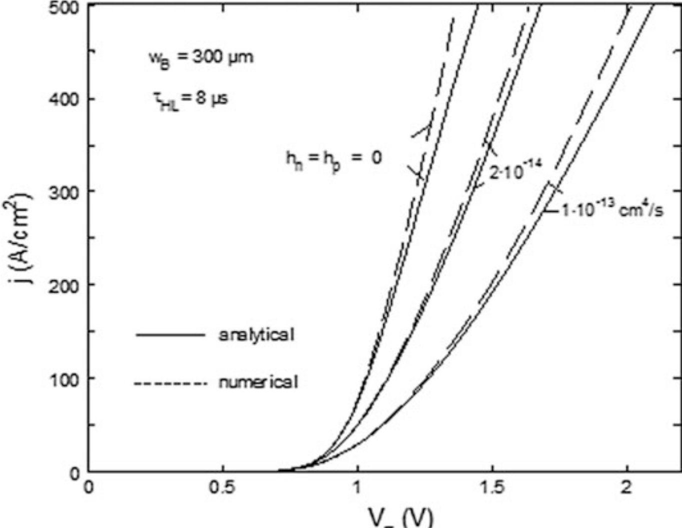


Fig. 5.9 Current voltage characteristics calculated with different emitter parameters

Sometimes the direct relationship between forward voltage and current density is desirable. To derive it, Eq. (5.57) is resolved first for $\bar{p}$, which yields

$$\frac{1}{\bar{p}} = \frac{q\, w_B}{2\, j\, \tau_{HL}} \left(1 + \sqrt{1 + \frac{2\, \tau_{HL}\, H}{q\, D_A\, \tanh^2(d/L_A)} \cdot j} \right) \tag{5.60}$$

Inserting this into (5.35) one obtains for the voltage drop across the middle region:

$$V_{drift} = \frac{w_B^2}{2\, (\mu_n + \mu_p)\, \tau_{HL}} \left(1 + \sqrt{1 + \frac{2\, \tau_{HL}\, H}{q\, D_A\, \tanh^2(d/L_A)} \cdot j} \right) \tag{5.61}$$

The junction voltage as a function of current density is obtained by insertion of (5.60) into (5.59). Hence the forward voltage $V_F = V_j + V_{drift}$ is given now as a function of the current density j.

Contrarily to the Hall approximation (5.39), the voltage drop V_{drift} depends now explicitly on the current density. For a sufficiently small emitter quantity H or small current density j, (5.61) reduces to (5.39). Since H cannot be made smaller than about 1×10^{-14} cm^4/s, however, the current dependent term in (5.61) becomes soon significant. The reference current density

$$j_0 \equiv \frac{q\, D_A\, \tanh^2(d/L_A)}{2\, \tau_{HL}\, H}$$

around which the current density j comes into play, is fairly small. If for example $H = 2 \times 10^{-14}$ cm^4/s, $d = 100$ µm, $\tau_{HL} = 4$ µs, $D_A = 15$ cm^2/s, j_0 amounts only to 11.1 A/cm^2. If H is made large for above mentioned reasons, j_0 will be still smaller. For $j \gg j_0$, (5.61) turns into

$$V_{drift} = \frac{w_B^2}{(\mu_n + \mu_p)\, L_A\, \tanh(d/L_A)\, \sqrt{2}} \cdot \sqrt{H \cdot j/q} \tag{5.62}$$

V_{drift} is here proportional to the square root of the current density, as far as the variation of the mobilities can be neglected. $\bar{p}$ is proportional to $\sqrt{j}$ in this case according to (5.60), hence $j/\bar{p}$ in (5.35) is also proportional to $\sqrt{j}$. This relationship is in accordance with measurements, which often yield a dependency $V_F - V_j \propto \sqrt{j}$ for power diodes.

Figure 5.10 shows the measurement of the forward characteristics of a fast 600-V diode and the result of a calculation with Eqs. (5.59)–(5.61). For the dependency of μ_n, μ_p, D_A on the carrier concentration, the Equations given in Sect. 2.6 and Appendix A are used. The calculation agrees with the measurement within a wide range of the current density.

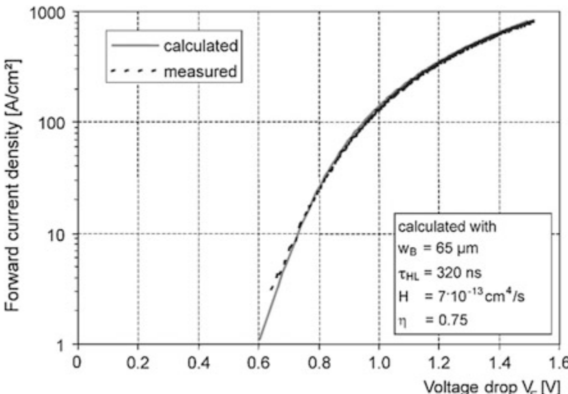


Fig. 5.10 Measured forward characteristics of a fast 600-V diode compared to calculation results with consideration of recombination in the emitter regions

If the carrier distribution is very inhomogeneous, i.e. very unsymmetrical or if it is characterized by a $d/L_A > 2$, the approximations (5.35) or (5.47), underlying our calculation, underestimates the voltage drop V_{drift} considerably. Carrying out the integral in (5.34) with the exact carrier distribution $p(x') = \cosh((x' - \Delta)/L_A)$ results in a voltage drop V_{drift} which has been given for the Hall case by (5.41). According to this equation, V_{drift} is by the factor

$$f = \left(\frac{L_A}{d}\right)^2 \cdot \sinh\left(\frac{d}{L_A}\right) \cdot \cosh\left(\frac{\Delta}{L_A}\right) \cdot arctg\left(\frac{\sinh(d/L_A)}{\cosh(\Delta/L_A)}\right) \tag{5.63}$$

higher than given by the approximation (5.40) (Δ denotes the distance of the minimum of $p(x)$ from the middle of the base). Since the carrier distribution has the same general *cosh*-form also if emitter recombination is present, the inhomogeneity factor (5.63) applies generally and has to be added correctly to the V_{drift} expressions (5.47), (5.58), (5.61) and (5.62). In Figs. 5.9 and 5.10 the factor has been taken into account.

The above equations can be used also for IGBTs to describe the I–V-characteristics in the saturated current range. The conduction behavior of several types of IGBTs is mainly determined by the p-emitter, since very shallow and low doped emitter regions are implemented. An IGBT-characteristic will be shown later in Fig. 10.2. In the range above the rated current one finds often a nearly linear or resistive characteristic.

In Fig. 5.8 it is observed that τ_{eff} for small $\bar{n}$, in the range of 10^{15} cm^{-3}, is nearly constant like the lifetime τ_{HL} in the base, but is essentially smaller than τ_{HL}. This result has been found for all samples till now. It has led to the conclusion that there is a linear recombination part entering the stored charge, which is not located in the volume of the base but must be located in the emitters or thin boundary layers of the base not resolved by the radiation profiles [Sco79, Coo83]. This recombination part called ‘linear emitter recombination’ has been mentioned already in Sect. 3.4 and is expressed by Eq. (3.112). In the later Fig. 5.33 in Sect. 5.7.4, the phenomenon appears again: The lifetime in the volume of the base is higher than given by Q_F/I_F

at small $\bar{p}$ where the quadratic emitter recombination is negligible. The sagging of the concentration towards the middle of the base is therefore smaller. For the incorporation of this effect into the theory of the I–V-characteristic we refer to the literature [Sco79].

Above calculations can be used to implement the parabolic shape of the forward characteristics of modern fast diodes in a circuit simulator. Such simulators need simplified models for the forward characteristics of diodes, and this description by a parabola is in most cases much closer to reality than the description with the Hall approximation, or the often used oversimplified description with a straight line.

5.4.6 Temperature Dependency of the Forward Characteristics

For low current, the forward voltage decreases with temperature, because the part of the voltage that drops at the pn-junction, i.e. the term $V_j(\mathrm{p}^+\mathrm{n})$ given in Eq. (5.31), decreases with increasing n_i^2. The diode at low-current condition can thus be used as a temperature sensor (see Fig. 3.12 and Fig. 11.19 in Chap. 11). At increased current density, the temperature dependency of V_{drift} predominates and here, opposing effects have to be considered:

- The mobilities decrease with temperature, (see Fig. 2.14 and Appendix A1). According to (5.47), this leads to an increase in V_{drift}.
- The carrier lifetime increases with increasing temperature which, in turn, leads to a decrease in V_{drift}.

Since both effects are opposing, the resulting behavior depends on the specific technology, especially of the temperature dependency of the effect of the used recombination centers.

By using radiation-induced recombination centers, a curve as is depicted in Fig. 5.11 on the right-hand side is measured. The lines for the characteristics at 25 and 150 °C intersect at 150–200 A/cm^2, a current density which is typical for rated current of a 1200 V fast diode.

Even though an intersection point is likewise found for diodes without recombination centers (rectifier diodes for grid frequency) and for diodes using gold as recombination center, this intersection occurs for those devices typically at a 3-times higher current density.

By using platinum combined with a not very weakly doped p-emitter, no intersection point is found in the relevant current range. An example is given in Fig. 5.11 (left). In this case, the forward voltage strongly decreases with temperature. This is advantageous in terms of conduction losses, but this temperature behavior is very unfavorable for parallel connection of diodes. As the result of

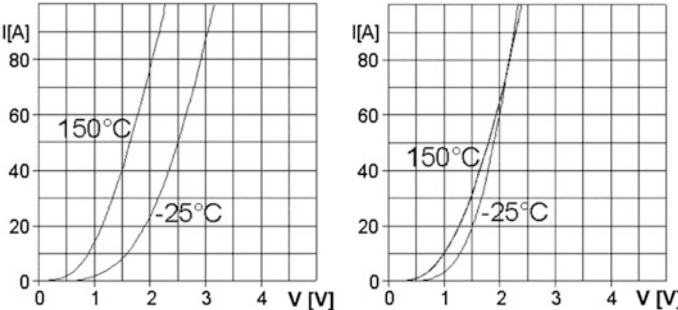


Fig. 5.11 Forward characteristics of fast 1200 V diodes and its temperature dependency. Left: Platinum-diffused diode. Right: Diode with radiation-induced recombination centers (CAL diode). Active area 0.32 cm^2

tolerances in the manufacturing process, there is always some variation of the forward voltage for different samples. In case of parallel connection, the diode with the lower voltage drop will attract more current. Consequently, it will generate higher conduction losses, its temperature will increase. That will result in a further reduction of the forward voltage, so that this diode will attract a further increased fraction of the current, and so on. A pronounced negative temperature dependency with a V_F(T)-decrease of more than 2 mV/K endangers thermal instability in parallel connection. A diode like that in Fig. 5.11 (right) is, on the other hand, well suited for parallel connection.

Diodes connected in parallel are coupled thermally:

- via the substrate for paralleling of several diodes in a module,
- via the heat sink with paralleling of modules.

In the case of a weakly negative temperature coefficient, this coupling is usually sufficient to avoid a thermal runaway of the diode with the lowest forward voltage. In the case of diodes with a negative temperature coefficient of <-2 mV/K, it is recommended to decrease the current load in parallel connection to a lower value as would result from the sum of the current ratings of the single diodes. This measure is known as “derating”.

5.5 Relation Between Stored Charge and Forward Voltage

Especially with fast diodes, a trade-off between the demands for fast switching—low stored charge etc.—and for low forward voltage has to be made. If the carrier lifetime is adjusted to be low, then the forward voltage increases according to Eq. (5.61). According to (5.37), the voltage V_{drift} and the stored charge Q_F are related by the equation

$$Q_F = \frac{w_B^2 \cdot I_F}{V_{drift}\left(\mu_n + \mu_p\right)} \tag{5.64}$$

The forward voltage has been partitioned into the parts V_{drift} in the base region and the voltage drops $V_j(p^+n)$ at the pn- and $V_j(nn^+)$ at the nn^+-junction (see Eq. 5.29). These terms are combined in Eq. (5.32) to the junction voltage V_j, $V_F = V_j + V_{drift}$. Replacing V_{drift}, (5.64) results in

$$Q_F = \frac{w_B^2 \cdot I_F}{\left(V_F - V_j\right)\left(\mu_n + \mu_p\right)} \tag{5.65}$$

Equation (5.65), like Eq. (5.35), is a good approximation for $w_B/L_A < 4$. This hyperbolic relation is shown in Fig. 5.11. In this figure, Q_F according to (5.65) is drawn for a fast 600 V diode with w_B = 65 µm. For comparison, the experimental results for Q_{RR} of a fast diode with this dimensioning are also shown. Q_{RR} distinguishes from Q_F by the amount of charge recombining during the measurement duration.

A hyperbolic relation as shown in Fig. 5.12 results for every technology; how far it can be shifted to lower values, however, is an evaluation criterion for the specific design. The base-width w_B contributes squared to Q_F as well as to V_F. Therefore, w_B must be kept as low as possible by taking into account all requirements to the diode.

Nevertheless, up to now nothing has been stated about the time-dependent waveform in which the stored charge emerges during the reverse-recovery procedure. This, however, is most significant and is to be discussed in the following paragraphs regarding the turn-off behavior. However, first the turn-on behavior will be investigated.

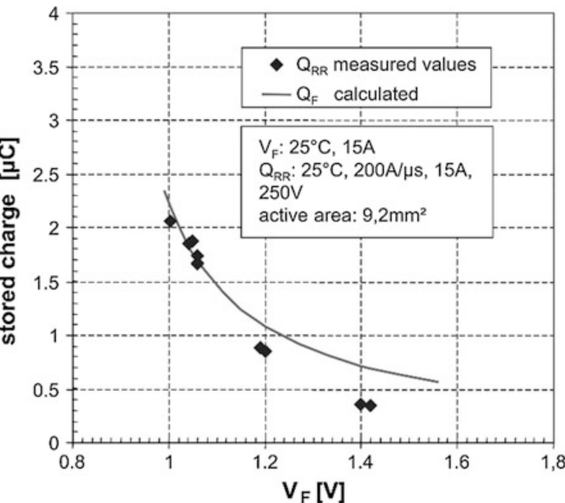


Fig. 5.12 Relation of stored charge to forward voltage for a fast 600-V diode

5.6 Turn-on Behavior of Power Diodes

At the transition of the diode into the conducting state, the voltage first increases to the turn-on voltage peak V_{FRM} (Forward-Recovery Maximum) before it drops down to the forward voltage. Figure 5.13 shows the definition of V_{FRM} and the turn-on time t_{fr}, in which t_{fr} is defined as the time interval between the instant of 10% forward voltage and the instant at which the voltage has dropped down again to the 1.1-fold value of the steady state forward voltage.

This old definition comes from a time in which thyristors were the dominating devices in power electronics; low current slopes were usual and V_{FRM} amounted to several volts. This definition does not pertain to freewheeling diodes and snubber diodes used in circuits with IGBTs as switching elements, because with them such steep slopes *di*/*dt* of the current occur that V_{FRM} may reach – for example in a poorly designed 1700 V diode – a value of 200–300 V which is more than 100-times the value of V_F. Reading out the time at the value of $1.1 \times V_F$ is no longer possible in the measurement.

A low V_{FRM} is one of the most important requirements for diodes in snubber- and clamping circuits, since these circuits work just after the diode has turned on.

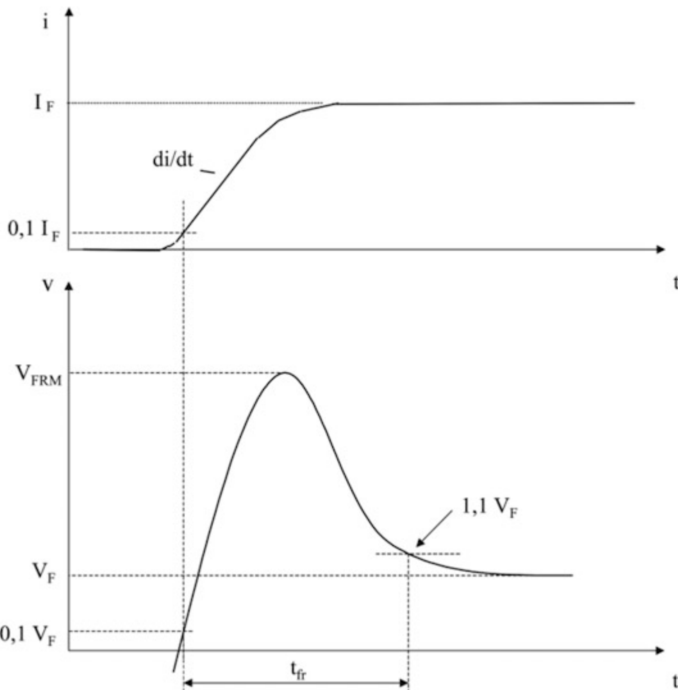


Fig. 5.13 Characteristic parameters of the turn-on behavior of power diodes

The forward-recovery voltage peak is an important feature also for freewheeling diodes designed for a blocking voltage of 1200 V and more. At turn-off of the IGBT, the freewheeling diode turns on; the *di/dt* at the IGBT turn-off creates at the parasitic inductance a voltage peak on which V_{FRM} is superimposed. The sum of both components can lead to a critical voltage peak.

The measurement of this behavior is not trivial, since the inductive component and V_{FRM} cannot be distinguished in an application-conform chopper circuit. The measurement is possible only at an open setup directly at the bond wires of the diode. Such a measurement is depicted in Fig. 5.14. Here, the turn-on of two diodes is shown, one of which (standard diode) is designed with a very wide w_B in order to achieve a soft recovery behavior at turn-off. For the CAL-diode presented in comparison, w_B is kept as low as possible. Under the same measurement conditions, i.e. the same parameters for controlling the IGBT turn-off, a V_{FRM} of 84 and 224 V results for the CAL-diode and the standard diode, respectively. It is possible to analyze the worst-case scenario. At a step function current form ($di/dt = \infty$), the maximally occurring voltage corresponds to the resistance of the base without carrier injection, multiplied by the current density:

$$V_{FRM} = \frac{w_B \cdot j}{q \cdot \mu_n \cdot N_D} \tag{5.66}$$

This equation can be used as long as $j < q \cdot N_D \cdot v_{sat}$. The design of a diode for higher blocking voltage requires a lower doping N_D and a wider base-width w_B, which increases the resulting voltage peak strongly. Figure 5.15 shows the results according to Eq. (5.66), whereby N_D and w_B were selected for fast diodes of different voltage ranges, and the mobility μ_n for $T = 400$ K from Fig. 2.12 was used.

For a diode designed for 600 V, a voltage peak amounting to only some 10 V can occur. For a diode designed for 1700 V, however, this peak may amount to more than 200 V, and for a diode designed for the voltage range of > 3000 V, more than 1000 V are possible. Moreover, the effects of recombination centers have not yet been considered in this calculation. It is known that the recombination center

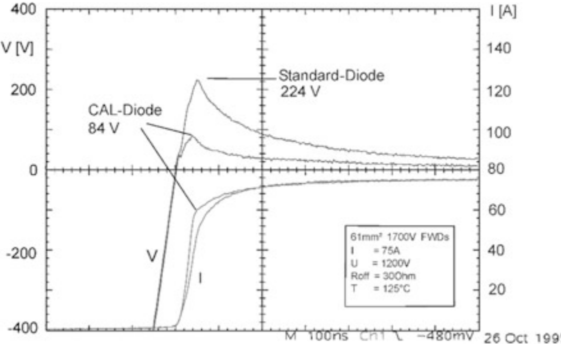


Fig. 5.14 Turn on of two diodes with different width w_B of the low-doped layer

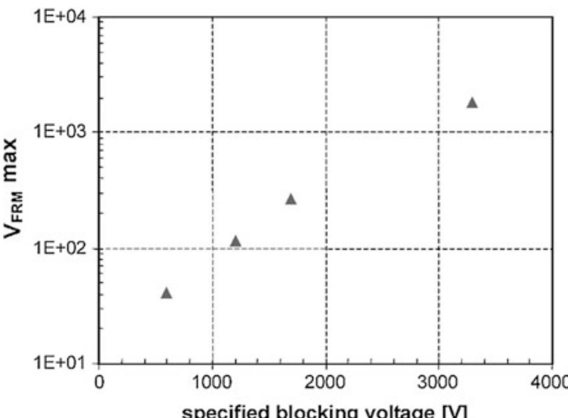


Fig. 5.15 Calculated worst-case voltage peak at turn-on of diodes at T = 125 °C. Compensation effects are not considered

gold has a compensating effect. In a low-doped n-region, the density of these acceptors compensates a part of the background doping density. The resulting decreased effective doping has to be applied in (5.66), and the voltage peak can then become significantly higher. In practice, there are no current slopes occurring with the form of a step-function, however, current slopes with di/dt in the range of > 2000 A/µs are to be expected in IGBT applications as shown in Fig. 5.14.

The importance of the turn-on behavior of freewheeling diodes was underestimated for a long time. Just after high-voltage IGBTs of > 3000 V were introduced and failures in the application occurred, the anti-parallel diode to the IGBT was found as a reason: if it creates a high forward-recovery peak V_{FRM}, this voltage peak is applied to the IGBT in the reverse direction. The reverse blocking capability is not specified for common IGBTs, since the collector-side pn-junction has no defined junction termination. To counteract this problem, more attention was paid to the turn-on of the freewheeling diodes.

Regarding switching losses, the turn-on behavior of the diode is not significant. Even if high voltage peaks occur, the turn-on process is very fast, and turn-on losses amount to only some percent of the turn-off losses or of the conduction losses of the diode. For thermal calculations, turn-on losses can be neglected in very most cases.

5.7 Reverse-Recovery of Power Diodes

5.7.1 Definitions

With the transition from the conducting to the blocking state, the charge stored in a diode has to be removed. This charge causes a current flow in the reverse direction of the diode. The reverse-recovery behavior signifies the time-dependent waveforms of this current and the corresponding voltage.

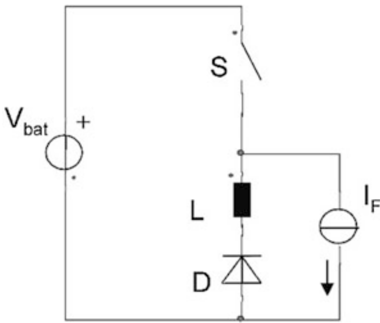


Fig. 5.16 Circuit for characterizing reverse-recovery behavior

The simplest circuit to measure this effect is the circuit according to Fig. 5.16, whereby S represents an ideal switch, I_F an ideal current source, V_{bat} an ideal voltage source, L an inductor and D the diode being considered. After closing the switch S, the progression of current and voltage as shown in Fig. 5.17 occurs at a the diode. Figure 5.17 exemplifies a diode with soft recovery behavior, whereas Fig. 5.18 depicts two examples of the current waveform of diodes with a snappy reverse-recovery behavior.

First, the definitions will be explained using the circuit in Fig. 5.16 and the waveform in Fig. 5.17. In the circuit of Fig. 5.16, after closing the switch S it holds

$$L \cdot \frac{di}{dt} + v(t) = -V_{bat} \tag{5.67}$$

where $v(t)$ is the time dependent voltage at the diode. First, during the current decay, the voltage at the diode is in the range of the forward voltage V_F which is in

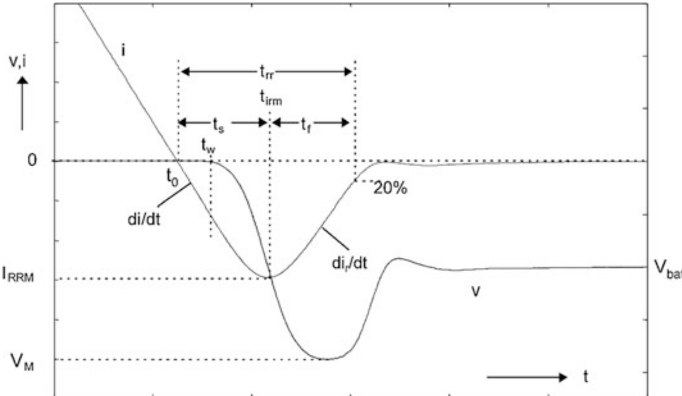


Fig. 5.17 Waveforms of current and voltage for a soft recovery diode during the reverse-recovery process in a circuit according to Fig. 5.16 and definitions of some characteristic values of the recovery behavior

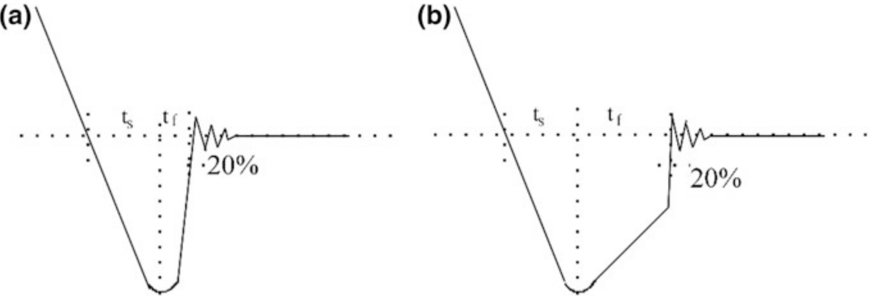


Fig. 5.18 Current waveform for two different possibilities of snappy reverse-recovery behavior

the range of 1 – 2 V, and $v(t)$ can be neglected. The current slope at commutation is determined by the voltage and inductance:

$$-\frac{di}{dt} = \frac{V_{bat}}{L} \tag{5.68}$$

The zero crossing point of the current occurs at t_0. At t_w the diode starts to take over the voltage; at this instant the pn-junction of the diode is free of charge carriers. At the same point in time, the current deviates from the linear slope. At t_{irm} the reverse current attains its maximum I_{RRM}. At t_{irm}, it holds that $di/dt = 0$ and from Eq. (5.67), $v(t) = -V_{bat}$ results.

After t_{irm} the reverse current decays down to the level of the static leakage current. The shape during this interval solely depends on the diode. If this decay is steep, a snappy reverse-recovery behavior is given. If this decay occurs slowly, however, a soft recovery behavior is indicated. This slope di_r/dt, which is often not linear, leads to an induced voltage $L \cdot di_r/dt$ that adds on the battery voltage.

The switching time t_{rr} is defined as the time between t_0 and the point in time at which the current has decayed down to the value of 20% of I_{RRM}. With the subdivision of t_{rr} in t_f and t_s as is shown in Fig. 5.17, formerly the following "soft factor" s was defined as the quantitative parameter of the reverse-recovery behavior:

$$s = \frac{t_f}{t_s} \tag{5.69}$$

whereby e.g. $s > 0.8$ signified that a diode can be called to be "soft".

This definition, however, is very insufficient. According to it, a current shape as in Fig. 5.18a would be snappy, but a current shape as in Fig. 5.18b would be accepted as soft. While in Fig. 5.18b $t_f > t_s$ is given and $s > 1$ according to (5.69), a very steep slope, a reverse current snap-off, occurs in a part of the reverse-recovery waveform.

The following definition of the soft factor is better:

$$s = \left| \frac{-\frac{di}{dt}\Big|_{i=0}}{\left(\frac{di_r}{dt}\right)_{\max}} \right| \tag{5.70}$$

The applied current slope must be measured at the zero crossing point, and the di_r/dt caused by the diode is measured at its maximal value. The measurement must be executed at less than 10% and at 200% of the specified rated current. For soft recovery the value of $s > 0.8$ is again required. With this definition, also a behavior like that shown in Fig. 5.18b is considered to be snappy. Additionally, this definition includes the observation that small currents are especially critical for reverse-recovery behavior.

The term di_r/dt determines the occurring voltage peak, and $v(t)$ in (5.67) has the maximal amplitude at the maximal slope:

$$V_M = -V_{bat} - L \cdot \left(\frac{di_r}{dt}\right)_{\max} \tag{5.71}$$

Thus, the voltage peak V_M occurring under special conditions or the induced voltage $V_{ind} = V_m - V_{bat}$ can be used as a quantitative definition for the reverse-recovery behavior. As conditions, V_{bat} and the applied di/dt must be indicated.

This definition, however, is also insufficient, because even more parameters have an influence on the reverse-recovery behavior:

1. The temperature: in most cases, high temperatures are more critical for the reverse-recovery behavior. For some fast diodes, however, room temperature or a lower temperature is a more critical condition for possible occurrence of snappy recovery behavior.
2. The applied voltage V_{bat}: higher voltage leads to worse recovery behavior.
3. The value of the inductor L: according to (5.71), with increased L the voltage at the diode is increased; this makes the conditions for the diode harder.
4. The commutation velocity di/dt: A rise in di/dt leads to greater danger of oscillations and snap-off of the current. The reverse-recovery behavior has an increased tendency towards snappy behavior.

All these different influences cannot be covered by a simple quantitative definition. The circuit according to Fig. 5.16 and the definitions according to Eqs. (5.69) and (5.70) can only be used to show the effect of different design parameters. In fact, the reverse-recovery behavior has to be evaluated by using the waveforms of current and voltage, which are measured under application-conform conditions.

The application-conform double-pulse measurement circuit is shown in Fig. 5.19. Compared to the circuit in Fig. 5.16, the ideal switch is replaced by a real switch, e.g. an IGBT. The ideal current source is replaced by an ohmic-inductive load consisting of R and L. The commutation velocity is given by the transistor; for

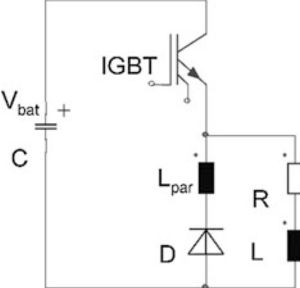


Fig. 5.19 Application-conform double-pulse circuit for measuring the reverse-recovery behavior

an IGBT it is adjustable by the resistor R_{on} in the gate circuit as described later in Chap. 10. V_{bat} is the battery voltage, which is assisted by a capacitor C. The wiring between capacitor, IGBT, and the diode together form a parasitic inductance.

In Fig. 5.20 the drive signals for the IGBT, the current in the IGBT and the current in the diode are shown for the double-pulse mode. By turning off the IGBT, the load current is transferred to the freewheeling diode. At the next turn-on of the IGBT, the diode is commutated – the point in time when the characteristic reverse-recovery of the diode occurs. At turn-on, the IGBT has to conduct additionally the reverse current of the freewheeling diode.

This reverse-recovery event is shown at higher time resolution in Fig. 5.21 for a soft recovery diode. Figure 5.21a presents the current- and voltage waveform in the IGBT and the resulting power loss $v(t)\cdot i(t)$ at the turn-on. In addition, Fig. 5.21b

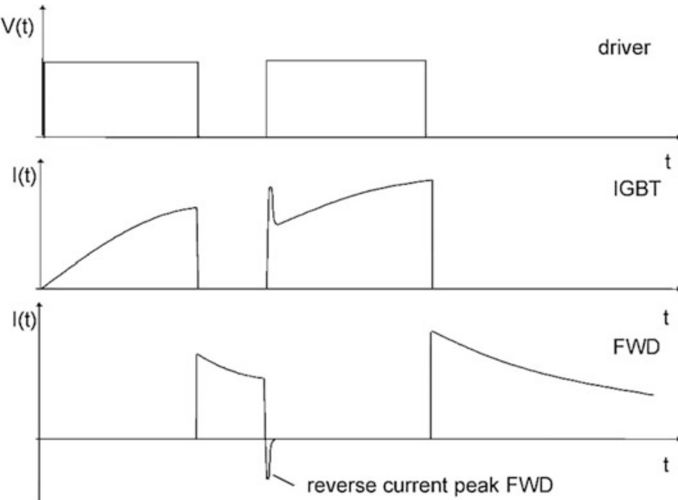


Fig. 5.20 Signal of the driver, current in the IGBT and in the freewheeling diode (FWD) at double-pulse measurement with the circuit according to Fig. 5.19

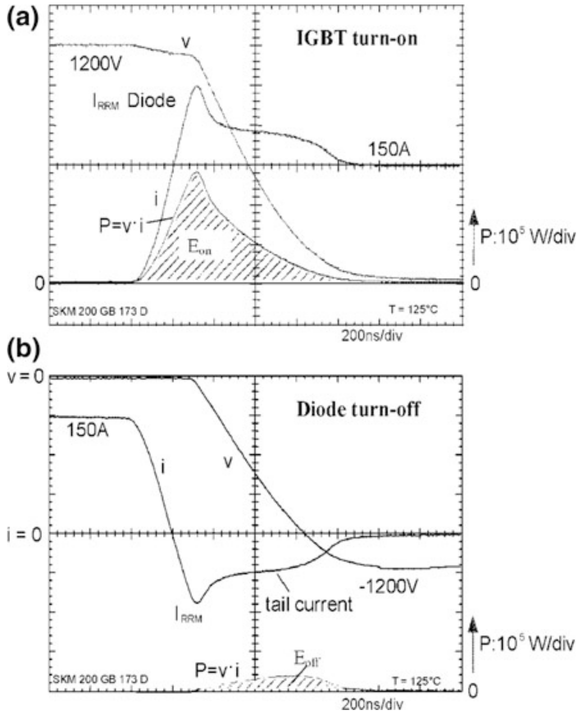


Fig. 5.21 Current waveform, voltage waveform and power losses at turn-on of the IGBT (**a**) and simultaneous turn-off of the diode (**b**) during the measurement of the diode recovery behavior in a double-pulse circuit according to Fig. 5.19

displays the current- and voltage waveforms for the freewheeling diode as well as the power losses in the diode.

While the IGBT has to conduct the reverse-recovery current maximum I_{RRM} of the freewheeling diode additionally to the load current, the voltage at the IGBT is still in the range of the battery voltage V_{bat} (1200 V in Fig. 5.21a). In this instant the maximum of the turn-on loss in the IGBT occurs.

The reverse-recovery current waveform of the diode can be divided into two phases:

1. The waveform until I_{RRM} and the subsequent decay of the reverse current with di_r/dt. In a soft recovery diode, $|di_r/dt|$ is in the range of $|di/dt|$. The reverse current peak I_{RRM} causes the most stress for the switching device.
2. The tail current phase during which the reverse current slowly phases out. A useful definition of the switching time t_{rr} is hardly possible for such a waveform. The tail current phase causes the main losses in the diode, since now there is a high voltage across the diode. Even though a snappy diode without tail current would feature less switching losses in the diode, it is detrimental for the application due to generated voltage peaks and oscillations. Slow and soft waveforms are desired. For the IGBT the tail current phase of the diode causes less stress, since in this phase the voltage at the IGBT has already decayed down to a low value.

The diode switching losses in Fig. 5.21b are represented on the same scale as those for the IGBT in Fig. 5.21a; the diode losses in the application are small compared to the losses in the IGBT. With regard to the total losses in the interaction of both devices, it is therefore important to keep the reverse-recovery current peak I_{RRM} low, and to take care that the main part of the stored charge of the diode is extracted in the tail phase. As the tail current causes the main part of the switching losses in the diode, it must also be limited. Typically, the switching losses in the diode are lower than those in the transistor (compare Fig. 5.21a, b). Regarding its contribution to the total losses, the most important characteristic for the diode is the reverse-recovery current peak I_{RRM} that must be as low as possible.

For a typical application in a current range of 100 A, in which the semiconductor devices of the chopper circuit are packaged inside of a single module, the parasitic inductance L_{par} is in the range of 40 nH or below. This leads to no significant overvoltage. Since there is no longer an ideal switch but rather a real transistor, the transistor still takes a part of the voltage during the reverse-recovery phase of the diode and lowers the applied battery voltage by $v_C(t)$. After turn-on of the transistor, (5.67) is now valid in a modified form:

$$L_{par} \cdot \frac{di}{dt} + v(t) = -V_{bat} + v_C(t) \tag{5.72}$$

The voltage occurring at the diode after I_{RRM} is now

$$v(t) = -V_{bat} - L_{par} \cdot \frac{di_r}{dt} + v_C(t) \tag{5.73}$$

where $v_C(t)$ is the voltage drop across the transistor in this phase. For soft recovery diodes it is typical that at moderate commutation velocities di/dt of up to 1500 A/µs and minimized parasitic inductance, the absolute value of the voltage at the diode $v(t)$ is smaller than V_{bat} and no voltage peak occurs.

As long as the circuit according to Fig. 5.19 is applied and the parasitic inductance is kept low, one can use the following definition:

A diode exhibits a soft recovery behavior if, under all relevant application conditions in an application conform circuit, no overvoltage is caused at the diode by a reverse-recovery current snap-off.

The relevant conditions cover the whole current range, all commutation velocities which can occur in the application and the temperature range from −50 °C up to 150 °C.

This definition is valid provided that there are not excessively high commutation velocities (> 6 kA/µs) or high parasitic inductances (> 50 nH) in the circuit. If L_{par} is increased and the switching characteristics of the IGBT approaches the characteristics of an ideal switch, which means that $v_C(t)$ approaches zero, then the circuit in Fig. 5.19 approaches the circuit in Fig. 5.16. In this case, voltages peaks are unavoidable also for soft recovery diodes.

In high-power modules with a rated current of 1200 A and more, 24 or more IGBT chips are connected in parallel. Such modules exhibit a large volume and it is very difficult to achieve a low parasitic inductance when connecting these modules to a three-phase inverter. In this range it makes sense to investigate the diode with an increased parasitic inductance. Thus, it should be investigated whether there are conditions at which a reverse current snap-off occurs. In such a reverse current snap-off, the waveform usually is similar to the waveform shown in Fig. 5.18b. The reverse current snap-off may even occur relatively late in the reverse recovery waveform, at the end of the tail current.

5.7.2 Reverse-Recovery Related Power Losses

The turn-off energy of the diode per switching event is generally (see Fig. 5.21b):

$$E_{off} = \int_{t_s + t_f} v(t) \cdot i(t) dt \tag{5.74}$$

A simplified estimation can be given for two cases. The first is the case of the circuit according to Fig. 5.16 and the waveform in Fig. 5.17. This waveform is drawn in a simplified way in Fig. 5.22.

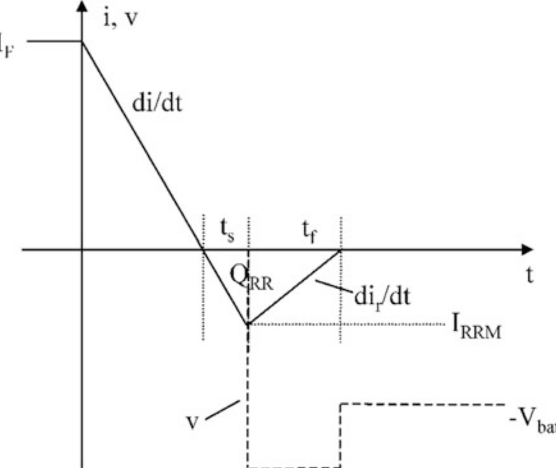


Fig. 5.22 Simplified waveform of current and voltage at the diode during turn-off in the circuit according to Fig. 5.16

During the time until current zero-crossing and during the time t_s, the voltage is simply assumed to be $v = 0$. For $t > t_s$ the diode takes over the voltage. If a linear decay of i_r during t_f is assumed, it holds during t_f that

$$i_r(t) = -I_{RRM} + \frac{I_{RRM}}{t_f} \cdot t \tag{5.75}$$

$$v = -V_{bat} - L \cdot \frac{di_r}{dt} = -V_{bat} - L \cdot \frac{I_{RRM}}{t_f} = const. \tag{5.76}$$

Thus, it follows:

$$E_{off} = \frac{1}{2} \cdot L \cdot I_{RRM}^2 + \frac{1}{2} \cdot V_{bat} \cdot I_{RRM} \cdot t_f \tag{5.77}$$

The first term on the right-hand side of (5.77) can be modified using Eq. (5.67). Accordingly, it holds that $L = -\frac{V_{bat}}{di/dt} = \frac{V_{bat}}{I_{RRM}/t_s}$. With this, one obtains from (5.77):

$$\begin{aligned} E_{off} &= \frac{1}{2} \cdot V_{bat} \cdot I_{RRM} \cdot t_s + \frac{1}{2} \cdot V_{bat} \cdot I_{RRM} \cdot t_f \\ &= \frac{1}{2} \cdot I_{RRM} \cdot t_{rr} \cdot V_{bat} = Q_{RR} \cdot V_{bat} \end{aligned} \tag{5.78}$$

The diode turn-off losses are thus directly proportional to Q_{RR}.

This simplified consideration is valid only for the circuit in Fig. 5.16 in which main effects are determined by the inductor L. Also for the second case, the application-conform circuit according to Fig. 5.19 with the waveforms according to Fig. 5.21, a simplified estimation of the switching losses is possible. The waveforms from Fig. 5.21 are illustrated in a simplified way in Fig. 5.23. Here, the parasitic inductance L_{par} is neglected, and the voltage slope at the diode is only determined by the voltage slope at the transistor $v_C(t)$. The voltage fall time of the transistor t_{fv} is assumed to be equal to the reverse current fall time t_f of the diode. The current- and voltage waveforms are again idealized by straight lines.

The reverse-recovery charge Q_{RR} of the diode is subdivided into the charge Q_{RS} occurring during the storage time t_s and the charge Q_{RF} occurring during the reverse current fall time t_f. It holds that $Q_{RR} = Q_{RS} + Q_{RF}$. The turn-off energy loss in the diode per switching event is then

$$E_{off} = \frac{1}{3} Q_{RF} \cdot V_{bat} \tag{5.79}$$

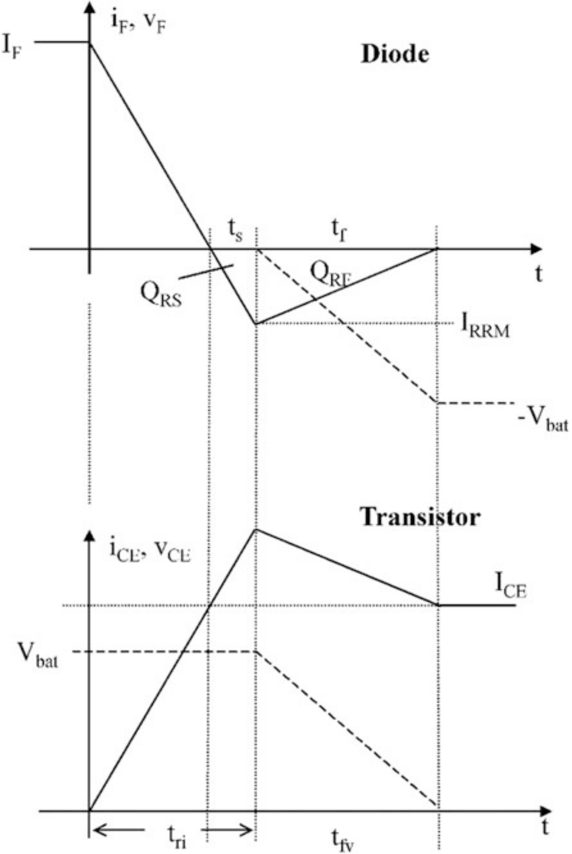


Fig. 5.23 Simplified waveforms of current and voltage at the diode and in the transistor for the circuit according to Fig. 5.19

The parameters Q_{RR} and I_{RRM} at specified conditions di/dt und V_{bat} are given in the data sheets of well specified, modern freewheeling diodes. From I_{RRM} and di_F/dt, one can calculate Q_{RS} as

$$Q_{RS} = \frac{1}{2} t_s \cdot I_{RRM} = \frac{1}{2} \cdot \frac{I_{RRM}}{di_F/dt} \cdot I_{RRM} = \frac{1}{2} \cdot \frac{I_{RRM}^2}{di/dt} \tag{5.80}$$

and because of $Q_{RF} = Q_{RR} - Q_{RS}$, it follows:

$$E_{off} = \frac{1}{3} V_{bat} \cdot \left(Q_{RR} - \frac{1}{2} \cdot \frac{I_{RRM}^2}{di/dt} \right) \tag{5.81}$$

If this result is compared with the result (5.78), the case of the inductor-determined turn-off process, then one can see: the turn-off energy loss is less than half of the estimation in Eq. (5.78). The transistor has relieved the stress for the diode by its $v_C(t)$ which drops across the transistor during the diode turn-off event and which slowly decays during the turn-on of the transistor.

Nevertheless, for this reduction of losses in the diode, the switching transistor has to pay the price during its turn-on. From the same simplified consideration in Fig. 5.23, it can be derived that assuming a freewheeling diode without reverse-recovery current maximum – and thereby without a stored charge – the idealized turn-on energy loss in the transistor would be

$$E_{on}(tr, id) = \frac{1}{2} \cdot (t_{ri} - t_s) \cdot I_{CE} \cdot V_{bat} + \frac{1}{2} \cdot t_{fv} \cdot I_{CE} \cdot V_{bat} \tag{5.82}$$

By the freewheeling diode, the following terms are generated additionally:

the dissipated energy because the current-increase time t_{ri} is prolonged by t_s; in this interval, the current as well as the voltage are high
the power losses caused by Q_{RS}
the power losses caused by Q_{RF}

This leads to additional losses ΔE_{on}—in the order of the enumeration:

$$\Delta E_{on} = t_s \cdot I_F \cdot V_{bat} + Q_{RS} \cdot V_{bat} + \frac{2}{3} Q_{RF} \cdot V_{bat} \tag{5.83}$$

Comparing this with Eq. (5.79), one can state that the losses in the transistor, generated by the diode, are higher than the losses in the diode itself by the first terms in Eq. (5.83) and additional twice of the diode losses according to Eq. (5.79) occurs as last term in Eq. (5.83).

The turn-on losses in the transistor amount to $E_{on}(tr) = E_{on}(tr,id) + \Delta E_{on}$. The sum of the losses caused by the diode in the transistor and in the diode is:

$$E_{off} + \Delta E_{on} = t_s \cdot I_F \cdot V_{bat} + Q_{RR} \cdot V_{bat} \tag{5.84}$$

This is significantly higher than the estimation in Eq. (5.78); in summary, an excessive price was paid for relieving the stress in the diode!

These estimations show that the requirement for diodes is a low reverse-recovery peak I_{RRM}, and thereby a t_s which is as low as possible.

In real circuits, a parasitic inductance that causes additional losses in the diode has to be considered. If the parasitic inductance dominates and if the relief of the diode by the voltage decay of the transistor can be neglected, then the losses in the diode again approach the situation that was described with Eq. (5.78). Comparatively high parasitic inductances have to be taken into account for example in some traction applications.

In the application, neither an ideal switch nor a circuit without any inductance can be presumed. Thus, Eq. (5.74) must be used for an exact determination of the switching losses. The waveforms of i and v are recorded with an oscilloscope, multiplied and integrated over the total switching time. The simplified estimation given here comply with measured values in a low-inductive circuit with an accuracy of ±20%.

5.7.3 Reverse Recovery: Charge Dynamic in the Diode

Figure 5.6 has shown the flooding of the base of the diode with free carriers for the forward conduction state. The reverse recovery behavior is determined by the internal behavior of the stored plasma and the time dependent shape during its removal. First, this shall be investigated qualitatively. Figure 5.24 depicts the simulation of the stored carriers in a snappy diode, whereas Fig. 5.25 shows the same for a soft-recovery diode.

At forward conduction, the n^--base of the diode is flooded with free electrons and holes in a range of 10^{16} cm^{-3}; they build up a neutral plasma in which the density of electrons n and of holes p are approximately equal. After commutation a neutral plasma zone with $n \approx p$ is still present in the diode until the time t_4. The removal of free carriers takes place towards the cathode by the electron current and towards the anode by the hole current, the plasma removal process occurs in the

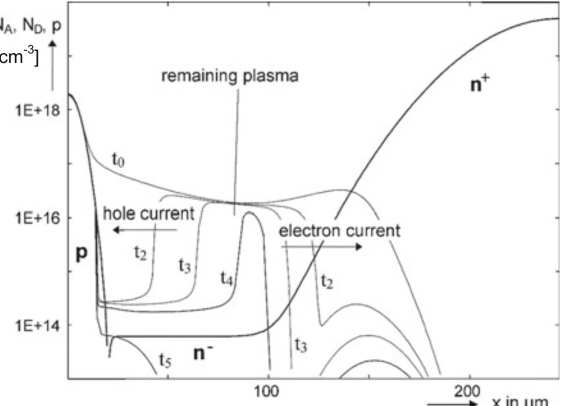


Fig. 5.24 Doping profile and concentration of the remaining plasma during reverse recovery in a snappy diode (ADIOS-Simulation)

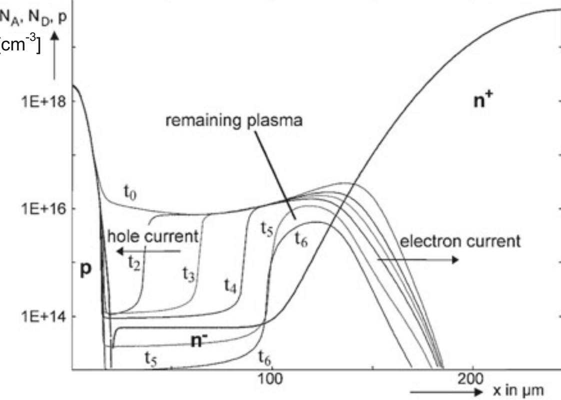


Fig. 5.25 Doping profile and concentration of the remaining plasma in a soft recovery diode (ADIOS-Simulation). Compared to Fig. 5.24 the difference is caused by an inhomogeneous carrier lifetime

outer circuit as reverse current. In the case of the snappy diode (Fig. 5.24), both edges of the plasma meet shortly after t_4, and suddenly the source for feeding the reverse current vanishes. The reverse current is interrupted abruptly, the reverse recovery behavior is snappy.

The time-dependent shape of the remaining plasma for a soft recovery diode is shown in Fig. 5.25. The doping profile is the same as in Fig. 5.24 in this example, but the on-state carrier concentration ($t = t_0$) is changed resulting from an inhomogeneous carrier lifetime, i.e. a low lifetime at the pn-junction, and a higher lifetime at the nn$^+$-junction. During the whole reverse recovery process, a neutral plasma remains inside the diode and feeds the reverse current. At the instant t_5 the diode has taken over the applied voltage. A plasma decay as shown in Fig. 5.25 leads to a tail current as depicted in Fig. 5.21b.

Whether soft recovery behavior is achieved depends on the time-dependent decay of the plasma and how this process is controlled suitably. It took a comparatively long time until the reverse recovery behavior was mastered.

For analyzing the dynamic behavior of the plasma, two cases will be investigated in more detail in the following. The first case refers to the effects during the increase in voltage and is based on the model of Benda and Spenke [Ben67]. With this model is investigated under which conditions a reverse current snap-off or a soft recovery behavior is to be expected. Thereafter, a further analysis is conducted for the case that the device has already taken the voltage with a soft-recovery shape of the current and that despite this, a snap-off does occur at the end of the tail current.

For investigating the first case based on the model of Benda and Spenke, the following simplifications are assumed:

- The pn- and nn$^+$-junctions are assumed to be abrupt, and the charge within the highly doped regions is neglected.
- The plasma edges are treated as abrupt.
- Since the amount of charge carriers removed by the current is much higher than the recombination of charge carriers, recombination in this time interval is neglected.
- Additionally, the hyperbolic asymmetric shape of the initial plasma (Fig. 5.6 respectively Eq. (5.27)) is, as a first step, simplified by a constant plasma density, whereby $n = p = \bar{n}$ is assumed.

Figure 5.26 illustrates this simplified model. In the plasma the current consists of electron current and hole current; it holds that $j = j_n + j_p$ whereby

$$j_p = \frac{\mu_p}{\mu_n + \mu_p} j$$
$$j_n = \frac{\mu_n}{\mu_n + \mu_p} j \tag{5.85}$$

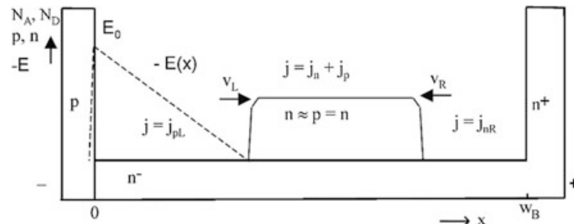


Fig. 5.26 Simplified drawing of the removal of the internal plasma in a diode

On the left side of the plasma zone, towards the pn-junction, a part of the n^{-}-base has become free of carriers and a space charge is building up. The current flows here as a hole current $j = j_{pL}$. From the continuity condition for the current at the border of plasma and space charge, it follows:

$$j_{pL} = j_n + j_p \tag{5.86}$$

For the difference of the hole current, it is therefore valid at this position that

$$j_{pL} - j_p = \Delta j_p = j_n \tag{5.87}$$

Using (5.85) it follows:

$$\Delta j_p = \frac{\mu_n}{\mu_n + \mu_p} j \tag{5.88}$$

Both sides of Eq. (5.85) multiplied by the time interval dt lead to a differential charge

$$\Delta j_p \cdot dt = \frac{\mu_n}{\mu_n + \mu_p} j \cdot dt \tag{5.89}$$

This differential charge corresponds to a charge $q \cdot \bar{n} \cdot dx$ stored in a volume element dx while the plasma zone is shortened by dx. Consequently, (5.89) becomes

$$q \cdot \bar{n} \cdot dx = \frac{\mu_n}{\mu_n + \mu_p} j \cdot dt \tag{5.90}$$

and for the velocity of the movement of the left-hand side border of the plasma to the right hand side, one obtains

$$|v_L| = \frac{dx}{dt} = \frac{\mu_n}{\mu_n + \mu_p} \cdot \frac{j}{q \cdot \bar{n}} \tag{5.91}$$

Analogously, this can be analyzed for the right side of the plasma zone. From the continuity condition for the current at the border between the plasma zone and the plasma-free zone at the right-hand side, it follows:

$$j_{nR} = j_n + j_p \tag{5.92}$$

and this leads to

$$|v_R| = \frac{\mu_p}{\mu_n + \mu_p} \cdot \frac{j}{q \cdot \bar{n}} \tag{5.93}$$

For silicon, with $\mu_n \approx 3\mu_p$, one obtains $v_L \approx 3v_R$: The plasma zone is removed from the side of the pn-junction thrice as fast as from the side of the nn^+-junction. With the assumed constant density of free carriers in the plasma, both plasma fronts will finally meet at

$$w_x = \frac{v_L}{v_L + v_R} \cdot w_B = \frac{3}{4} w_B \tag{5.94}$$

The electric field is built up by the space charge in the part of the n^--layer between the pn-junction and border of the plasma. The electric field cannot penetrate the plasma zone, since this is neutral. Hence, the electric field will have a triangular shape. The voltage across the device corresponds to the area below the line $-E(x)$ in Fig. 5.26. The plasma-free zone at the right side does not contribute to the voltage in this first approximation.

Therefore, the device can take over as much voltage as possible with a width w_x before a reverse current snap-off occurs. This voltage shall be designated as threshold voltage for snappy behavior V_{sn}. Since electrons are not present in the space-charge region, one obtains

$$V_{sn} = \frac{1}{2} \frac{q \cdot N_D}{\varepsilon} w_x^2 \tag{5.95}$$

The current j in this region flows as a hole current, and the holes have the same polarity as the positively charged donor ions of the background doping. Equation (5.95) is valid as long as the density of holes p, which are traveling through the space charge, is low, $p \ll N_D$, and can be neglected. With increasing p, N_D has to be replaced by $N_{eff} = N_D + p$. Therefore, the gradient of the electric field and thereby the area below the line $-E(x)$ is increased by the current. The device can thus take over more voltage. Maximally, this voltage can take the value at

which avalanche breakdown sets in. Rewriting Eq. (5.1), one obtains the maximal voltage V_{sn} as a function of w_x to be

$$V_{sn} = \left(\frac{w_x^6}{2^4 \cdot B}\right)^{\frac{1}{7}} \tag{5.96}$$

This is the maximal voltage for triangular electric field shape. Equation (5.96) predicts a higher value for V_{sn} than Eq. (5.95). In most cases, the prediction from Eq. (5.95) is too low, especially for switching with high *di/dt* which is typical in applications with IGBTs as switching devices; at these conditions the hole concentration p cannot be neglected. Equation (5.96) is closer to the experimental observation. It should be noted here that V_{sn} is always significantly lower than the breakdown voltage V_{BD} of the device. Since with a wide w_B also increases w_x, this shifts the snap-off of the reverse current to a higher voltage.

Earlier, some suggestions to achieve soft-recovery behavior in fast diodes have been based on this movement of the plasma fronts and proposed a wide w_B [Mou88]. Even some recent solutions are applying this approach. Here, the diode must be designed for triangular field shape (NPT), and w_x is determined using Eq. (5.1). Moreover, the doping is chosen as high as possible for the respective voltage according to Eq. (3.80), w_x results from (3.85) or (5.1) and finally, according to Eq. (5.94), $^1/_3 w_x$ is added to determine w_B.

As was already mentioned in Sect. 5.3 in the investigations on dimensioning of fast diodes, the minimal width of the n^--layer of a NPT diode is the $2^{2/3}$-fold of the minimal width of the n^--layer of a PT diode – the diode with the smallest width of the n^--region for the required voltage. Compared to the minimal width $w_{Bmin} = w_{B(PT,lim)}$ of Eq. (5.12), the suggestions above lead to

$$w_{\mathrm{B}} = \frac{1}{0.63} \cdot \frac{4}{3} w_{\mathrm{B\,min}} \cong \sim 2,1 \cdot w_{\mathrm{B\,min}} \tag{5.97}$$

This results in a significant increase in the forward voltage drop, to which w_B contributes with the power of two or even exponentially. To avoid this high forward voltage, the charge-carrier lifetime can be increased, but this contradicts the requirement that the diode must be fast. Hence, further measures are necessary to achieve soft-recovery behavior, especially under condition of fast switching events in IGBT circuits. With the behavior of the internal plasma as in Fig. 5.25, a soft recovery is obtained without making w_B as thick.

Up to now, the plasma profile was assumed to be homogeneous and constant over the base of the diode, which contradicts reality. However, this simplified approach allows useful conclusions to be drawn also for an inhomogeneous distribution. According to Fig. 5.5 or Eq. (5.27), the plasma density is increased at the pn-junction because of the different mobilities of electrons and holes. This is now taken into account by the densities $\overline{n}_L$ at the side of the pn-junction and $\overline{n}_R$ at the

side of the nn+-junction. The term $\overline{n}_L$ stands for the average density of carriers close to the pn-junction. Division of (5.93) with (5.91) leads to

$$\frac{v_R}{v_L} = \frac{\mu_p}{\mu_n} \cdot \frac{\overline{n}_L}{\overline{n}_R} \tag{5.98}$$

Furthermore, the quantity $\eta = p_L/p_R$ introduced Sect. 5.4.5 in Eq. (5.55) shall be used, and we approximate that this proportionality is valid within some distance to the respectively junction, e.g. $\eta = \overline{n}_L/\overline{n}_R = \overline{p}_L/\overline{p}_R$. Inserted into (5.94), this results in:

$$w_x = \frac{1}{1 + \frac{\mu_p}{\mu_n} \cdot \eta} \cdot w_B \tag{5.99}$$

If the profile in Fig. 5.6 is simply expressed with $\eta = 2$, then (5.99) yields

$$w_x = \frac{3}{5} \cdot w_B \tag{5.100}$$

The diode would snap-off even at a lower voltage than estimated with Eq. (5.94), or a diode with abrupt highly doped boarder regions and homogeneous lifetime in the base would have to be made even thicker than according to Eq. (5.97)!

On the other hand, if the profile in Fig. 5.6 can be inverted to maintain a higher density of carriers at the nn+-junction compared to the pn-junction, this will be advantageous [Sco89]. If, for example $\eta = 1/3$, then Eq. (5.99) leads to $w_x = 0.9 \cdot w_B$, and w_B would have to be widened by a much smaller amount to obtain a sufficient w_x. Such a distribution can be achieved by special structures of the p-emitter, or by using a p-region which is much lower doped than the n+-region, or by an inhomogeneous carrier lifetime which, at the pn-junction, is much lower than deep in the base. Such measures are used in modern fast diodes (see Sect. 5.7.4).

The above analytical approach using a simplified model gives us a good understanding of the physics of the recovery process. Considering the results of the numerical simulation shown in Fig. 5.24, one notices the following deviations from the simple model:

- The p and n+ region usually have diffusion profiles; hence, the pn- and nn+-junction are not abrupt.
- The very low gradient of the doping concentration at the nn+-junction in the figure results in a later start of the carrier removal at the nn+-junction than at the pn-junction. Therefore, the low doping gradient at the nn+-junction is advantageous for the recovery behavior. The doping gradient at the pn-junction, on the other hand, should be as abrupt as possible, because this involves an early rise in the reverse voltage and a reduced peak reverse current.

- Of course, the gradients at the edges of the plasma cannot be infinite. However, the non-abrupt transition to the depletion regions does not greatly affect the velocity at the edges, because after a time interval dt the form of the carrier concentrations at the boundaries is approximately unchanged and only the thickness of the uniform plasma region is reduced.

Another aspect not included in the above model is avalanche generation which often takes place during fast switching, since free carriers, still present in the space-charge region, enhance the electric field. This “dynamic avalanche” will be discussed later in Chap. 13.

Now the second case will be considered: the device has successfully sustained the interval of the voltage increase, whereby the reverse recovery in this interval was soft. The device has taken over the applied voltage, but the plasma was not completely removed. Resulting from the remaining plasma, a tail current still flows. This tail current flows through the space charge as hole current, there holds $j = j_p$. Under the given condition of a high field, the holes flow with the drift velocity $v_{d(p)}$, which approximates the saturation drift velocity v_{sat}. Accordingly, the hole density in the space charge is

$$p = \frac{j}{q \cdot v_{sat}} \tag{5.101}$$

This influences the gradient of the electric field in the space charge region as

$$\frac{d\mathsf{E}}{dx} = \frac{q}{\varepsilon}(N_D + p) \tag{5.102}$$

This situation is illustrated in Fig. 5.27 for a diode with low base-doping N_D. The voltage can be assumed to be constant in the investigated time interval: consequently, the area under $-\mathsf{E}(x)$ is constant. The hole density is one factor in $d\mathsf{E}/dx$,

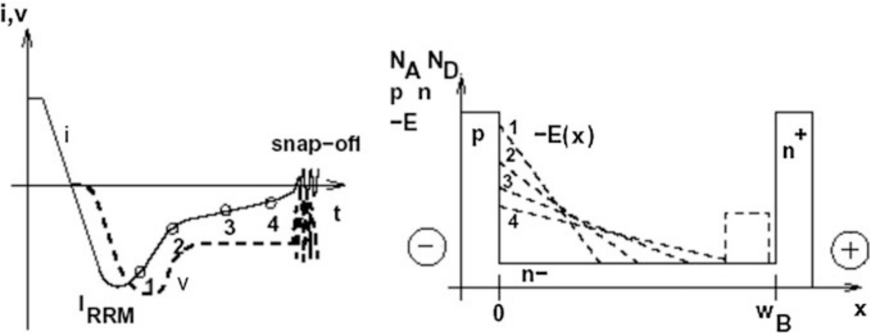


Fig. 5.27 Reverse current snap-off at the end of the tail current. Left: Waveform of current and voltage. Right: Electric field at different points in time

and p is determined by the hole current extracted from the remaining plasma. This hole current leads to a removal of the remaining plasma. By applying Eq. (5.101), (5.91) is changed to:

$$|v_l| = \frac{dx}{dt} = \frac{\mu_n}{\mu_n + \mu_p} \cdot \frac{p \cdot v_{sat}}{\bar{n}} \tag{5.103}$$

As p and j decrease, $d\mathsf{E}/dx$ becomes lower, and the space charge layer widens. However, if the space charge reaches the end of the base while still a significant current is flowing, the source of the current will suddenly vanish and the current will then snap-off. The electrical field springs from a triangular to a trapezoidal shape.

To avoid this effect, the device must be capable of taking the space charge at a given voltage without a punch of the space charge to the n^+-layer. The voltage limit, at which the space charge reaches the n^+-layer at a given background doping N_D and a base-width w_B, is:

$$V_{sn} = \frac{1}{2} \frac{q \cdot N_D}{\varepsilon} w_B^2 \tag{5.104}$$

For this case which is described in more detail in [Fel04], w_x can be set to w_B. As long as the battery voltage V_{bat} stays below V_{sn}, no reverse current snap-off will occur. In the static case or upon occurrence of voltage peaks, the diode can bear much more voltage, since then a trapezoidal space charge can be built up, and the breakdown voltage V_{BD} is much higher than V_{sn}.

The simplified description up to now can not give an answer how a diode transits from the plasma in Fig. 5.26 to that in Fig. 5.27, since only the field-induced drift components of the current were considered. Taking into account the diffusion components, the movement of the plasma layer backwards to the cathode can be described. The diffusion components on the right side of the plasma become dominating in the tail current phase, and the remaining plasma can move from the position in Fig. 5.26 to that in Fig. 5.27. This is explained in [Bab08].

The effect of reverse current snap-off in the tail phase can especially occur with diodes designed for higher voltages (> 2000 V). In such applications, often a low di/dt is applied, because the circuit contains significant parasitic inductances. This is the case in some applications for very high power control. Usually in such applications, the battery voltage is limited to 66% of the breakdown voltage V_{BD} for which the device is designed. In order to keep V_{sn} high, the doping N_D is not allowed to be too low. This, however, contradicts the demands of cosmic ray stability, see Sect. 12.8. Hence, an optimal trade-off has to be found.

5.7.4 Fast Diodes with Optimized Reverse-Recovery Behavior

Recombination centers are implemented in all fast silicon pin-diodes. The characteristics of the employed recombination centers gold, platinum and radiation induced centers have been described in Sects. 2.7 and 4.9. With the density of recombination centers, the charge carrier lifetime and thereby the stored charge Q_{RR} is decreased. However, there is no direct relationship between the concentration of recombination centers – as long as their axial distribution is kept constant – and the form of the reverse current decay. Therefore, it cannot be determined by the recombination center density whether the behavior will be soft or snappy.

It was already shown that the recovery behavior will become soft, if only the width w_B of the lowly doped base region is wide enough. Yet this leads to very high forward-conduction losses and/or switching losses, which can not be accepted in most cases. The modern concepts aim to adjust soft recovery behavior without using a strongly enhanced base-width.

5.7.4.1 Diodes with a Doping Step in the Low-Doped Layer

To avoid an excessively wide w_B and to minimize the disadvantages resulting from it, in 1981 an n^--layer with a step in the doping density was suggested by Wolley und Bevaqua [Wol81]. The doping profile of this diode is shown in Fig. 5.28. Approximately in the middle of the base, the doping is augmented by a factor of 5 – 10. Such layers are manufactured using a two-step epitaxy process.

When the space charge builds up and the field penetrates the higher doped layer n_2, it decreases there with a steeper gradient. At turn off, the remaining plasma zone is located in the layer n_2. The voltage which the device can take corresponds to the area below the line $-E(x)$. This area is larger than that for a triangular field shape. The threshold voltage for possible snappy reverse-recovery behavior is shifted to a higher value. This measure is nowadays often used for diodes manufactured in epitaxial technology in the voltage range of up to 600 V. For diodes of higher voltage, especially above 1200 V, the fabrication of an epitaxial layer with the

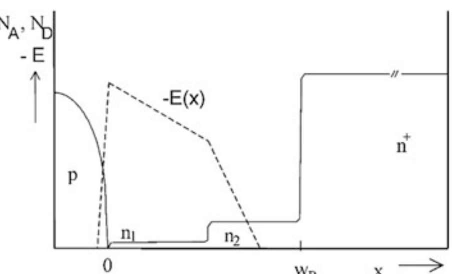


Fig. 5.28 Diodes with a step in the doping concentration of the low-doped layer

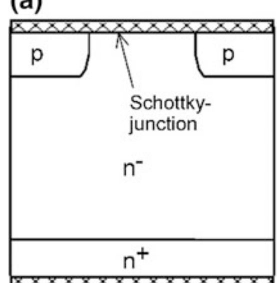
necessary thickness is too laborious. For soft-recovery behavior under all conditions relevant for the application, usually one of the following concepts is used.

5.7.4.2 Diodes with Anode Structures for Improving the Recovery Behavior

It was already shown that the carrier concentration in the flooded base of the pin-diode with highly doped boarder regions is higher at the pn-junction than at the nn^+-junction (see Fig. 5.6). This is disadvantageous for the reverse-recovery behavior. Therefore, concepts were developed to invert this distribution: the concentration shall be higher at the nn^+-junction than at the pn-junction. This principle was explained already in [Sco89], and there were several approaches to realize this by using structures in the p-anode layer.

For example, Schottky junctions cannot inject holes. Thus, it is obvious that implementation of Schottky junctions on a part of the area will lead to the desired distribution, considering the average concentration over the area. This measure is often discussed in the literature.

The "Merged Pin-Schottky" (MPS) diode consists of sequenced p-layers and Schottky regions [Bal98] (Fig. 5.29a). The distance between the p-layers is chosen to be so small that, in case of blocking voltage, the Schottky junction is shielded from the electric field, and only low field-strength occurs at its position. Consequently, the high leakage current of a Schottky junction is avoided. Figure 5.30 presents the forward characteristics of a MPS diode similar to that of [Bal98] calculated with the device simulator TCAD DESSIS [SYN07]. The structure with w_B = 65 µm is designed for a blocking voltage of 600 V. Compared here are the forward characteristics of the pin region, of the Schottky region and of the parallel connection of both in a MPS diode. At low current density the MPS diode approaches the characteristics of a Schottky diode. In the range of 200 A/cm^2, the typical current density of fast 600-V diodes at the rated current, the advantage caused by the Schottky regions is only small. At increased current density, the MPS diode has a higher forward voltage because of the loss of area of p-emitter regions.



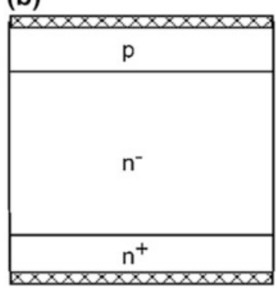


Fig. 5.29 p-emitter for improving the reverse-recovery behavior: **a** Emitter structures of the Merged Pin/Schottky diode **b** uniformly reduced p-doping, low emitter efficiency

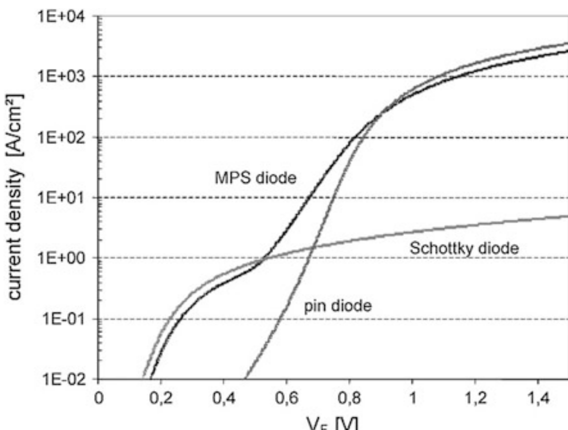


Fig. 5.30 Forward characteristics of a MPS diode with 50% Schlottky area. The characteristics of the pin-region and of the Schottky region are shown in comparison. Simulation using TCAD DESSIS

If the MPS diode is designed for a blocking voltage of 1000 V and greater, then the range of lower forward-voltage drop shifts to lower current density, because then the voltage drop in the low-doped base region predominates. But the effect, that the area of the p-region is reduced and thereby the injection of carriers at the anode side of the device is reduced, maintains its function. An inverted distribution of free carriers is formed (see Fig. 5.31, right-hand side).

Since the p-layers must be arranged closely to shield the Schottky regions from the electric field, the possibility is limited to make their share of the surface area small. A further idea to improve the MPS diode is the "Trench Oxide Pin Schottky" (TOPS) diode [Nem01] which was presented by Fuji. The structure is shown in Fig. 5.31. The p-anode regions are placed at the bottom of the trench cells. A Schottky contact is formed at the semiconductor surface. Hole injection takes place only by the p-layers that are connected via the resistor of the polysilicon layer in the trench. Thus, the integral hole injection over the area is strongly reduced.

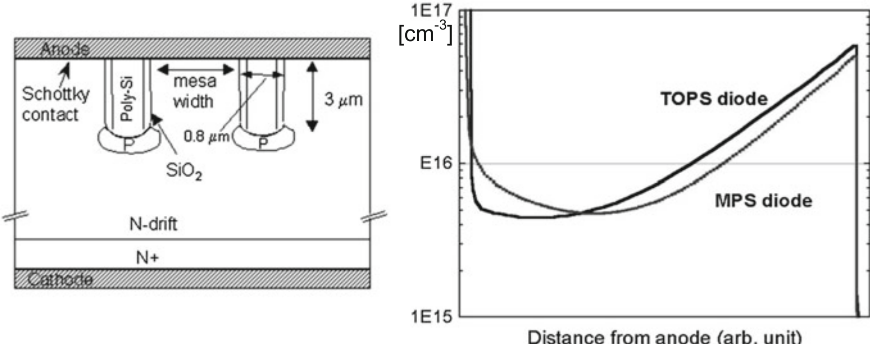


Fig. 5.31 Trench Oxide Pin Schottky (TOPS) diode. Structure (left) and vertical distribution of holes in the forward-conduction state (right). Figure taken from [Nem01] © 2001 IEEE

This results in a profile of free carriers as is shown in Fig. 5.31 on the right side, which exhibits a stronger inversion of the internal plasma.

A soft-recovery behavior can be expected from this plasma profile. The p-regions at the bottom of the trenches effectively shield the surface against the electric field; therefore, the leakage current is kept low with a narrow arrangement of the trenches.

There are several further concepts of structures at the anode side, among them, also structures with diffused p^+- and n^+-regions. What they all have in common is the reduction of the area of layers that are injecting holes and, thus, the reduction of the density of free carriers at the pn-junction.

5.7.4.3 The EMCON-Diode

Instead of reducing the emitter area by emitter structures, an homogeneous p-layer with high emitter recombination, as is shown in Fig. 5.29b, can also lead to the desired inverted distribution of the plasma. This concept is applied in the "Emitter Controlled" (EMCON) Diode [Las00]. It needs much less effort in production compared to the MPS diode or to the TOPS diode.

The EMCON diode utilizes a p-emitter of low emitter efficiency. The emitter parameter h_p, introduced in Eq. (3.98), can be expressed for a p-emitter which is not too highly doped as follows:

$$h_p = \frac{D_n}{p^+ \cdot L_n} \tag{5.105}$$

To reduce the emitter efficiency γ according to Eq. (3.100), h_p must have a high value. According to Eq. (5.105), this can be done if the doping density of the emitter p^+ is chosen low and also if the effective diffusion length L_n is adjusted to a low value. Both measures are used in the EMCON diode. p^+ must be sufficiently high to avoid a punch-through of the electric field to the semiconductor surface. For a thin p-layer, L_n is approximately the same as the penetration depth of the p-layer w_p, and also this depth is small in an EMCON diode. Under this condition it holds that

$$h_p = \frac{D_n}{p^+ \cdot w_p} = \frac{D_n}{G_n} \tag{5.106}$$

where $G_n = p^+ \cdot w_p$ is the Gummel-number of the emitter under the condition of an abrupt emitter [Sze81]. The term G_n represents the number of doping atoms per area. For the diffused emitter of the EMCON diode, it is more precisely expressed as:

$$G_n = \int_0^{w_p} p(x)dx \tag{5.107}$$

Using the Gummel-number according to (5.107) increases h_p compared to the abrupt emitter. High h_p and therewith low γ leads to a decreased p_L, as is necessary to create the inverted distribution. According to (5.52), h_p is the dominating factor for the emitter recombination. High h_p means that a significant share of the total recombination takes place inside the p-emitter or at the surface.

An emitter of high doping and high efficiency is applied at the cathode side of the EMCON diode; hence, the plasma density at this side is high.

Figure 5.32 depicts the turn-off waveform of an EMCON-HE diode. The measured turn-off behavior is compared to the numerical device simulation. Here, the simulated waveform agrees sufficiently well with the measured characteristic. The numerical simulation enables visualization of the effects inside the device. The simulated distribution of the density of free carriers is shown in Fig. 5.33 for the indicated time steps in Fig. 5.32.

At forward conduction the diode is flooded with free carriers. In Fig. 5.33 the rhombic dots show the measured carrier distribution in an EMCON diode at forward conduction as is obtained by the internal laser deflection method [Deb96]. The density of free holes calculated with the device simulator (line A in Fig. 5.33 for instant A in Fig. 5.32) is in good agreement with the measured density. The hole density represents the plasma density; $n \approx p$ applies for the regions of high flooding below the lines A, B, C etc. For the initial distribution at instant A, it holds that $\eta = p_L/p_R \approx 0.25$; this is attained by a strong emitter recombination in the anode.

During commutation and change of the polarity of the voltage (C, D, E in Fig. 5.32 and 5.33), the hole current flows to the negatively poled anode on the left side, and the electron current flows to the positively poled cathode. As is illustrated in Fig. 5.33, the removal of the stored charge occurs for the instants B, C, D, E. At the instant C, the diode has reached the maximal reverse current. After that, a current can still flow resulting from the removal of the still existing plasma which ensures a soft recovery behavior.

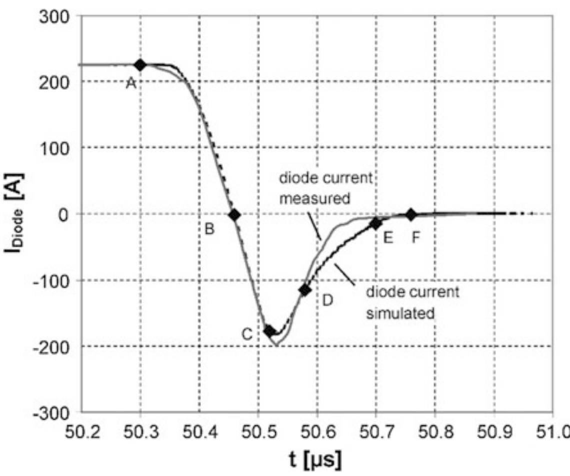


Fig. 5.32 Turn off behavior of a 1200 V EMCON-HE diode, measured and simulated at 600 V, 25 °C, 225 A/cm^2

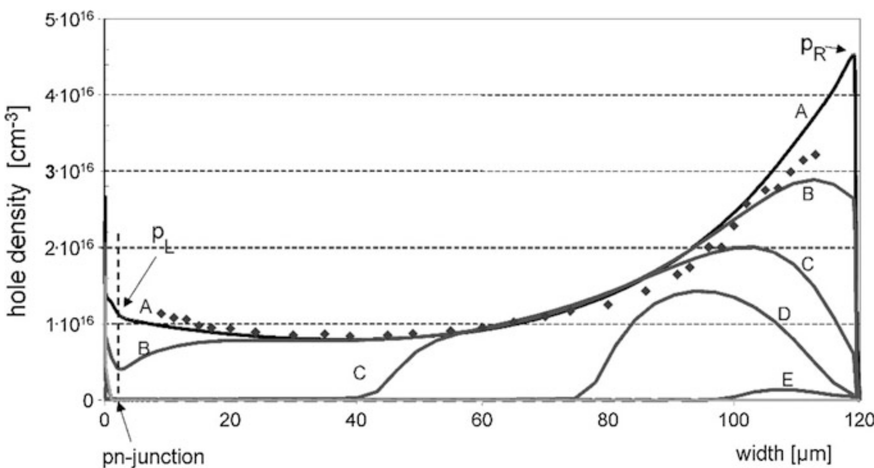


Fig. 5.33 Carrier distribution in an EMCON diode for forward conduction and during commutation. Measured plasma distribution in the forward-conduction state (rhombuses), simulated hole distribution for the forward-conduction state (line A), removal of the internal stored free-carriers (on example of holes) during commutation (lines B–E)

The low p-emitter efficiency of the EMCON diode leads to the disadvantage of a reduced surge current capability. A structure proposed in [Sco89], the so-called SPEED-diode containing high p-doped regions, removes this drawback to a mayor part.

5.7.4.4 The CAL-Diode

For the carrier lifetime adjustment in the diodes, often a platinum diffusion is used, whereby the vertical profile of the recombination centers (Fig. 4.28) results from the diffusion process and cannot be modified. The advantage of an adjusted profile of recombination centers, created by implantation of light ions, was already recognized early [Sil85, Won87]. But at that time this technology, which requires particle accelerators of energy of up to 10 meV, was only available for research issues. The situation changed in the early 90s of the last century. Then, the interest of basic researchers in high-energy physics migrated to the GeV region, and particle accelerators in the medium energy range became available for semiconductor production.

The first diode that reached the maturity of a series product with this technology was the so-called "Controlled Axial Lifetime" (CAL) diode [Lut94], designed for a blocking voltage of 1200 V in 1994. Meanwhile several manufacturers have applied this concept, whereby devices of up to 9 kV have been realized and are commercially available.

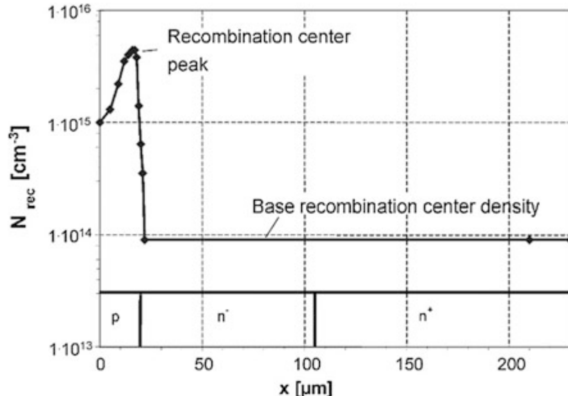


Fig. 5.34 Profile of recombination centers in the CAL diode (scheme)

Figure 5.34 illustrates the profile of recombination centers in the CAL diode. The recombination center peak is created by implanting He^{++}-ions. From this implantation, a profile of lattice defects similar to those in Fig. 4.19 results. Its depth can be adjusted by the energy of the helium-implantation, and its peak density by the dose. In combination with this implantation, the base recombination center density is adjusted as homogeneous across the base, preferably by electron irradiation. Three degrees of freedom – the recombination center peak depth, its height and the base recombination center density – are available for adjusting the reverse-recovery behavior.

It is most suitable to locate the recombination center peak close to the pn-junction. The main requirement is a low reverse-recovery current peak I_{RRM}. For this the pn-junction must be free of carriers at an early moment of the reverse recovery period. The relation between forward voltage V_F and I_{RRM} is better, the closer the recombination center peak is located to the pn-junction. In the CAL diode, the peak of radiation-induced recombination centers is located in the p-layer close to the pn-junction as is shown in Fig. 5.34. Consequently, the main part of multi-vacancies, which act as generation centers (see Sect. 4.9), is outside of the space charge, and the arrangement leads to a low leakage current.

An inverted plasma distribution in the on-state is achieved by this arrangement of the recombination center peak. The on-state plasma distribution shown in Fig. 5.25 is calculated with a recombination center profile according to Fig. 5.34.

The reverse-recovery behavior of the CAL diode was already displayed in Fig. 5.21b. The reverse recovery current peak, which can be adjusted by the height of the recombination center peak, is reduced, and the main part of the stored charge is extracted during the tail current phase. The tail current can be adjusted by the base recombination center density. An increased base recombination center density shortens the tail current time, but this has the drawback of increased forward voltage. With the given degrees of freedom, the reverse recovery behavior can be adjusted in a wide range. Thus, a diode can be designed that exhibits soft-switching behavior under all conditions relevant in the application, especially also at low current.

5.7.4.5 The Hybrid Diode

Modern fast diodes have been considerably optimized with the described concepts. Thus, there is only a small difference between the CAL diode and the EMCON HE diode in the voltage range of 1200 V with respect to the relationship of forward voltage drop V_F and reverse-recovery charge Q_{RR} [Lut02]. Moreover, there are indications that the limits for possible optimization of fast 1200 V rated diodes based on silicon have almost been approached. However, by parallel- and series connection of different diode designs, these limits given for a single device can be exceeded to some degree.

The hybrid diode [Lut00] is shown in Fig. 5.35a. It consists of a parallel connection of two diodes with contrary switching behavior, whereby on the one hand, a snappy diode D_E is used whose low-doped middle layer is designed as thin as possible; this design is also known as a “Punch-Through” (PT)-diode. On the other hand, a soft-recovery diode D_S is used. With the correct adjustment of the characteristics of both diodes in parallel connection, the low forward-voltage drop of a PT-diode combined with soft-recovery behavior of the soft diode can be obtained simultaneously.

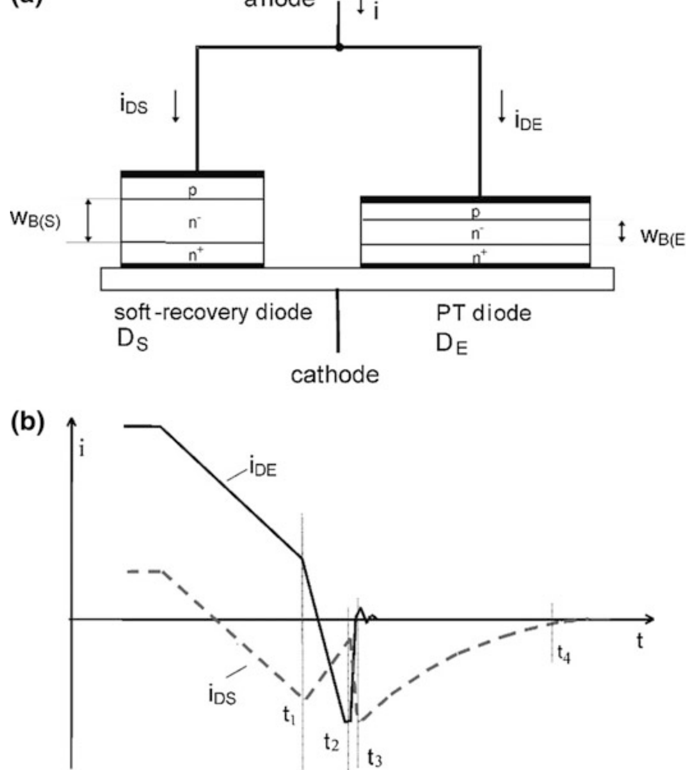


Fig. 5.35 Hybrid diode: **a** Structure; **b** Current waveform in both partial diodes during reverse recovery

The function principle is presented in Fig. 5.35b. The snappy diode D_E carries the main part of the forward current, whereas the soft diode D_S carries a smaller part. At commutation, the current i_{DS} through D_S crosses zero first, and reaches at the instant t_1 its turning point in the reverse current. At this moment, the diode D_E is still carrying forward current. At t_1, the pn-junction of D_S is free of carriers. Then, the diode D_E is commutated with an increased slope di/dt. The total current is still impressed by the outer circuit.

At t_2 the pn-junction of D_E is free, and similarly the reverse recovery current maximum of D_E is reached. Between t_2 and t_3 the hard snap-back of the reverse current i_{DE} in D_E occurs. The reverse current i_{DS} in D_S, which contains still stored plasma, increases with the same slope. The total current as the sum of i_{DS} and i_{DE} shows no reverse current snap-off. Therefore, no high voltage peak is induced. The remaining plasma in the soft-recovery diode D_S is removed between t_3 and t_4. The behavior of the combined arrangement is soft.

For an effective function of the hybrid diode, the diode D_S has to deliver sufficient charge after the reverse-current snap-off in D_E. Thus, the soft diode D_S must be flooded with sufficient plasma, and for this, it must carry between 10 and 25% of the total forward current. The forward voltage drop of both diodes must match to achieve this. In the practical realization, a very fast epitaxial diode with low width w_B of the middle layer is used as diode D_E. Its forward voltage at rated current is in the range of 1.1 V. As diode D_S, a CAL diode with $^1/_6$ of the area of D_E is connected in parallel. This special CAL diode has the double width w_B compared to the Diode D_E, but by implementing a low recombination center density in the base, its forward-voltage drop is adjusted to achieve the desired current distribution between both diodes. D_S features a large tail current, but it is capable of surely overtaking the reverse current snap-off of the epitaxial diode with a six-fold area.

The hybrid diode is proven in application in drives for fork lifts and other electrical vehicles. In these vehicles the battery voltage is typically in the range of 80 V. The applied step-down converters are built up using MOSFETs. To achieve low conduction losses, the MOSFETs must be designed for a voltage not higher than 160–200 V. For the induced voltage peak according to Eq. (5.71),

$$V_{ind} = -L_{par} \cdot \left(\frac{di_r}{dt}\right)_{\max} \tag{5.108}$$

only 80 V margin are available for V_M if the rated voltage of 160 V should not be exceeded for a battery voltage of 80 V. The controlled current is typically high in these applications, i.e. in the range of 200 – 700 A. High currents lead to a large footprint of the modules, and a significant parasitic inductance L_{par} can hardly be avoided. Therefore in Eq. (5.108), the term di_r/dt, the slope of the current after the reverse-current maximum, has to be low to keep the induced voltage in the range that is allowed in such applications. The hybrid diode has proven its capability under these hard requirements.

5.7.4.6 The Tandem Diode

The tandem diode consists of a series connection of two fast diodes in a common housing. An example [IXY00] is shown in Fig. 5.36. The configuration is intended for the application in a step-up converter for Power-Factor-Correction. The concept of the tandem diode aims to offer diodes with a reverse-recovery charge Q_{RR} as low as possible for the application at very high switching frequencies.

Equation (5.37) describes an approximate relationship between the stored charge and the base-width w_B of a diode. The minimal base-width, however, is determined by the required blocking voltage. To express the base-width as function of the desired voltage, we assume the case of a moderate PT-dimensioning, as it was given in (5.15)

$$w_B = \chi \cdot V_{BD}^{\frac{7}{6}} \quad \text{with} \quad \chi = 2.3 \times 10^{-6} \text{cmV}^{-\frac{7}{6}}$$

Inserting this into Eq. (5.37), one obtains

$$Q_F = I_F \frac{\chi^2}{V_{drift}\left(\mu_n + \mu_p\right)} \cdot V_{BD}^{\frac{7}{3}} \tag{5.109}$$

Now the recovery charge of the diodes connected in series, measured as reverse current integral, is equal to the stored charge of a single diode in the series, hence $Q_{RR}^{(series)} = Q_{RR}$. Furthermore, $Q_{RR} = Q_F$ can be assumed as long as the recombination during the reverse recovery is low. However, the total forward voltage of a series of n diodes is

$$V_F = n \cdot \left(V_{drift} + V_j\right) \tag{5.110}$$

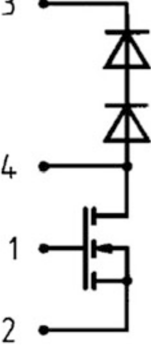


Fig. 5.36 Configuration of a tandem diode and a MOSFET intended for the application as step-up converter

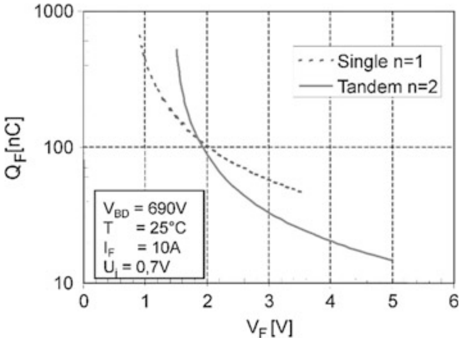


Fig. 5.37 Relation between stored charge and forward voltage for a single diode and for a tandem diode

Expressing the drift voltage of a single diode V_{drift} by the total forward voltage V_F, Eq. (5.109) yields

$$Q_F = I_F \cdot \frac{\chi^2}{(\mu_n + \mu_p)} \cdot \frac{\left(\frac{V_{BD}}{n}\right)^{\frac{7}{3}}}{\left(\frac{V_F}{n} - V_j\right)} \tag{5.111}$$

This relation between stored charge and forward voltage is presented in Fig. 5.37 for a single diode (n = 1) and for a tandem diode with n = 2. For a forward voltage higher than approx. 1.9 V, the tandem diode shows a better trade-off between V_F and Q_F. The tandem diode recommends itself for applications where very high frequencies and low switching losses are required as the primary feature, but where the current is low and conduction losses are less important. If a forward voltage of 3 V is allowed, a significant reduction of the stored charge can be achieved with the tandem diode. The conduction losses are shared between two devices, which is advantageous regarding the thermal management.

Since a PT-dimensioning was assumed for the partial diodes in the above equations, no soft-recovery behavior can be expected under these conditions. But in the intended applications with voltages of around 600 V, the converters are usually build very compactly, so that low parasitic inductances can be realized. In combination with the typically applied low currents, the soft-recovery requirements can be reduced, also considering that soft recovery leads to switching losses during current decay. In the considered voltage range, no RC-network for symmetrical voltage sharing is necessary using modern devices. Competing with the tandem diode are diodes made from GaAs or Schottky diodes from SiC that are also used for very high frequencies.

5.7.5 MOS-Controlled Diodes

The idea underlying the MOS Controlled Diode (MCD) is to improve diode properties by introducing a third electrode, a MOS-gate. Here, we have to anticipate

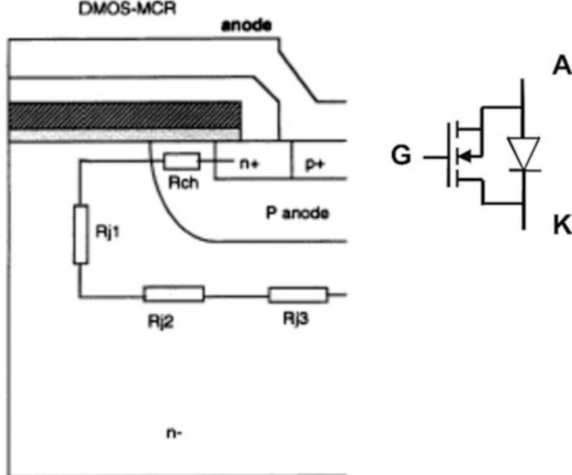


Fig. 5.38 Basic structure of the MOS Controlled Diode (MCD) and equivalent circuit. Figure taken from [Hua94]

some concepts of Chap. 9 about the MOSFET – the Metal Oxide Semiconductor Field Effect Transistor. The basic form of the MCD has the same structure as a power MOSFET. As is shown in Fig. 5.38, the MOSFET contains a $pn^{-}n^{+}$ diode, which consist of the n^{+} drain region, the weakly doped n-region and the p-well which is connected with the source metallization, the anode of the diode [Hua94]. The diode lies parallel to the MOS-channel and conducts if the voltage changes its polarity; it is an "anti-parallel" diode. The voltage at the drain electrode is negative in the diode conducting state. The corresponding equivalent circuit is shown in Fig. 5.38 at the right-hand side.

By opening the channel of the MOSFET with a positive gate voltage, a current path parallel to the pn-junction of the diode is created. If the voltage drop along the channel is smaller than the set-in junction voltage of the diode (≈ 0.7 V at room temperature), almost the whole current is flowing through the channel, and the MCD operates in a unipolar mode without carrier injection. Therefore, during commutation no stored charge of injected carriers is to be extracted. Operating in this mode, the structure was called a "synchronous rectifier" [Shm82]. At a channel voltage higher than the set-in junction voltage, charge carriers are injected by the pn-junction. This range of operation, however, is usually excluded when speaking of a synchronous rectifier.

Normally, the MCD is used in a different manner: since the channel is closed during most of the time of forward conduction, the MCD works like a pin-diode. The channel is opened a short time before commutation, so that the pn-junction is nearly shortened via the n-channel. Therefore, the injection by the anode emitter is

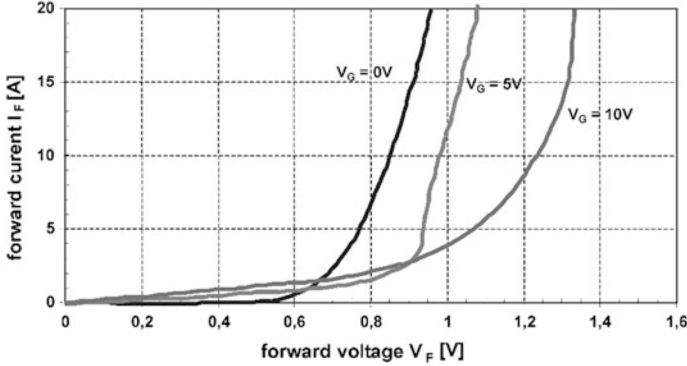


Fig. 5.39 Characteristics of the 1000 V MOSFET IXFX21N100Q in the 3rd quadrant

strongly decreased. This can be expressed also using an effective emitter efficiency of the p-well *and* the channel. Analogously to Eq. (3.96), this can be written

$$\gamma = 1 - \frac{j_n}{j} \tag{5.112}$$

where j_n now denotes the electron current through the channel. If j_n is increased by opening the channel shortly before commutation, γ is strongly decreased and hence also the carrier density is decreased at the anode side of the diode structure. In this way, a drastic reduction of the recovery charge can be achieved. In addition, the inverted carrier distribution leads to a soft-recovery behavior.

The parallel operation of the pin-diode and the MOS-channel is illustrated in Fig. 5.39. A conventional 1000 V MOSFET was found which shows the effect, i.e., the IXFX21N100Q of the manufacturer IXYS. With the channel closed (V_G = 0 V), the current is very small below the set-in junction voltage of approx. 0.6…0.7 V. With opened channel (V_G = 5 V and V_G = 10 V), significant current flows already at smaller voltages where a resistive characteristic is observed. This is caused by the channel resistance, R_{ch}, and resistance components arising from a vertical (R_{j1}) and lateral current paths (R_{j2}, R_{j3}) through the weakly doped n region. On account of R_{ch} the total resistance decreases with increasing gate voltage. This is the operation range of the synchronous rectifier. In the range above 0.7 V, the forward voltage drop at open channel is significantly higher than that in the case of closed channel, as can be seen especially for the higher gate voltage. During the operation cycle of the MCD, the diode works only shortly before turn-off in this part of the characteristics.

In order that opening of the channel results in effective shunting of the current, the voltage drop caused by the external current in the channel and the other resistance components must be smaller than the set–in junction voltage [Hua94]. This holds only up to a critical current I_{crit} given by (at room temperature):

$$I_{crit} = \frac{0.7V}{R_{CH} + R_{j1} + R_{j2} + R_{j3}} \tag{5.113}$$

I_{crit} should be as high as possible and, hence, the resistances R_{Ch} and R_{j1}, R_{j2}, R_{j3} must be as small as possible. An MCD structure optimized for this requirement is shown in Fig. 5.40 [Hua95]. By using a trench structure, the resistances R_{j1} and R_{j2} are removed completely. If additionally the n-region directly under the p-layer is provided with a higher doping, depicted in Fig. 5.40 as n-buffer, also the resistance R_{j3} should be small. But the allowed buffer doping at this position is very closely limited, because it reduces the blocking capability.

The MCD must be switched from the state of fully flooded base, at which no voltage is applied at the gate, to the state of a lower plasma density in the base by applying a voltage to the gate. In this process the main part of the stored charge must be removed. If this is removed by recombination, one has to consider a time of some 10 µs, which is too long for practical application. An n-buffer additionally makes the extraction of holes more difficult. Therefore, an additional p-layer is implemented at the cathode side, which facilitates the extraction of holes to the negatively poled cathode easier.

With these measures, the stored charge before commutation can be reduced significantly; in device simulation even a reduction by a factor of 20–40 was shown [Hua94]. Nevertheless, the stored charge cannot be completely eliminated.

The fundamental disadvantage of all the previously shown variants is that before a blocking voltage is applied to the device, the channel must be closed again. Figure 5.41 depicts an MCD as replacement for a diode in a commutation circuit with an IGBT as switch. The current flows in the freewheeling path, before the

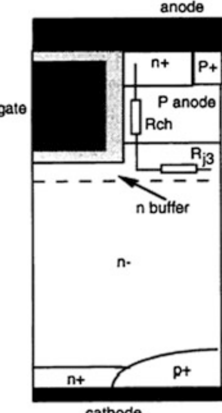


Fig. 5.40 Trench-MCD cell which uses an additional buffer layer at the anode side. Figure from [Hua94]

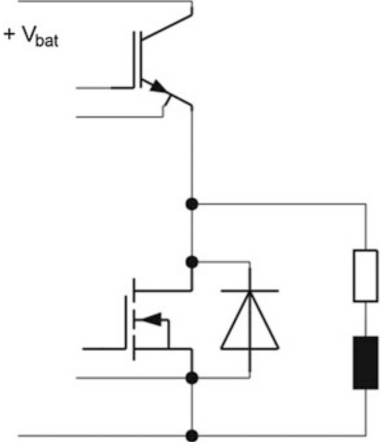


Fig. 5.41 MCD in a commutation branch with IGBT as switching device

diode is commutated from the conducting to the blocking mode at the next turn-on of the IGBT. If the channel of the MCD is open while voltage is applied, the current will not flow into the load but via the channel; a short in the circuit occurs in the commutation branch. Therefore, the channel must be closed beforehand; this point in time must be met with an accuracy of approximately 100 ns. It is hard to ensure this with a driver circuit, however. This disadvantage – that when a positive gate voltage is applied, no blocking capability is given – is the main factor that impedes the implementation of MCD and Trench-MCD in practical applications.

An example for a modified solution without this fundamental disadvantage is the Emitter Controlled Diode (ECD), which is shown in Fig. 5.42 [Dru01]. A low-doped p-zone is connected to the p^+-zone or to the p-well. The highly doped n^+-zone is arranged in the p-well; the path to the low doped p zone is controlled by a MOS-channel. The channel can be prolonged by a very shallow n^+-layer above the low-doped p-layer. Geometry and doping of the respective layers are chosen such that if a positive voltage is applied to the gate and the channel is open, the current takes the path via the low-doped p-layer. For this, the sum of all voltage

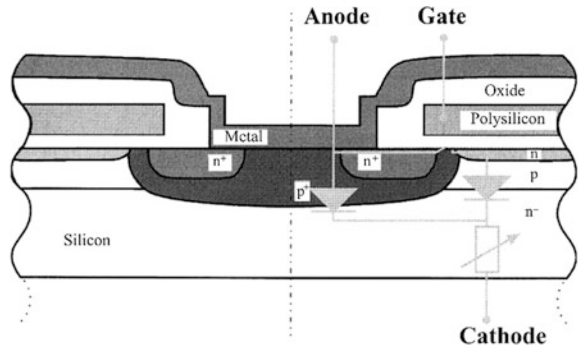


Fig. 5.42 Emitter Controlled Diode (ECD). Figure according to [Dru03]

drops across this path – i.e. the junction voltage at the pn^--junction, the voltage drop in the channel and further resistive parts in the path – must be lower than the junction voltage of the p^+n^--junction.

Without positive voltage at the gate, a $p^+n^-n^+$-structure is given in which a plasma distribution with increasing density to the pn-junction arises, similar to Fig. 5.6. In the case of a channel opened by a positive voltage at the gate, an inverted plasma distribution is given similar to Fig. 5.33 or Fig. 5.31 (right side). The turn-off process is executed only from the open-channel mode; in this mode a turn-off with soft-recovery behavior can be expected. In Fig. 5.43 the internal plasma distribution is compared for both modes. The special progress of the MCD is that, also with open channel, a structure with a blocking capability is given.

The ECD is explained in detail in [Dru01] and in [Dru03] and has not been realized in practice as yet. Nevertheless, it is possible that this idea or similar ideas are used as a foundation for future optimization of diodes in the voltage range of > 3 kV. Even though the necessary effort seems to be high at first glance, one has to consider that, nowadays in applications with IGBTs in this voltage range, the possible switching frequencies are limited especially by the reverse-recovery behavior of the diode. A progress in diodes can lead to an advantage on the system level that might justify the necessary effort.

Derived from the function of the MCD is also the Inverse Injection Dependency of Emitter Efficiency (IDEE) diode [Bab10]. It uses the implementation of n-channels, however there is no gate. The anode layer is highly doped. The therewith achieved high emitter efficiency would cause a snappy reverse recovery behavior, therefore the emitter efficiency is reduced by a parallel electron current. For this, the p-layer is interrupted and a shallow n^+-layer is implemented to ensure an ohmic contact. The structure is shown in Fig. 5.44.

The n-channels are in its dimensioning so narrow that is case of blocking mode the electric field is shielded and is of sufficient distance to the n^+-layers at their

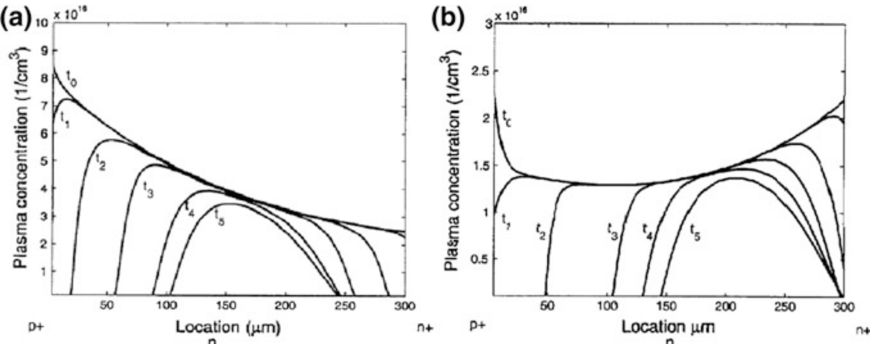


Fig. 5.43 Plasma distribution in the ESD during the reverse-recovery process: **a**) $V_G = 0$, closed channel, **b**) positive gate voltage, open channel. Figure taken from [Dru01] © 2003 IEEE

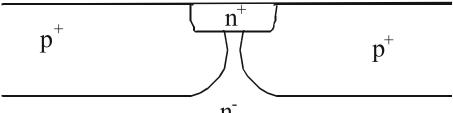


Fig. 5.44 Anode structure of the Inverse Injection Dependency of Emitter Efficiency (IDEE) diode

interface to the contact, see Fig. 5.44. The mode of function is that in forward conduction mode an electron current flows across the channels. This current is limited by the resistance of the channels.

If the voltage drop along the channel and below the p^+-layer, see Eq. (5.113), is larger than the set-in junction voltage of the pn-junction of the diode (≈0.7 V at room temperature), the pn-junction will inject carriers. In this rough simplification, the potential at the junction remains at the junction voltage and electron current across the channel is given by

$$j_n(x_p) = \frac{V_{bi}}{R_{CH} \cdot A_{CH}} \tag{5.114}$$

with R_{CH} as resistance of the channel and A_{CH} as its area. It remains the same even if the current density of the diode is increased. Therefore, at increased current density the share of electron at the total current is reduced, and with this the emitter efficiency γ according to Eq. (5.112) should increase. For usual pn-junctions this is opposite: The emitter efficiency is reduced with increasing current density, see Sect. 3.4. If Eq. (3.96) is rewritten for a p-emitter, it reads

$$\gamma = 1 - j_n(x_p)/j = 1 - q \cdot h_p \cdot p_L^2/j \tag{5.115}$$

Since p_L is approximately proportional to j and is contained with the power of 2 in Eq. (5.115), it results in a decrease of γ proportional to j.

Additionally, Fig. 5.8 has shown a decreased effective lifetime τ_{eff} with increased carrier density, caused by emitter recombination which is stronger with higher carrier density.

The dependency of the emitter efficiency on the current extracted from the simulation results is shown in Fig. 5.45. The conventional p-doping controlled diode (reference) shows the typical strong decay of γ with increasing currents. The IDEE-diode shows low γ at low current. The desired inverted dependency of γ with the current density is not completely, however in tendency achieved. The IDEE diode shows increasing values of γ at high currents.

The IDEE-diode gives a potential of further improvement of the reverse-recovery behavior at low current densities as well as the possibility of lower forward voltage at high current. This leads in the same way to a higher surge-current capability. Surge current behavior is discussed later in Sect. 12.2.

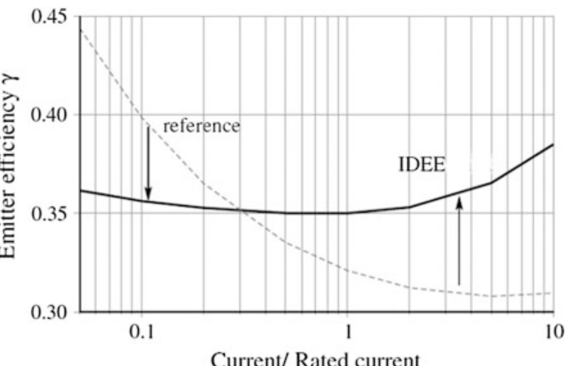


Fig. 5.45 Emitter efficiency γ of the anode-side emitter depending on forward current for a usual diode with homogeneous p-layer (reference) and for the IDEE diode. Figure from [Bab10] © 2010 IEEE reprint with permission

5.7.5.1 Diodes with Cathode Side Hole Injection

While all measures up to now modified the anode side, it was found that also the cathode side, the nn$^+$-junction, can be used to improve the reverse recovery behavior. Structures which inject at reverse recovery additional holes from the cathode side are the Field Charge Extraction (FCE) structure [Kop05] and the Controlled Injection of Backside Holes (CIBH) structure [Chm06]. Since there are additional p-layers at the cathode side, no electric field can build up between the plasma and said layers. It can be achieved that the plasma does not detach from the cathode. Instead of Eq. (5.93), $|v_r| = 0$ applies and from Eq. (5.94) results $w_x = w_B$. This extends the voltage range of soft recovery. Moreover, it was shown for the CIBH-diode that, if at the end of the reverse recovery process the space charge punches to the n$^+$-layer (see Eq. (5.104)), the backside p-layers inject additional holes and damps possible oscillations [Fel08].

Since these structures were developed with the intention to increase the capability of the diode to withstand dynamic avalanche, they will be described in more detail in Sect. 13.4.

5.8 Outlook

For the soft-recovery behavior of fast diodes, appropriate solutions have been found in the voltage range smaller than 2000 V. Even though further optimization is possible, there are indications that the design is already approaching the limits of what is possible for pin-diodes in silicon. Still there is a potential of hybrid structures for improving diodes.

For the voltage range of 3000 V and more, considerable work still has to be done to realize diodes with satisfactory reverse-recovery behavior in applications with very high power. These applications combine switching slopes, which are much steeper than formerly occurring using thyristors and GTOs, with a significant parasitic inductance in the commutation circuit. For these applications the devices

have to be optimized to avoid voltage peaks and oscillations even under such hard conditions. Furthermore, dynamic ruggedness is very important; this will be explained in Chap. 12.

Regarding applications with high switching frequencies, Schottky diodes are the better choice. Schottky-diodes from GaAs are available as single diodes for 300 V, and they may be developed to reach the voltage range of 600 V. Though Schottky diodes in SiC have already been established for the range of 600 V and 1200 V (see Chap. 6), they can also be designed for a higher voltage range. Because of problems involving material quality and defects, SiC devices are still available only with comparatively small area. For applications in the wide field of motor drives, they should become available in sufficient area (of up to 1 cm^2) to carry enough current per device and at costs that are competitive to silicon devices.

Moreover, there are encouraging results in research regarding pin-diodes in SiC; in particular, here blocking voltages far above 10 kV are possible with a single diode. First results seem to show that the physics of reverse recovery, which was previously explained, can be used in a similar way for analysis and optimization of SiC pin-diodes [Bar07]. However, applications in the high-power range require high currents and, thus, large device areas too.

Therefore, diodes from silicon will probably dominate the market for a long time, and still further work for optimizing them is necessary. In the field of fast diodes, Si and SiC will most likely exist in parallel for some time.

References

[Bab08] Baburske, R., Heinze, B., Lutz, J., Niedernostheide, F.J.: Charge-carrier plasma dynamics during the reverse-recovery period in p+-n–n+ diodes. IEEE Trans. Electron Device **ED-55**(8), 2164–2172 (2008)

[Bab10] Baburske, R., Lutz, J., Schulze, H.-J., Siemieniec, R., Felsl, H.P.: A new Diode Structure with Inverse Injection Dependency of Emitter Efficiency (IDEE), Proceedings of the ISPSD Hiroshima, p 165–168 (2010)

[Bal87] Baliga, B.J.: Modern Power Devices. Wiley, New York (1987)

[Bal98] Baliga, B.J.: Power devices. In: Sze, S.M. (ed.) Modern Semiconductor Device Physics, Wiley, New York (1998)

[Bar07] Bartsch, W., Thomas, B., Mitlehner, H., Bloecher, B., Gediga, S.: SiC-powerdiodes: design and performance. Proceedings European Conference on Power Electronics and Applications EPE (2007)

[Ben67] Benda, H.J., Spenke, E.: Reverse recovery process in silicon power rectifiers. Proc. IEEE, **55**(8) (1967)

[Chm06] Chen M, Lutz J, Domeij M, Felsl HP, Schulze, HJ: A novel diode structure with Controlled Injection of Backside Holes (CIBH). Proceedings of the ISPSD, Neaples. pp. 9–12 (2006)

[Coo83] Cooper, R.N.: An investigation of recombination in Gold-doped pin rectifiers. Solid-State Electron. **26**, 217–226 (1983)

[Deb96] Deboy, G., et al.: Absolute measurement of carrier concentration and temperature gradients in power semiconductor devices by internal IR-Laser deflection. Microelectron. Eng. **31**, 299–307 (1996)

[Dru01] Drücke, D., Silber, D.: Power Diodes with Active Control of Emitter Efficiency. Proceedings of the ISPSD, Osaka, pp. 231–234 (2001)

[Dru03] Drücke, D.: Neue Emitterkonzepte für Hochspannungsschalter und deren Anwendung in der Leistungselektronik, Dissertation, Bremen (2003)

[Fel04] Felsl, H.P., Falck, E., Pfaffenlehner, M., Lutz, J.: The influence of bulk parameters on the switching behavior of FWDs for traction application. Proceedings Miel 2004, Niš/ Serbia & Montenegro (2004)

[Fel08] Felsl, H.P., Pfaffenlehner, M., Schulze, H., Biermann, J., Gutt, T.,. Schulze, H.J., Chen, M., Lutz, J.: The CIBH diode – great improvement for ruggedness and softness of high voltage diodes. ISPSD 2008, Orlando, Florida, pp. 173–176 (2008)

[Hal52] Hall, R.N.: Power rectifiers and transistors. Proc IRE **40**, 1512–1518 (1952)

[Hua94] Huang, Q.: MOS-Controlled Diode – A New Class of Fast Switching Low Loss Power Diode" VPEC, pp. 97–105 (1994)

[Hua95] Huang, Q., Amaratunga, G.A.J.: MOS Controlled Diodes – A New Power Diode. Solid State Electr. **38**(5), 977–980 (1995)

[IXY00] IXYS data sheet FMD 21–05QC (2000)

[Kop05] Kopta, A., Rahimo, M.: The Field Charge Extraction (FCE) Diode – A Novel Technology for Soft Recovery High Voltage Diodes. Proceedings of ISPSD Santa Barbara. pp. 83–86 (2005)

[Las00] Laska, T., Lorenz, L., Mauder, A.: The Field Stop IGBT Concept with an Optimized Diode. Proceedings of the 41th PCIM, Nürnberg (2000)

[Lut94] Lutz, J., Scheuermann, U.: Advantages of the New Controlled Axial Lifetime Diode. Proceedings of the 28th PCIM, Nuremberg (1994)

[Lut00] Lutz, J., Wintrich, A.: The hybrid diode – mode of operation and application. Eur. Power Electron. Drives J. **10**(2) (2000)

[Lut02] Lutz, J., Mauder, A.: Aktuelle Entwicklungen bei Silizium-Leistungs-dioden. ETG-Fachbericht 88, VDE-Verlag Berlin (2002)

[Mou88] Mourick, P.: Das Abschaltverhalten von Leistungsdioden, Dissertation, Berlin (1988)

[Nem01] Nemoto, M., et al.: Great improvement in IGBT turn-on characteristics with trench oxide PiN Schottky Diode. Proceedings of the ISPSD, Osaka (2001)

[Sco69] Schlangenotto, H., Gerlach, W.: On the effective carrier lifetime in psn-rectifiers at high injection levels. Solid-State-Electron. **12**, 267–275 (1969)

[Sco79] Schlangenotto, H., Maeder, H.: Spatial composition and injection dependence of recombination in silicon power device structures. IEEE Trans. El. Dev. **26**(3), 191–200 (1979)

[Sco82] Schlangenotto, H., Silber, D., Zeyfang, R.: Halbleiter-Leistungsbauelemente - Untersuchungen zur Physik und Technologie. Wiss. Ber. AEG-Telefunken **55 Nr.**, 1–2 (1982)

[Sco89] Schlangenotto, H., et al.: Improved recovery of fast power diodes with self-adjusting p emitter efficiency. IEEE El. Dev. Lett. **10**, 322–324 (1989)

[Shm82] Shimada, Y., Kato, K., Ikeda, S., Yoshida, H.: Low input capacitance and low loss VD-MOSFET rectifier element. IEEE Trans. Electron Dev. **29**(8), 1332–1334 (1982)

[Sil85] Silber, D., Novak, W.D., Wondrak, W., Thomas, B., Berg, H.: Improved Dynamic Properties of GTO-Thyristors and Diodes by Proton Implantation. IEDM, Washington (1985)

[SYN07] Advanced tcad manual. Synopsys Inc. Mountain View, CA. Available: http://www.synopsys.com (2007)

[Sze81] Sze, S.M.: Physics of Semiconductor Devices. Wiley, New York (1981)

[Wol81] Wolley, E.D., Bevaqua, S.F.: High speed, soft recovery epitaxial diodes for power inverter circuits. IEEE IAS Meeting Digest (1981)

[Won87] Wondrak, W., Boos, A.: Helium implantation for lifetime control in silicon power devices. Proc. of ESSDERC 87, Bologna, pp. 649–652 (1987)

Chapter 6
Schottky Diodes

Schottky diodes are formed by a metal-semiconductor junction, whose rectifying behavior was described early [Sch38, Sch39]. They are unipolar devices, which means that only one type of carriers is available for the current transport. If they are designed for large blocking voltages, the resistance of the base will increase strongly due to the lack of charge carrier modulation. Schottky power diodes have been used for a long time, but in the last years they have gained an increased importance in following areas:

- Si Schottky diodes in the voltage range up to approximately 100 V to be used as freewheeling diodes for MOSFETs. Their advantage is the low junction voltage and the absence of stored charge. At turn-off from conducting to blocking mode, only the capacitive recharge of the junction capacitance needs to be considered. This makes them very useful for very high switching frequencies.
- Schottky diodes from semiconductor materials with wide bandgap. With these materials, much higher blocking voltages are possible due to the higher critical fields. Because of the much smaller possible junction voltage compared with wide-gap pn-diodes, Schottky diodes have become an attractive alternative.

6.1 Energy Band Diagram of the Metal-Semiconductor Junction

Figure 6.1 shows the band structure of a contact between a metal and an n-type semiconductor. We focus to this case which because of the higher mobility of electrons than of holes is solely used for power Schottky diodes. An ideal contact with negligible influence of surface states is considered. The diagram is determined by the work functions Φ_M and Φ_S of the metal and the semiconductor and the electron affinity χ of the semiconductor. The work function of a body not in contact with others is defined as the difference between the vacuum level, i.e. the energy far

© Springer International Publishing AG 2018

J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8_6

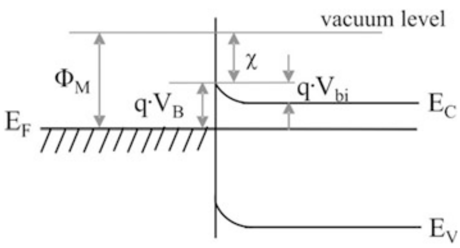


Fig. 6.1 Band diagram of a metal-semiconductor junction in thermal equilibrium. Surface charges are neglected. Figure according to [Spe65]

outside, and the Fermi level; the electron affinity χ of a semiconductor is the difference between the vacuum level and the conduction band edge. Since in a non-degenerate semiconductor the Fermi level is below the conduction band edge, χ is smaller than the work function Φ_S. If now a metal and an n-type semiconductor with $\Phi_S < \Phi_M$ are put into contact, the Fermi level in the semiconductor is initially higher than in the metal, since the vacuum level remains constant (the space outside is field-free in thermal equilibrium). Hence electrons flow from the semiconductor into the metal leaving a positive space charge region in the semiconductor formed by depleted donors. This is connected with a bend-up of the bands and continues until equilibrium is reached and the Fermi level is constant throughout the structure, see Fig. 6.1. At the interface then the following energy step occurs:

$$qV_B = \Phi_M - \chi \tag{6.1}$$

This barrier, which electrons moving from the metal to the semiconductor have to overcome, is called contact barrier or Schottky barrier. In the semiconductor, the bend-up of the bands means that a built-in voltage V_{bi} is formed over the space charge region:

$$qV_{bi} = \Phi_M - \Phi_S = \Phi_M - (\chi + (E_C - E_F)) \tag{6.2}$$

Compared with the isolated semiconductor the electron energy in the neutral part of the n region is reduced by qV_{bi}, the neutral n region is on a positive potential compared to the metal. Although electrons moving from the semiconductor to the metal have to surmount a smaller barrier (qV_{bi}) than electron moving from the metal to the semiconductor, the two currents compensate each other in thermal equilibrium, since the relevant carrier density in the conduction band due to its distance from the Fermi level is smaller than in the metal.

The Eqs. (6.1) and (6.2) hold also if $\Phi_M < \Phi_S$. Then no depletion layer is formed in the semiconductor, but instead an enhancement layer of electrons. According to (6.2) the built-in voltage is negative then. For $\Phi_M = \chi$ the Schottky barrier qV_B vanishes after (6.1). In this case, the contact behaves like an ohmic contact, since the contact resistance $R_c = (dj/dV)^{-1}$ is very small (see [Sze81], p. 304). Usual *ohmic contacts* differ from Schottky contacts (i) by a large number of interface states with levels in the bandgap resulting in a high surface recombination velocity, which in Schottky junctions is negligible; (ii) usual ohmic contacts are formed with

a highly doped semiconductor with a doping > 5×10^{18} cm^{-3}. In this range the barrier is so thin that charge carriers can easily tunnel through.

6.2 Current-Voltage-Characteristics of the Schottky Junction

If a voltage V is applied to the metal relative to the semiconductor, the potential difference V_{bi} between the semiconductor and the metal is reduced for $V > 0$ and enhanced for $V < 0$. This is connected with a narrowing or widening of the space charge region in the semiconductor, respectively. The barrier for the electrons flowing from the semiconductor to the metal is changed to $q(V_{bi} - V)$, whereas the contact barrier qV_B to be surmounted by electrons flowing from the metal to the semiconductor remains constant. Hence for $V > 0$ the electron current flowing from the semiconductor to the metal soon predominates strongly, the junction is conducting. For $V < 0$, the voltage-independent electron current flowing from the metal to the semiconductor predominates, and since this current is small due to the higher barrier qV_B, the junction is blocking in this direction. Assuming that the electrons obey the Boltzmann distribution and both the currents, from the semiconductor to the metal and from the metal to the semiconductor, are proportional to the electron concentration at the interface, the I–V characteristic is obtained as [Sze81]:

$$j = j_s \cdot \left(e^{\frac{q \cdot V}{k \cdot T}} - 1 \right) \tag{6.3}$$

In this general form the equation is identic with Eq. (3.50) of the ideal pn-junction, see also Fig. 3.10. However the saturation current density j_s is quite different, namely given by

$$j_s = A^* \cdot T^2 \cdot e^{-\frac{q \cdot V_B}{kT}} \tag{6.4}$$

where qV_B is the contact barrier defined in (6.1) and A^* a semiconductor specific constant, the so called effective Richardson constant. The characteristic is independent of the bandgap and doping density of the semiconductor. For a metal on n-semiconductor contact, V is the voltage applied to the metal relative to the semiconductor, for a metal contact on a p-semiconductor the polarity is inverse.

The effective Richardson constant for some semiconductors is given in Table 6.1. In the inset table of Fig. 6.2, the contact barriers of important contact metals for Si Schottky diodes are compiled. Mainly silicides are used for the contact, since these facilitate tuning and optimization of the contact properties. The saturation current densities calculated by Eq. (6.4) are given also, to show the very strong increase with decreasing barrier height. j_s is in all cases several orders of magnitude higher than for pn-junctions in Si. The shown I–V characteristics obtained with these data illustrate on the other hand that the threshold voltage

decreases with decreasing V_B. That the threshold voltage for listed contact metals is significantly smaller than for pn-junctions, is the reason why Schottky junctions are applied at all.

Similar to pn-junctions the saturation current increases strongly with temperature. For Cr_2Si for instance the threshold voltage (if defined as the voltage at 5 A/cm^2) is only 0.2 V, but the leakage current is about 2 mA/cm^2 at 300 K and about 1 A/cm^2 at 400 K. The Cr_2Si Schottky contact is therefore only suited for very low blocking voltages, e.g. for diodes in switched mode power supplies for low voltages. For 100 V Schottky diodes, typically PtSi is used as contact material.

In blocking direction the ideal I–V characteristics according to (6.3), (6.4) is observed only for relatively small voltages. For larger reverse biasing the blocking current is enhanced by the image force [Sze81]. This leads to a lowering of the contact potential by

$$\Delta\Phi = \sqrt{\frac{q \cdot \mathsf{E}}{4 \cdot \pi \cdot \varepsilon}} \tag{6.5}$$

where E is the field strength at the contact. Hence instead of (6.4) the saturation current density at higher reverse voltage is given by:

$$j_s = A^* \cdot T^2 \cdot e^{-\frac{q \cdot (V_B - \Delta\Phi)}{kT}} \tag{6.6}$$

Table 6.1 Effective Richardson constant A^* for different semiconductors of n-type

Si	110 A/(cm^{2}K^2)	[Sze81]
GaAs	8 A/(cm^{2}K^2)	[Sze81]
SiC	400 A/(cm^{2}K^2)	[Tre01]

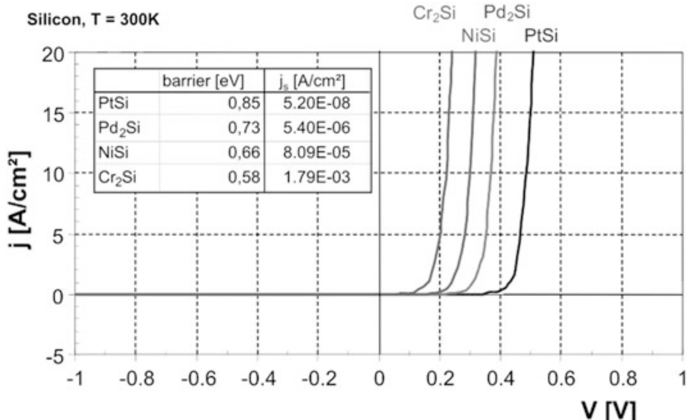


Fig. 6.2 Calculated I-V characteristics for Si Schottky diodes for different contact materials according to data from berndes, IXYS semiconductor GmbH [Ber97]. The characteristics and saturation current densities in the inset refer to 300 K

The increase of $\Delta\Phi$ with the applied reverse voltage is the main reason for the "soft" reverse characteristics of Schottky diodes. The effect becomes especially noticeable for Schottky diodes with low barrier heights or Schottky contacts on wide-gap semiconductors. At $E = 2 \times 10^5$ V/cm the barrier lowering amounts to 50 meV.

In forward direction, Eq. (6.3) is valid at higher current densities only, if the voltage V is defined as the voltage across the junction, excluding the resistive voltage drop over the neutral semiconductor. The whole forward voltage as function of j is given at higher current densities as the sum of the junction voltage obtained from (6.3) and the ohmic voltage drop:

$$V_F = \frac{kT}{q} \ln\left(\frac{j}{j_s} + 1\right) + r_\Omega j \tag{6.7}$$

where $r_\Omega = R_\Omega \cdot A$ is the resistance times active area A. Since the resistance of unipolar devices is constant, the I–V characteristic is completely given by (6.7). Because conductivity modulation does not take place, the resistance is much larger than for a pin diode with equal breakdown voltage. R_Ω is called also on-resistance, since relevant only for the forward voltage or the on-state. In Sect. 6.4 R_Ω is discussed in detail.

Also an ideality factor n is sometimes introduced in the exponent of Eq. (6.3) to obtain a better agreement with experimentally measured values:

$$j = j_s \cdot \left(e^{\frac{q \cdot V}{n \cdot k \cdot T}} - 1\right) \tag{6.6b}$$

The value for n is between 1 and 2, but for good Schottky diodes a value of 1.02–1.06 can be obtained. The main reasons for a non-ideal behavior, i.e. $n \neq 1$, are states at the interface.

6.3 Structure of Schottky Diodes

Figure 6.3 shows the basic structure of a Schottky diode. The low doped n^- layer, which must sustain the reverse voltage, is grown on the n^+ substrate by epitaxy. Real Schottky diodes need additional junction termination structures, not shown in the figure. Commonly used terminations are field plates, the diffusion of potential rings, or a JTE-structure. Sometimes also combinations of field plate and potential ring or JTE-structures are implemented. For details of these junction terminations, see Chap. 4.

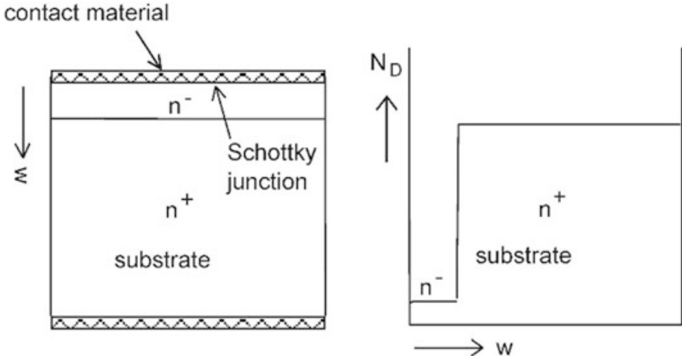


Fig. 6.3 Basic structure of a Schottky diode

6.4 Ohmic Voltage Drop of a Unipolar Device

The resistance of the low-doped middle layer in a unipolar device is given by the following expression:

$$R_\Omega = \frac{w_B}{q \cdot \mu_n \cdot N_D \cdot A} \tag{6.8}$$

where as previously w_B denotes the thickness of the layer, N_D the doping concentration (carrier density), μ_n the electron mobility and A the active area of the device. Because with increasing blocking voltage the required base width increases and the allowed doping concentration decreases, the resistance increases strongly with breakdown voltage. For a triangular field distribution, the base width for a breakdown voltage V_{BD} is $w_B = 2V_{BD}/\mathsf{E}_c$, the maximal doping concentration according to (3.74) is $N_D = \varepsilon \mathsf{E}_c^2/(2qV_{BD})$, where E_c is the critical field of the semiconductor. Substituting this into (6.8) one obtains for the resistance times area.

$$r_\Omega = R_\Omega A = \frac{4V_{BD}^2}{\mu_n \mathsf{E}_c^3} \tag{6.9}$$

If the dependence of μ_n on N_D and hence on V_{BD} is left out of account, the resistance is proportional to the square of the breakdown voltage, the material parameter E_c enters with the third power. The latter is the cause why the on-resistances of unipolar devices of different semiconductors such as Si and SiC is extremely different. The critical field decreases slightly with breakdown voltage as given by Eq. (3.88). Inserting (3.88) one obtains:

$$r_\Omega = \frac{1}{\varepsilon \mu_n} \left(\frac{B}{n+1} \right)^{\frac{3}{n-1}} (2V_{BD})^{\frac{2n+1}{n-1}} \tag{6.10}$$

Due to the decrease of E_c the resistance increases yet stronger than with the square of V_{BD}. This holds even taking the increase of μ_n with V_{BD} into account. The mobility is obtained inserting the doping concentration given by (3.87) as function of the voltage into the formula (2.35) using the parameters of Appendix A. If the real base width is equal to the unhindered extent of the space charge region (NPT-case), the resistance is necessarily given by (6.8) and, as far as (3.88) is applicable, by Eq. (6.10).

Passing to PT-design with trapezoidal field shape, the resistance for a given V_{BD} changes, since w_B and N_D will decrease differently from their NPT-values. At moderate PT the resistance has a minimum, as can be calculated expressing N_D from (5.7) by w_B assuming E_c as independent of N_D and inserting it into (6.8). With μ_n assumed constant also, differentiation for constant V_{BD} yields that R_Ω is minimal at [Hu79]

$$w_B = \frac{3}{4} \left\{ \frac{2V_{BD}}{E_c} \right\}, \qquad N_D = \frac{8}{9} \left\{ \frac{\varepsilon E_c^2}{2qV_{BD}} \right\} \tag{6.11}$$

This means that the field E_w at the nn^+-junction is one third of the critical field: $E_w = E_c/3$. With (6.11) in (6.8) the resistance is reduced to 27/32 = 0.844 of the value given by (6.9) for NPT-design.

The *dependence* of the resistance on N_D or w_B for a given breakdown voltage [Dah01] can be calculated solving Eq. (5.7) first for w_B which yields:

$$w_B = \frac{2V_{BD}}{E_c + \sqrt{E_c^2 - 2\frac{q}{\varepsilon} V_{BD} N_D}} \tag{6.12}$$

The equation holds up to the N_D-value for NPT, where the square root term vanishes. E_c depends on N_D according to Eq. (3.78)

$$E_c = \left(\frac{q(n+1)N_D}{\varepsilon B} \right)^{\frac{1}{n+1}}$$

which holds also for moderate PT with $E_w \lesssim E_c/2$, see Sect. 5.3. Substituting this for E_c in (6.12) and inserting the equation into (6.8) one obtains the resistance as function of N_D for a given V_{BD}. The result for a Si Schottky diode with 240 V blocking voltage is shown in Fig. 6.4. Besides the resistance r_Ω also the base width is plotted versus N_D. For a good approximation of α_{eff} for this voltage the exponent $n = 6$ was used, which according to (3.83) leads to $B = 5.45 \times 10^{-30}$ $\mathrm{cm^{n-1}/V^n}$.

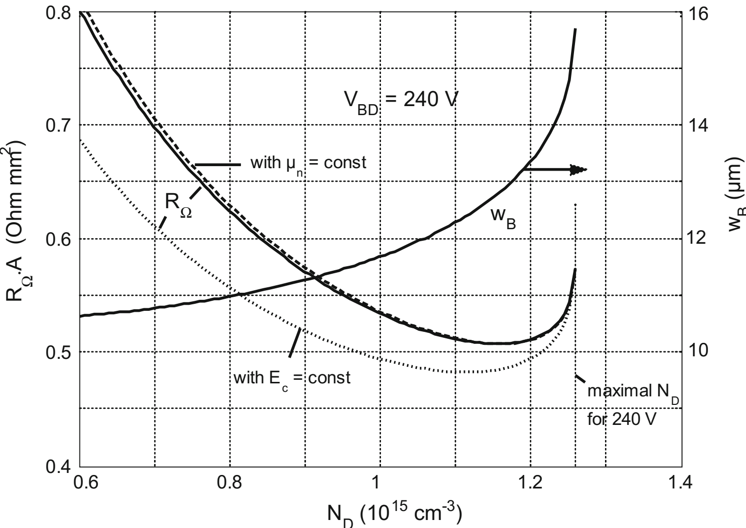


Fig. 6.4 Calculated resistance R_{Ω} and base-width w_B for a 240-V silicon Schottky diode in dependency of the base doping

In the first range below the maximal doping (NPT-doping), w_B decreases very strongly, thus overcompensating the decrease of N_D. For r_{Ω} three lines are shown: The solid line shows the exact result allowing not only for the variation of E_c but also the dependence of μ_n on N_D. The dashed curve is calculated with variable E_c but constant mobility μ_n using the value 1360 $cm^2/(Vs)$ for the doping concentration at the minimum of R_{Ω}. As is seen, the variation of μ_n has not a significant effect in the considered range. For the dotted curve the critical field was assumed constant as for the Eq. (6.11) using the NPT-value given by (3.88) (μ_n being variable). As the Figure shows, the resistance obtained with this approach is clearly too small. According to the solid curve the lowest possible resistance of a 240 V-device is $0.508\ \Omega\ mm^2 = 0.89$ times the resistance in NPT-layout. Furthermore, the N_D-value of minimal R_{Ω} is located closer to the N_D-value for NPT than according to (6.11).

This approach yields a considerably smaller resistance than more correctly calculated with variable E_c (solid curve). The calculation for other blocking voltages leads to similar results. This is supported by analytical expressions for the minimum resistance and the N_D-value of the minimum, taking the variation of E_c into account (in contrast to (6.11)). From Eqs. (5.7) and (3.78) one determines first the derivative dw_B/dN_D for constant V_{BD} and uses this for the differentiation of (6.8). The result of this somewhat lengthy calculation is that for minimal R_{Ω} the field strength E_w at the nn^+ junction in relation to the critical field E_c (at the metal semiconductor junction) must obey the condition:

$$\frac{\mathsf{E}_w}{\mathsf{E}_c} = \frac{n-1}{n+1}\,\frac{1}{3} \tag{6.13}$$

The field ratio varies from 1/3 for large n (constant E_c) over 1/4 for n = 7 to 1/5 for $n = 4$, an exponent value appropriate for 50 V-devices. The decrease of E_c with decreasing N_D moves the minimum nearer to NPT. From the known ratio $\mathsf{E}_w/\mathsf{E}_c$ one can determine the doping concentration and base width at the minimum of R_Ω and hence the minimal resistance itself. The results are in close agreement with Fig. 6.4. The factor between the minimum resistance and the NPT-value is nearly independent of the blocking voltage. For Si-devices the minimum resistance is (88 ± 2) % of the NPT-value. Hence using Eq. (6.10) with $n = 7$ one obtains for the minimal resistance of unipolar Si devices:

$$R_{\Omega,\min} = 0.88 \cdot \frac{2 \cdot B^{\frac{1}{2}} \cdot V_{BD}^{\frac{5}{2}}}{\mu_n \cdot \varepsilon \cdot A} \tag{6.14}$$

Figure 6.5 shows the minimal resistance and the appropriate width of the middle layer of a unipolar Si-device as a function of the breakdown voltage V_{BD}. The NPT-resistance was calculated from (6.10) with $n = 6$ for this voltage range and is multiplied then with the factor 0.88 as described above. The base width was calculated multiplying the NPT-value by 0.817, a factor derived from Eq. (6.13) together with (3.78) (see also Fig. 6.4). The resistance is growing in this range nearly with the power of 2.5 of the breakdown voltage. As a result, unipolar devices for high blocking voltage will have a very high on-state resistance.

For silicon Schottky diodes, often a deviation from the unipolar resistance behavior is observed. To achieve sufficient blocking capability, a p-doped potential ring is usually implemented at the edge of the active area, similar to Fig. 4.24 in Chap. 4. In this case a pn-junction is connected in parallel to the Schottky junction, and at a forward voltage in the range of the diffusion voltage of the pn-junction this junction starts to inject charge carriers which reduce the resistance. Because the

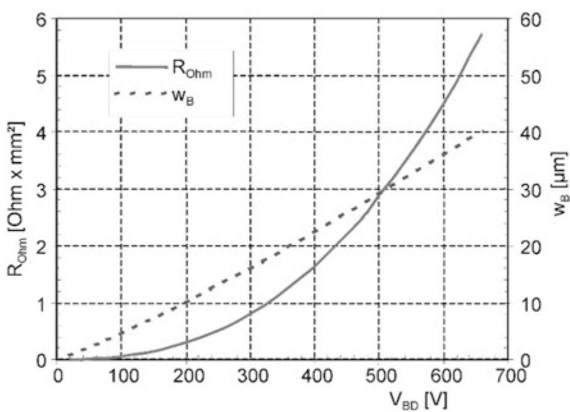


Fig. 6.5 Minimum resistance and appropriate width of the middle layer of a unipolar Si-device in dependence of breakdown voltage

injection increases with temperature owing to the decrease of the threshold voltage, a decrease of the resistance with temperature is found in such cases, whereas an increase is expected from the decrease of the mobility. For Schottky diodes based on SiC, the pn-junction of which begins later to inject, the resistance shows the temperature dependence as determined by the mobility μ_n. Details on this issue will be discussed below, see Fig. 6.9.

6.4.1 Comparison of Silicon Schottky Diodes and pin Diodes for Rated Voltages of 200 and 100 V

Because some spread of data is to be expected from tolerances in used materials and fabrication processing steps (roughly 10% as a role of thumb) and to have some safety for tolerances in the measurement technique (up to 10%), the calculation for a 200 V rated diode is done for a breakdown voltage of 240 V. From Eq. (6.15) a value for $R_{\Omega}{\cdot}A$ = 0.51 Ω.mm^2 is obtained. At a current density of 1.5 A/mm^2 – a typical current density at rated current – a voltage drop across the epi region of 0.77 V results. As Schottky contact material PtSi can be used. This material leads to a threshold voltage of 0.5 V, and therefore a forward voltage $V_F = V_S + R_{\Omega}{\cdot}A{\cdot}j$ = 1.27 V is to be expected. In practice, however, less than 0.9 V is measured. This discrepancy can be explained by the bipolar effect, as discussed above.

In comparison, a fast epitaxial pin diode designed for a rated voltage of 200 V can be fabricated so that a forward voltage smaller than 1 V is reached at the assumed current density, although the junction voltage is in the range 0.7 – 0.8 V. The stored charge to be extracted during commutation is comparable to the capacitive charge of the Schottky diode.

For a rated voltage of 100 V, however, the Schottky diode designed with similar tolerances shows a resistance $R_{\Omega}{\cdot}A$ = 0.082 Ω mm^2. With the PtSi barrier again the resulting on-state voltage is $V_F = V_S + R_{\Omega}{\cdot}A{\cdot}j$ = 0.62 V. Such a low value cannot be reached with a pin-diode. For still lower voltages, the advantage of the Schottky diode compared to a pin diode will be even higher.

6.5 Schottky Diodes Based on SiC

6.5.1 SiC Unipolar Diode Characteristics

The first SiC Schottky diodes were developed by the German company SiCED and introduced on the market by Infineon for rated voltages of 600 and 1200 V. SiC Schottky diodes are superior to solutions using bipolar Si-diodes in all applications that require a high switching frequency (switched mode power supplies, power factor correction, etc.) [Zve01].

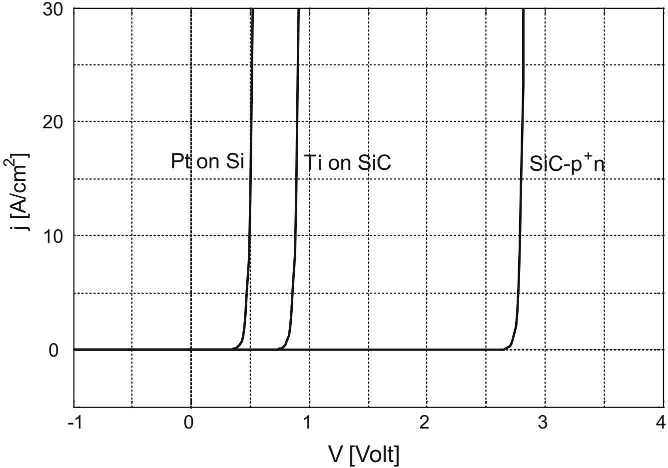


Fig. 6.6 Ideal low-current characteristics of different junctions based on SiC and Si

The used SiC polytype is 4H which has a bandgap of 3.26 eV. For a SiC pn-junction the high bandgap leads to a threshold voltage of about 2.7 V, as can be seen from Fig. 3.11. The high forward voltage drop is a clear disadvantage of SiC pn- and pin-diodes, which is why diodes in SiC are usually Schottky-diodes or Merged pin Schottky diodes. In Fig. 6.6 the forward characteristic of a SiC-Schottky diode with a Ti-contact, which has a contact barrier qV_B = 1.27 eV, is compared in a range of lower current densities with characteristics of a SiC pn-junction (parameters as in Fig. 3.11) and a Si-Schottky junction with PtSi-contact (qV_B = 0.85). The ideal characteristics are assumed with effective Richardson constants taken from Table 6.1. The advantage of the SiC-Schottky-junction against the pn-junction is obvious.

For Schottky-junctions on SiC typically metals with a higher contact barrier are used than for Si, because the higher field at reverse bias results in a strong lowering of the barrier by the image force. The Ti-contact is often used for high-voltage Schottky diodes.

One of the main advantages 4H-SiC compared to silicon is a well ten times higher critical field strength for avalanche breakdown. As has been expressed in Eq. (2.101), the effective ionization rate is found to obey the power dependency $\alpha_{eff} = B|\mathsf{E}|^n$ with n = 8.03 and $B = 2.185 \times 10^{-48}$ cm^{n-1}/V^n [Bar09]. On the base of these parameters, specified Eqs. (3.91) and (3.92) have been given for the critical field and breakdown voltage as functions of N_D. The latter relationship has been compared in Fig. 3.19 with that of silicon. Equations containing n and B explicitly such as (3.86–3.88) are of course also immediately applicable to SiC. At V_B = 1500 V Eq. (3.88) yields for the critical field of 4H-SiC $\mathsf{E}_c = 2.64\,MV/cm$, which is a factor 11.8 higher than obtained for Si with n = 7, $B = 2.107 \times 10^{-35}$ cm^{n-1}/V^n. This passes in a reciprocal manner to the base width $w_B = 2V_{BD}/\mathsf{E}_c$ which is by the same factor *smaller* than for a silicon device. A plot

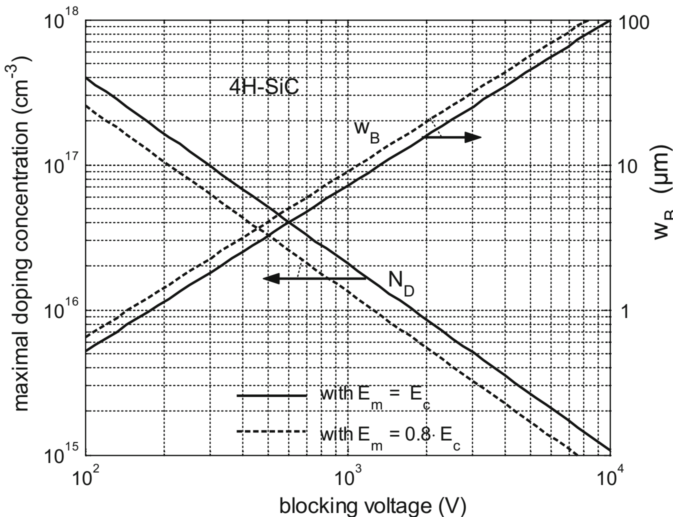


Fig. 6.7 Doping concentration and base width versus usable blocking voltage for a SiC-diode with triangular field distribution. For the solid lines the allowed maximum field E_m is put equal to the critical field E_c, for the dashed curves E_m is put equal to 0.8 E_c

of the doping concentration and base width versus the "usable" blocking voltage is shown in Fig. 6.7. The solid curves are obtained if the usable blocking voltage is equated with the breakdown voltage V_{BD} ($\equiv V_B$). For the dashed curves the usable blocking voltage is defined as the voltage at which the field reaches 80% of the critical field. This corresponds to practical guidelines for the design, which have to take into account that SiC-devices exhibit soft reverse characteristics with substantially increasing current appreciably below the breakdown voltage, as shown further below in Fig. 6.9. Therefore the field strength is limited to a value considerably below E_c. The respective curves in Fig. 6.7 were calculated replacing the critical field in $w_B = 2V_{BD}/E_c$ and $N_D = \varepsilon E_c^2/(2qV_{BD})$ by 0.8· E_c and substituting then E_c by (3.88). Even with such cutback, the necessary base width of SiC devices is nearly an order of magnitude smaller and the allowed doping density N_D two orders higher than for silicon (see Fig. 3.17).

Both inserted into (6.7) results in a unipolar on-resistance which is nearly three orders of magnitude smaller than for Si-devices. This follows already from Eq. (6.8) because of the one order higher critical field. For a definite calculation of R_Ω one needs the electron mobility. According to [Scr94] (see Appendix A2) the electron mobility in 4H-SiC depends at 300 K in the following manner on the doping density:

$$\mu_n = \frac{947}{1 + \left(\frac{N_D}{1.94 \cdot 10^{17} cm^{-3}}\right)^{0.61}} \frac{cm^2}{Vs} \tag{6.15}$$

μ_n is somewhat smaller than in Si, and since the doping of SiC-devices is higher, its decrease with increasing doping (decreasing V_{BD}) comes stronger into play. Substituting N_D in (6.15) by (3.87) and inserting the obtained μ_n into (6.9), the resistance is obtained as function of V_{BD}. This holds directly for the case of NPT. As discussed above, a slight reduction of R_Ω is possible by PT-layout. To obtain the minimal resistance attainable the NPT-value is to be multiplied by a factor which is nearly independent of the blocking voltage, see Sect. 6.4. Calculations like those for Fig. 6.4 using $n = 8.03$ yield that this factor is $0.868 \pm 1\%$ for the whole voltage range above 500 V. Hence multiplying (6.10) with this factor and inserting the mentioned n-value one obtains for the minimal resistance of SiC-devices:

$$R_{\Omega,\min} = 0.87 \cdot \frac{2^{2.43} \cdot \left(\frac{B}{9.03}\right)^{0.43} \cdot V_{BD}^{2.43}}{\mu_n \cdot \varepsilon \cdot A} \tag{6.16}$$

The minimal resistance $R_{\Omega,min}$ obtained in this way is shown in Fig. 6.8 as function of V_{BD} (using the relative dielectric constant $\varepsilon_r = 9.66$). For comparison also the equal dependency for silicon is depicted. The curves are approximately straight lines on this double-logarithmic scale. The resistance for SiC is at 1200 V by a factor 740 lower than for Si, the blocking voltage for a given R_Ω is well a decade higher. The allowed field has been assumed in the calculations to be equal to E_c. The curves represent hence lowermost limits for R_Ω which are referred to as the silicon carbide respectively silicon 'unipolar limits'. As shown by the experimental points depicted in the figure, practical SiC-devices do not reach this limit. The resistance of the purely unipolar Schottky diode taken from [Pet01] lies an order above the theoretical limit.

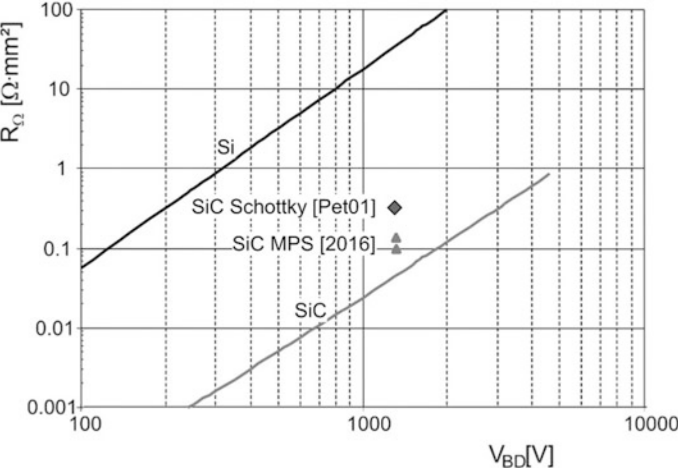


Fig. 6.8 Unipolar resistance $R_{\Omega,\ min}$ for Si and SiC as function of the breakdown voltage. Measurements from [Pet01] and actual MPS diodes from 2016 production are indicated for comparison

The graphs for the “unipolar limit” vary in the literature by some small amount. Possibly some authors use E_c as constant and do not consider its doping dependency. The field shape and doping are taken into account here by using the respectively approach for the ionization rates for Si and for SiC.

Figure 6.9, taken from [Pet01], shows the IV characteristics of a SiC Schottky diode rated for 1200 V with 10 mm^2 active area. The characteristics for 25 and 125 °C are shown. The forward characteristic has a resistive behavior, the resistance extracted from the characteristic at 25 °C is depicted in Fig. 6.8 as [Pet01]. The threshold voltage at 25 °C is in agreement with Fig. 6.6. With increasing temperature, the on-resistance increases in accordance with the decrease of the electron mobility. In reverse direction the current shows a significant increase already appreciably below the breakdown voltage, i.e. the characteristics are soft, and the soft behavior increases with temperature. The softness of the characteristic limits the usable blocking voltage which decreases with increasing temperature. In theory the breakdown voltage should increase with the temperature due to the decreasing avalanche coefficients and increasing critical field strength. The softness is probably mainly caused by lattice defects and/or surface defects. An additional cause is the image force introduced by Eqs. (6.5) and (6.6). Because of the high electric field this results in a significant reduction of the Schottky barrier in SiC-devices. For a field of 1 × 10^6 V/cm Eq. (6.6) yields $\Delta\Phi = 0.122$ V, which according to (6.5) results in an enhancement of the saturation current at 300 K by the factor 112, at 400 K by a factor 34.4.

The charge necessary for a given on-conductance is extracted (to a large part) at commutation into the reverse direction and appears as reverse current integral. The resistance and the charge $Q = q \cdot N_D \cdot A \cdot w_B$ are connected by

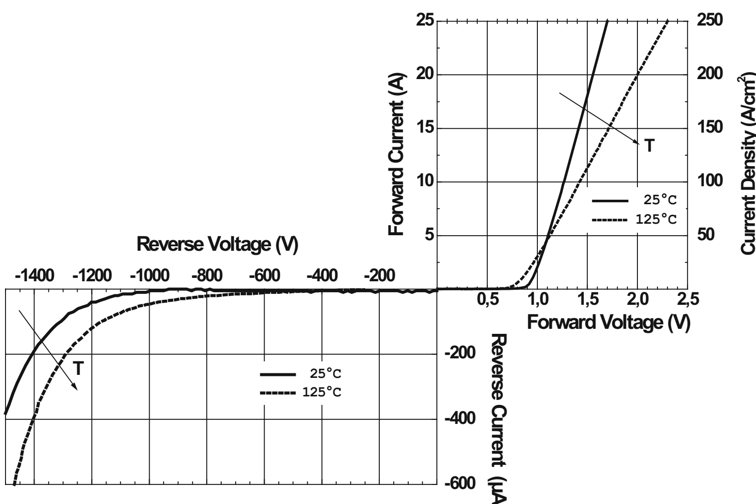


Fig. 6.9 IV-characteristic of a 1200-V SiC Schottky diode with 10 mm^2 active area at 25 °C and at 125 °C. Figure taken from [Pet01]

$$G = \frac{1}{R_\Omega} = \frac{q\mu_n N_D A}{w_B} = \frac{\mu_n Q}{w_B^2} \tag{6.16}$$

The inverse relationship between the on-resistance and the stored charge is similar to pin-diodes. In contrast to the latter the charge of a Schottky-diode does not increase with temperature, which is advantageous with respect to dynamical properties, but disadvantageous with respect to the forward voltage. For SiC devices with their small base width the charge is very small for a given R_Ω. Nevertheless it can be advisable in dimensioning of SiC Schottky-diodes to make the resistance R_Ω not *unnecessarily* small and the charge Q accordingly high. We will come back to the dynamical behavior of SiC-Schottky diodes in Sect. 6.5.3.

6.5.2 Merged Pin Schottky (MPS) Diodes

To reduce the influence the image force and improve the surge current capability, merged pin-Schottky structures (MPS-diodes) have been introduced [Bjo06, Rup06, Sin00]. Most Schottky diodes already contain a p-zone at the edge of the active area for field termination as shown in Fig. 6.10a. In Fig. 6.10b additional high p-doped areas are implemented, which act as p-emitters and inject carriers if the forward voltage drop is above the junction voltage of the SiC pin-diode. This is the case in the range of 2.8 V, as discussed above. An additional solution is the implementation of p-layers in analogy with silicon merged pin-Schottky (MPS) diodes as shown in Fig. 5.28. Such a structure is shown in Fig. 6.10c. These p-layers are separated by a relatively small distance of n^--material with the aim to completely shield the Schottky contact from the high electric field in reverse bias.

The surge-current behavior of these structures was investigated in [Hei08c]. The I–V characteristic of diodes with the three structures in Fig. 6.10 is shown in Fig. 6.11. The structure with p^+-regions (b) shows the lowest forward voltage drop for high current. The effect of the pin-diode is clearly noticeable for currents above

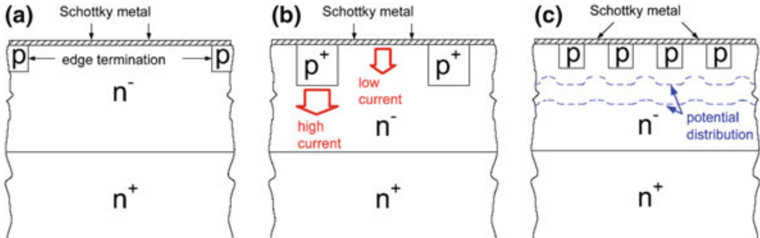


Fig. 6.10 Structure of **a** conventional SiC Schottky-diode, **b** MPS diode structure with p^+-regions to improve the surge-current capability [Bjo06] and **c** MPS diode structure with p-regions to shield the Schottky junction from high electric fields [Sin00]. Figure similar to [Hei08b]

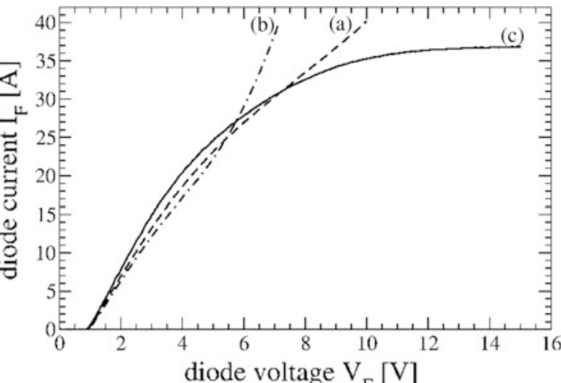


Fig. 6.11 Forward IV-characteristic for SiC diodes for the three structures shown in Fig. 6.10. Specified voltage 600 V, rated current 4 A for all three diodes. Figure similar to [Hei08b]

20 A, when the p-layers inject current. This structure shows a high surge-current capability. A similar effect, although not as strong, can also be seen for structure (a), where the pin-contribution caused by the edge region injects carriers. Structure (c) shows no pin-behavior at all. Although it contains p-layers, they do not inject carriers and it remains unipolar. Moreover, the I–V characteristic deviates from the ohmic behavior at currents above 20 A. At high forward voltage, a significant electric field arises across the middle layer, which has a thickness of only some µm. The increased field will then cause the electron mobility to decrease and will result in an increased forward voltage drop.

A further significant contribution is the temperature increase during measurement. Therefore, the surge-current capability is limited for this structure. However, a structure as displayed in Fig. 6.10c is appropriate to shield the Schottky contact from high electric fields. The highest electric field occurs at the edges of the p-zones. The surface between them is released. Therefore, this structure avoids the reverse bias leakage current of the Schottky junction.

Figure 6.12 shows a top view of the arrangement of p-regions and Schottky zones of a recent MPS-diode [Rup12, Rup14]. The Schottky regions are formed as hexagons, with a pin-region in the middle, with some larger-area pin-regions and with a pin-region of the largest size at the edge of the active area. The total Schottky area is about 50%. The structure combines the advantages of Fig. 6.11b, c.

SiC MPS diodes for 1200 V demonstrate a kind of triggering of the bipolar part at surge-current conditions. Figure 6.13 shows a measurement and electro-thermal simulation of a 1200 V SiC MPS diode with 10A rated current. The current density at rated current is 340 A/cm^2. The simulation simplifies a 10 µm wide 2-dimensional structure with 50% Schottky-area and 10 µm extension since simulation of the complex 3D-structure in Fig. 6.12 is due to extensive computing time, which is not possible, yet.

In the measurement, close to the marked point I, the p-regions start to inject and a branch with negative differential resistance starts to form. In the simulation with a simplified 2D-stucture, this is a pronounced triggering point. The device strongly

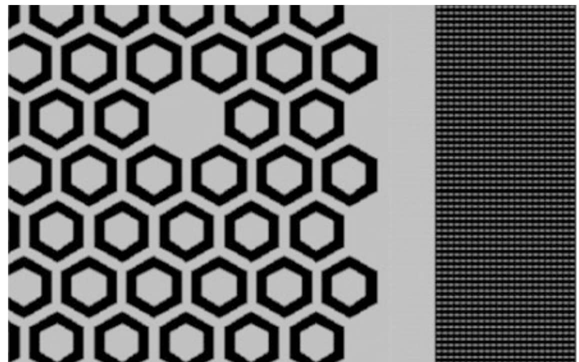


Fig. 6.12 Top view of the structure of an MPS-diode. Black: Schottky regions. Grey: pin-regions. Dotted area on the right side: Junction termination. Figure from R. Rupp, Infineon

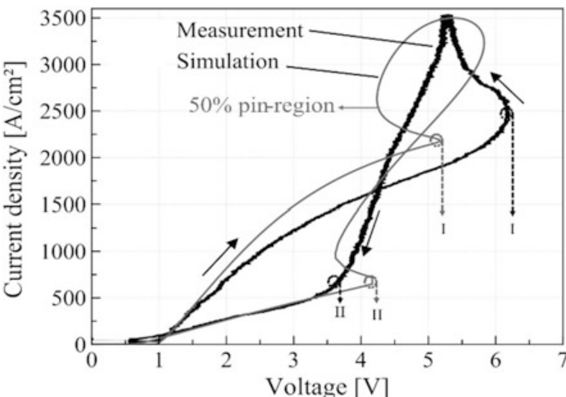


Fig. 6.13 Surge-current pulse, half-sinus 10 ms applied to a 1200 V SiC MPS diode with 10 A rated current, T = 25 °C, measurement and simulation. © 2016 IEEE. Reprinted, with permission, from [Pal16]

heats up and reaches up to 800 K. At high temperature the mobilities strongly decrease, and the descending branch follows another line. In simulation the mobility model with the temperature dependent parameters from [Scr94] is used. The extinction point at II is pronounced in simulation, in the measurement it is blurred due to the more complex structure with different sizes of p-layers.

The behavior of MPS diodes strongly depends on the device geometry, e.g. the size of the pin-area relative to the Schottky area [Fic14, Pal16]. The triggering of the minority carrier injection can be approximated if it is considered that in the Schottky-mode the electron current must flow laterally below the p-layer, see Fig. 6.14.

The voltage drop of lateral current flow across the length L can be calculated according to [Ogu04] as

$$V_p = \int_0^L R \cdot j_p \cdot y \cdot dy = \frac{1}{2} \cdot R \cdot j_p \cdot L^2 \tag{6.17}$$

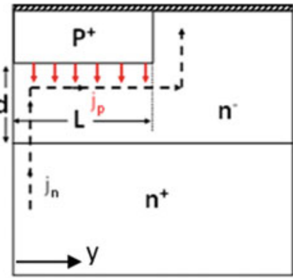


Fig. 6.14 Simplified drawing for the triggering state in an MPS diode. © 2016 IEEE. Reprinted, with permission, from [Pal16]

Here R is the sheet resistivity in $\Omega/\square$ and L is the lateral length of the pin-region. R can be approximated by ρ/d for a not too thin epi-layer thickness d, $d > ½\ L$. If it is assumed that the p-layers inject as V_p is equal to the difference in built-in voltage of the pn-junction and threshold of the Schottky-junction $V_{bi} - V_S$, this results in [Pal16]

$$j_p = \frac{2 \cdot (V_{bi} - V_S)}{R \cdot L^2} \tag{6.18}$$

Since $\rho = 1/(q \cdot \mu_n \cdot N_D)$ contains the bulk mobility μ_n in SiC which decreases strongly with temperature, this explains the triggering at lower current for the high temperature branch in Fig. 6.13. With these simplifications one can approximate the triggering state with the size of the pin-area in an MPS diode. For the structure in Fig. 6.12, it was found that p-areas start to contribute subsequently to the current conduction: first the pin-region at the edge termination triggers, followed by the additionally large p^+ cells, as last the small p^+ cells contribute [Rup12].

Further considerations on the triggering of the p^+-injection are reported in [Hua16], an extended model is given in [Niw17]. Work on high-voltage MPS diodes is presented in [Niw17]. The devices are fabricated with an epitaxial p^+-anode layer, and intensively investigated is the “snapback” phenomenon which is visible in the measurement in Fig. 6.13 at 6.2 V. In [Niw17] p-layers with wide L (see Fig. 6.13) of 50–150 µm are used, and for L, in case it is large enough, the transition from the Schottky-region the pin-operation becomes smooth. As design guideline of snapback-free hybrid operating MPS diodes, $L/d > 1.5$ is given. MPS diodes with breakdown voltages up to 11.3 kV are shown.

SiC MPS diodes have shown a high capability to withstand stress in avalanche, as shown for 600 V in [Rup06] and 1200 V in [Rup14]. The key enabler is the intentional field crowding at the bottom of the implanted p-layers in the cell structure, and then avalanche breakdown occurs at the edges of the p-layers (see Fig. 6.10c). Additionally, it is necessary to balance the breakdown in the junction termination with the breakdown in the cell field and to design the cell field in a way to have a lower breakdown voltage than the junction termination [Dra15].

For MPS diodes with such design, an increase of the breakdown voltage with temperature is found. A linear increase of V_{BD} with 0.35 V/K is measured, see Fig. 6.15. This differs from the Schottky-diode in Fig. 6.9. Obviously, the effect of

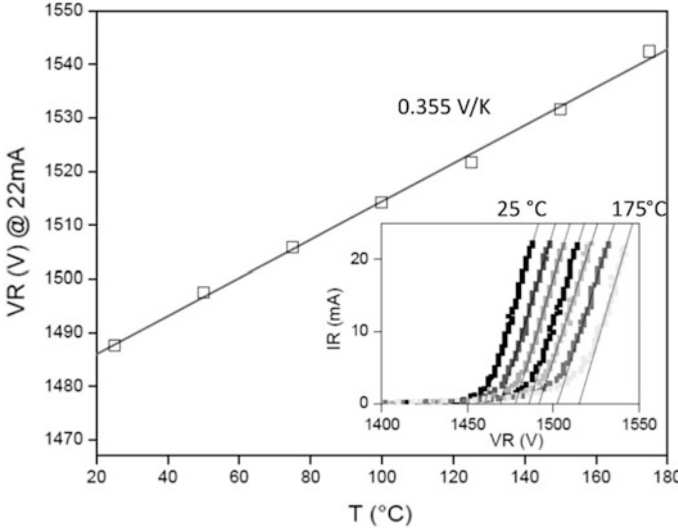


Fig. 6.15 Temperature dependency of the breakdown voltage in a 1200 V MPS diode. Figure from R. Rupp, Infineon

increased leakage current by the image force is reduced if the Schottky junction is released from high electric field. This can be done by a narrow distance of the p-layers in MPS diodes.

6.5.3 Switching Behavior and Ruggedness of SiC Schottky and MPS Diodes

SiC-Schottky-diodes are in competition with fast silicon pin-diodes, they can gain a market by a better switching behavior. A great advantage of SiC-diodes is their much smaller charge necessary for a given on-state conductance G, since this charge appears during commutation as reverse current integral. In a chopper circuit the reverse current pulse of the diode during turn-on of the switching element adds to the forward current in the latter, hence the lower charge leads to reduced turn-on losses in the switch.

For the NPT-case (triangular field) the ratio between the charge and conductance is

$$\frac{Q}{G} = QR_{\Omega} = \frac{qN_D A w_B}{q\mu_n N_D A / w_B} = \frac{w_B^2}{\mu_n} = \frac{4V_{BD}^2}{\mu_n \mathsf{E}_c^2} \tag{6.19}$$

The figure-of-merit of a semiconductor for a high on-state conductance per stored charge is hence $\mu_n \mathsf{E}_c^2$ [Sco91], a figure which includes static *and* dynamical properties. In a form considering also the mobility of holes in pin-diodes (μ_n replaced by $\mu_n + \mu_p$) it can be used for an approximate comparison of SiC Schottky diodes with Si pin-diodes. Assuming a factor 10 in the critical field and a factor 1/2 from mobilities, the product $Q{\cdot}R_\Omega$ of a 4H-SiC Schottky-diode is obtained to be a factor of 50 smaller than in a Si pin-diode. Hence for equal resistance of the base region the *charge* in the SiC-diode is by the same factor smaller than the stored charge in the Si pin-diode. Whereas the product $Q{\cdot}R_\Omega$ for a given material and breakdown voltage has a fixed value, the factors Q and R_Ω can be changed (in an inverse manner) via the active area A of the device. A lower limit to the area is set up by the losses to be conducted away.

A disadvantage of power Schottky diodes are oscillations of reverse voltage and current produced after commutation by the capacitance of the Schottky junction C_j and the stray inductance L_p, as shown in [Sco91]. If only Si is replaced by SiC in high-current applications, huge problems with oscillations may occur [Win12, Lut14]. The capacitance of the SiC-device is given by

$$C_j = \sqrt{\frac{q \cdot \varepsilon \cdot N_D}{2 \cdot (V_j + V_r)}} \cdot A \tag{6.20}$$

and has its maximum C_j (0 V) when the applied voltage V_r is zero. Due to the two decades higher doping N_D it is one decade higher than in the case of a Si-diode with the same active area. A tail current which occurs in Si bipolar devices and which damps oscillations is missing with SiC. The oscillations can be damped making the oscillation period $\tau_{osc} \sim \sqrt{L_p}$ not larger than the turn-on time of the switching element in the chopper circuit, which time should be small also, however. This is a challenge for a new packaging technology which requires much lower parasitic inductance. There is much effort invested today in the development of packages with very low inductance. In a package rated for 400 A with L_p of only 1.4 nH, the oscillations are strongly reduced [Bec16]. Special care has to be taken in case of paralleling, since non-symmetric inductivities in front of parallel connected chips will lead to bad dynamic current sharing and internal LC-oscillations.

MPS diodes show a high dynamic ruggedness at turn-off under overload conditions. Measurements of a 5 A MPS diode in a double-pulse circuit according to Fig, 5.19 executed with an IGBT rated to 90 A and 1200 V at $R_G = 20\ \Omega$ are displayed in Fig. 6.16 [Fic15]. Up to a current of 25 A the diode shows the typical fast-switching characteristics of a Schottky diode with no bipolar charge, and the occurring reverse recovery charge is only caused by the space charge capacity and independent from the forward current. At 50 A, however, the tendency to a bipolar reverse-recovery behavior can be presumed. At this forward current, the first p-doped regions are already active and the diode is slightly bipolar. When imposing a forward current of 75 A before the turn-off, the diode becomes bipolar. This is reflected in the switching behavior, which is similar to the behavior of a pin-diode.

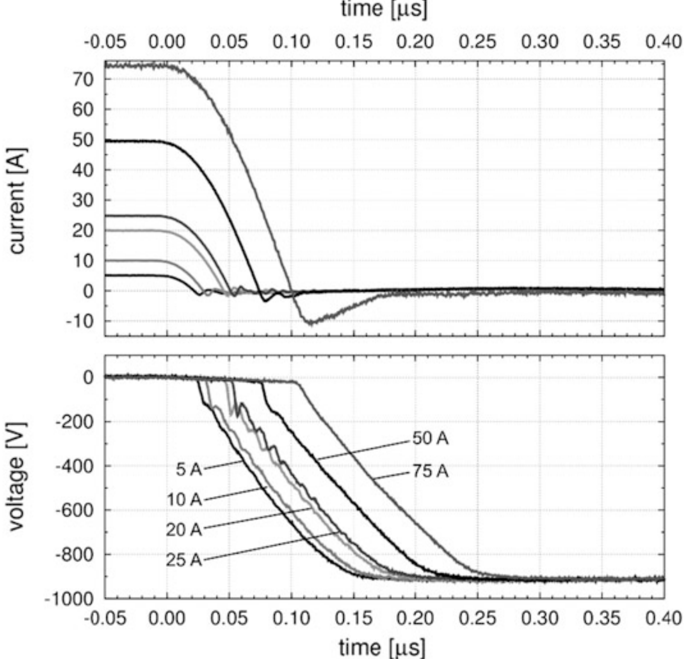


Fig. 6.16 Turn-off behavior of an MPS diode with rated current 5 A, T = 25 °C. The forward current varied between one to fifteen times the rated current. © 2015 Elsevier. Reprinted from [Fic15] with permission from Elsevier

The diodes were able to withstand turn-off at high di/dt and dv/dt executed by setting $R_G = 1\ \Omega$ at more than fifteen times the rated current. Above approximately eight to ten times the rated current, their switching behavior equals that of a pin-diode.

SiC MPS diodes have become mature devices. They are used in the 600 V range in combination with Silicon MOSFETs for high-frequency applications in Power Factor Correction (PFC) circuits and in DC-DC converters. In application as freewheeling diodes for IGBTs, SiC MPS diodes allow to reduce turn-on losses of the IGBT due to the very low reverse recovery charge, see Sect. 5.7.2.

SiC Schottky diodes are preferred for several applications that involve high switching frequencies. SiC devices are still young and, when compare them to Si devices, one must remember that the silicon technology has been improved for several decades to reach the device performance achieved today. For silicon carbide there still is a high potential for optimization of crystal quality as well as for processing technology and device design. It must be considered, however, that SiC is a challenge for the packaging technology because of the higher stiffness compared to Si, respectively the higher Young's modulus of SiC compared to Si.

Therefore, replacements of Si by SiC must consider reliability, especially power cycling capability (see Chap. 12 on reliability).

References

[Bar09] Bartsch, W., Schoerner, R., Dohnke, K.O.: Optimization of bipolar SiC-diodes by analysis of avalanche breakdown performance. In: Proceedings of the ICSCRM 2009, paper Mo-P-56 (2009)

[Bec16] Beckedahl, P., Buetow, S., Maul, A., Roeblitz, M., Spang, M.: 400 A, 1200 V SiC power module with 1nH commutation inductance. In: Proceedings of the international conference on integrated power systems (CIPS), (2016)

[Ber97] Berndes, G., Strauch, G., Mößner (IXYS Semiconductor GmbH), S.: "Die Schottky-Diode - ein wiederentdecktes Bauelement für die Leistungshalbleiter-Hersteller". Kolloquium Halbleiter-Leistungsbauelemente, Freiburg (1997)

[Bjo06] Bjoerk, F., Hancock, J., Treu, M., Rupp, R., Reimann, T.: 2nd generation 600 V SiC Schottky diodes use merged pn/Schottky structure for surge overload protection. In: Proceedings of the APEC (2006)

[Dah01] Dahlquist, F., Lendenmann, H., Östling, M.: A high performance JBS rectifier – design considerations. Mater. Sci. Forum **353–356**, 683 (2001)

[Dra15] Draghici, M., Rupp, R., Gerlach, R., Zippelius, B.: A new 1200 V SiC MPS diode with improved performance and ruggedness. Mater. Sci. Forum **821–823**, 608–611 (2015)

[Fic14] Fichtner, S., Lutz, J., Basler, T., Rupp, R., Gerlach, R.: Electro-thermal simulations and experimental results on the surge current capability of 1200 V SiC MPS diodes. In: Proceedings of the 8th international conference on integrated power systems (CIPS), pp. 438–443 (2014)

[Fic15] Fichtner, S., Frankeser, S., Rupp, R., Basler, T., Gerlach, R., Lutz, J.: Ruggedness of 1200 V SiC MPS diodes. Microelectron. Reliab. **55**(9–10), 1677–1681 (2015)

[Hei08b] Heinze, B., Baburske, R., Lutz, J., Schulze, H.J.: Effects of metallisation and bondfeets in 3.3 kV free-wheeling diodes at surge current conditions. In: Proceedings of the ISPS, prague (2008)

[Hei08c] Heinze, B., Lutz, J., Neumeister, M., Rupp, R.: Surge current ruggedness of silicon carbide Schottky- and merged-PiN-Schottky diodes. In: Proceedings ISPSD 2008. Orlando, Florida, USA (2008)

[Hu79] Hu, C.: A parametric study of power MOSFETs. Record of 1979 IEEE power specialists conference, pp. 385–395 (1979)

[Hua16] Huang, Y., Erlbacher, T., Buettner, J., Wachutka, G.: A trade-off between nominal forward current density and surge current capability for 4.5 kV SiC MPS diodes. In: Proceedings of the ISPSD, Prague, pp. 63–66 (2016)

[Lut14] Lutz, J., Baburske, R.: Some aspects on ruggedness of SiC power devices. Microelectron. Reliab. **54**, 49–56 (2014)

[Niw17] Niwa, H., Suda, J., Kimoto, T.: Ultrahigh-voltage SiC MPS diodes with hybrid unipolar/bipolar operation. IEEE Trans. Electron Devices **64**(3), 874–881 (2017)

[Ogu04] Ogura, T., Ninomiya, H., Sugiyama, K., Inoue, T.: 4.5 kV injection en-hanced gate transistors (IEGTs) with high turn-off ruggedness. IEEE Trans. Electron Devices **51**, 636–641 (2004)

[Pal16] Palanisamy, S., Fichtner, S., Lutz, J., Basler, T., Rupp, R.: Various structures of 1200 V SiC MPS diode models and their simulated surge current behavior in comparison to measurement. In: Proceedings of the ISPSD, Prague, pp 235–238 (2016)

[Pet01] Peters, D., Dohnke, K.O., Hecht, C., Stephani, D.: 1700 V SiC Schottky diodes scaled up to 25 A. Mater. Sci. Forum **353–356**, 675–678 (2001)

[Rup06] Rupp, R., Treu, M., Voss, S., Björk, F., Reimann, T.: '2nd Generation' SiC Schottky diodes: a new benchmark in SiC device ruggedness. In: Proceedings of the ISPSD, pp. 1–4 (2006)

[Rup12] Rupp, R., Gerlach, R., Kabakow, A.: Current distribution in the various functional areas of a 600 V SiC MPS diode in forward operation. Mater. Sci. Forum **717**, 929–932 (2012)

[Rup14] Rupp, R., Gerlach, R., Kabakow, A., Schörner, R., Hecht, C., Elpelt, R., Draghici, M.: Avalanche behaviour and its temperature dependence of commercial SiC MPS diodes: influence of design and voltage class. In: Proc. of the 26th ISPSD, pp 67–70 (2014)

[Sch38] Schottky, W.: Halbleitertheorie der Sperrschicht. Naturwissenschaften **26**, 843 (1938)

[Sch39] Schottky, W.: Zur Halbleitertheorie der Sperrschicht- und Spitzengleichrichter. Zeitschrift für Physik **113**, 376–414 (1939)

[Sco91] Schlangenotto, H., Niemann, E.: Switching properties of power devices on silicon carbide and silicon. EPE-MADEP (Symposion on materials and devices for power electronis), Firenze, pp. 8–13 (1991)

[Scr94] Schaffer, W.J., Negley, G.H., Irvine, K.G., Palmour, J.W.: Conductivity anisotropy in epitaxial 6H and 4H SiC. Mater. Res. Soc. Symp. Proc. **339**, 595–600 (1994)

[Sin00] Singh, R., et al: 1500 V 4 Amp 4H-SiC JBS diodes. In: Proceedings of the ISPSD, Toulouse (2000)

[Spe65] Spenke, E.: Elektronische Halbleiter. Springer, Berlin Heidelberg New York (1965)

[Sze81] Sze, S.M.: Physics of semiconductor devices. Wiley, New York (1981)

[Tre01] Treu, M., Rupp, M., Kapels, H., Bartsch, W.: Mater. Sci. Forum **353–356**, 679–682 (2001)

[Win12] Wintrich, A.: Elektronikpraxis Mai (2012)

[Zve01] Zverev, I., et al: SiC Schottky rectifiers: performance, reliability and key application. In: Proceedings of the 9th EPE, Graz (2001)

Chapter 7
Bipolar Transistors

The transistor was invented in 1947 first in form of a point contact transistor, whose emitter and collector were formed by sharp metal wires on a germanium block as base [Bar49, Sho49]. Soon it was clear that the metal semiconductor junctions at the point contacts can be replaced by two closely coupled pn-junctions. The first paper on a bipolar transistor of silicon with diffused emitter and base was published in 1956 [Tan56]. For a bipolar transistor as power switch, the emitter and base are fine interdigidated structures whose distance in the range of 30 μm have to be mastered; the technology for this was available in the 1970s. For a certain time, the bipolar transistor was the most important switching device in power electronics. But already at the end of the 1980s, the IGBT was introduced (see Chap. 10), and the IGBT began to replace the bipolar power transistor. Nowadays power converters are no longer equipped with bipolar transistors. In niche markets, e.g. as line deflection transistors in TVs, they have survived. However, recently, activities to develop bipolar SiC transistors were started.

7.1 Function of the Bipolar Transistor

The bipolar transistor has an npn- or pnp-structure. Therefore, it comprises two subsequent pn-junctions. The power transistor, with the exception of the voltage range below 200 V, always has an npn-structure as schematically shown in Fig. 7.1.

When a positive voltage is applied to the collector C, the pn-junction between base B and collector is biased in blocking direction and the pn-junction between base and emitter E is biased in forward direction. With open base, the electron density is low in the base. The base p-doping is in the range between 10^{16} and 10^{17} cm^{-3}; a density $n_{p0} = n_i^2/p$ in the range of 10^4 cm^{-3} can be achieved from the

© Springer International Publishing AG 2018
J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8_7

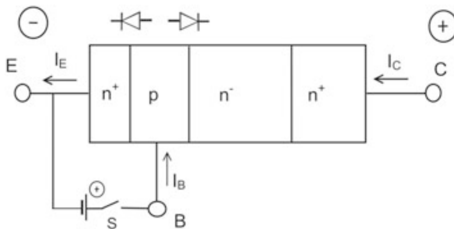


Fig. 7.1 npn power transistor, schematically

relation (2.6). In spite of a high voltage at the collector, the transistor carries only a very small current; it is in the blocking state.

If the switch S is closed and a positive current I_B is injected at the base contact, the n^+p-junction is biased in forward direction, and so the base is flooded with electrons. But with the base current I_B, not only the emitter current is increased. The electrons in the p-base now have a high gradient of carrier density in the direction of the blocking base-collector junction. They diffuse in this direction and into the low-doped n^--layer. If an electric field is build up, they are accelerated by the field towards the collector.

The current gain α in a common-base circuit is defined by

$$I_C = \alpha \cdot I_E + I_{CB0} \tag{7.1}$$

with the leakage current I_{CB0}, measured between base and collector with an open emitter. Furthermore, a current gain β in a common-emitter circuit is defined by

$$I_C = \beta \cdot I_B + I_{CE0} \tag{7.2}$$

with the leakage current I_{CE0}, measured between emitter and collector with an open base. According to Fig. 7.1, I_C is the load current which is to be controlled, therefore β is the current gain of the load current in relation to the controlling base current.[1]

If the relation

$$I_E = I_C + I_B \tag{7.3}$$

resulting from Fig. 7.1 is used and the leakage currents I_{CE0} and I_{BE0} are neglected, Eq. (7.2) is solved for β and finally inserted in Eq. (7.3), it follows that

$$\beta = \frac{I_C}{I_B} = \frac{I_C}{I_E - I_C} = \frac{I_C/I_E}{1 - I_C/I_E} = \frac{\alpha}{1 - \alpha} \tag{7.4}$$

[1]Strictly speaking it must be distinguished between the DC current gain $A = I_C/I_E$ and the small-signal current gains $\alpha = \Delta I_C/\Delta I_E$, the same holds for β. This is neglected in the following simplified treatment.

The same result is also obtained if the exact definitions (7.1) and (7.2) are used; then the leakage currents must be converted. Equation (7.4) can be solved for α which results in

$$\alpha = \frac{\beta}{\beta + 1} \tag{7.5}$$

The closer α approaches 1 the higher is the current gain β of the collector current. For example $\alpha = 0.95$ results in $\beta = 19$.

7.2 Structure of the Bipolar Power Transistor

Figure 7.2 shows the structure of a power transistor. The emitter regions are mostly arranged in stripes, the width of the emitter fingers of power transistors is typically in the range of 200 μm. Base and emitter fingers are arranged in an alternating sequence, interleaved like the teeth of two combs.

The collector region is divided into a low-doped n^--layer to take over the electric field and an adjacent higher-doped n^+-layer. The diffusion profile of the bipolar transistor along the vertical line A–B through an emitter region in Fig. 7.2 is shown in Fig. 7.3. A diffusion profile of this type features the "triple-diffused" bipolar power transistor. The term "triple-diffused" stands for the subsequent diffusion of the deep collector layer, followed by the p-base layer and finally by the n^+-emitter diffusion. In this case, the n^+-layer is a deep-diffused layer with a Gauss-shaped profile of the doping atoms. This deep-diffused layer can also be replaced by an epitaxial layer; then a transistor fabricated form epitaxial wafers is given. But the abrupt junction from the n^+-layer to the n^--layer in epitaxial transistors leads to some disadvantages; see later paragraphs on second breakdown.

7.3 I–V Characteristic of the Power Transistor

In Fig. 7.4, a measurement of the forward I–V characteristic of a power transistor is shown. Already at a low collector voltage, e.g. at 0.4 V, a comparatively high current density is reached. This is not possible for a device where a pn-junction is forward-biased, because of the junction voltage in the range of 0.7 V. In a bipolar transistor in this operation mode, both pn-junctions are forward-biased, and the voltage at the pn^--junction is opposed to that at the n^+p-junction. This region of the forward characteristics, which features very low voltage drop, is denoted as saturation region.

The quasi-saturation region, in which the current slightly increases with increasing voltage, follows next to the saturation region. For a higher voltage – not shown in Fig. 7.4 – the bipolar transistor enters the active region. There, the

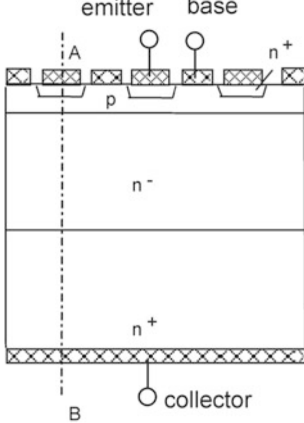


Fig. 7.2 Structure of a power transistor

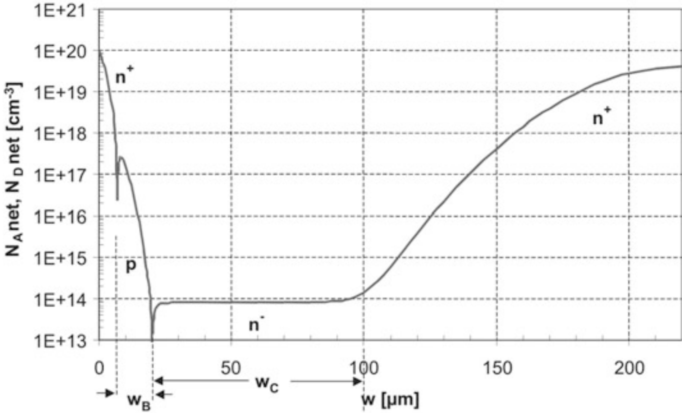
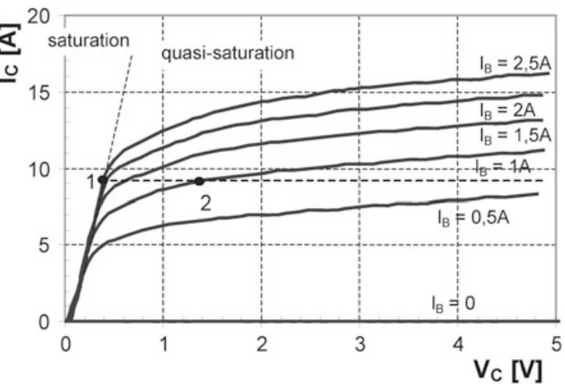


Fig. 7.3 Doping profile of a 1200 V bipolar power transistor along the line A–B in Fig. 7.2

Fig. 7.4 Forward I–V characteristic of a bipolar transistor type BUX 48A



collector current is nearly constant for a given base current and independent of the increasing collector voltage.

From this shape of the forward characteristic, the short circuit capability of the transistor results. The current is limited, even if there is a short circuit at the load. If in a basic circuit according to Fig. 5.18 a short circuit occurs shorting the load consisting of R and L, then the voltage across the transistor increases until the applied external voltage V_{bat} drops across the transistor. The value of the short-circuit current which occurs in short circuit mode is given by the transistor I–V characteristic and the applied base current. Very high losses are generated in this operation mode, however if the short circuit is detected within some µs by appropriate supervising functions in the driver circuit, and is turned-off by the driver, the device will survive this event.

7.4 Blocking Behavior of the Bipolar Power Transistor

In Eq. (7.1), I_{CB0} is the leakage current measured between base and collector. To determine the leakage current between collector and emitter at open base, Eq. (7.1) can be used. At open base, $I_C = I_E = I_{CE0}$ holds, and from (7.1) results that

$$I_{CE0} = \alpha \cdot I_{CE0} + I_{CB0} \tag{7.6}$$

solved for I_{CE0}:

$$I_{CE0} = \frac{I_{CB0}}{1 - \alpha} \tag{7.7}$$

The leakage current between collector and emitter is therefore always higher than the leakage current between collector and base. With $\alpha = 0.9$, I_{CE0} is ten times larger than I_{CB0}.

Also the onset value of avalanche breakdown for a voltage between collector and emitter is lower at open base compared to a voltage between collector and base. To calculate this, Eq. (7.1) must be extended by the effect of avalanche multiplication. The current αI_E enters the space charge from the collector side and is enhanced by the electron multiplication factor M_n. The leakage current I_{CB0}, mainly created by generation in the space charge, is enhanced by its multiplication factor M_{SC} [compare Sect. 3.3 on avalanche breakdown, especially Eq. (3.68)]. It holds that

$$I_C = M_n \cdot \alpha \cdot I_E + M_{SC} \cdot I_{CB0} \tag{7.8}$$

At open base, $I_C = I_E$ is given, which for the collector current results in

$$I_C = \frac{M_{SC} \cdot I_{CB0}}{1 - M_n \cdot \alpha} \tag{7.9}$$

Therefore, the collector current grows to infinity for $M_n \cdot \alpha = 1$ or it holds

$$M_n = \frac{1}{\alpha} \tag{7.10}$$

for avalanche breakdown between collector and emitter, whereas the avalanche breakdown between collector and base arises only if M_n grows to infinity. For high current gain (α close to 1), the blocking capability at open base is strongly reduced. For $\alpha = 0.9$, it is sufficient as the condition for avalanche breakdown that the multiplication factor M_n grows to $1/0.9 = 1.11$.

Because the ionization factors for electrons are much higher than that for holes, $\alpha_n > \alpha_p$, M_n grows much faster with the electric field (see Fig. 3.15). M_n determines the avalanche breakdown and the approximation by an effective multiplication factor M or an effective ionization rate α_{eff} is no longer valid. The breakdown voltage V_{CE0} is considerably smaller than V_{CB0}. Figure 7.5 shows the difference between V_{CB0} and V_{CE0} for a commercially available bipolar transistor. For this example, V_{CE0} amounts about 60% of V_{CB0}.

For a pnp-transistor, we obtain in analogy

$$M_p = \frac{1}{\alpha} \tag{7.11}$$

Because of $M_p < M_n$, the difference from V_{CE0} to V_{CB0} will be smaller for a pnp-transistor.

In Sect. 3.3, with Eq. (3.69) an often used approximation for the multiplication factor at a voltage $V < V_{CE0}$ was given [Mil85]

$$M = \frac{1}{1 - (V/V_{CB0})^m} \tag{7.12}$$

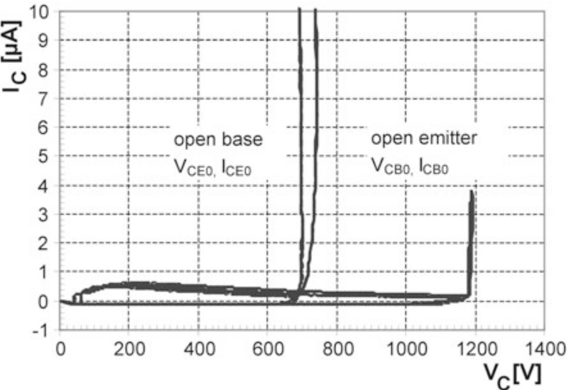


Fig. 7.5 Blocking characteristic of the transistor BUX 48A with open base and with open emitter

where V is an applied voltage smaller than V_{CE0}. With (7.10), it leads for $V = V_{CE0}$ to

$$V_{CE0} = (1-\alpha)^{\frac{1}{m}} \cdot V_{CB0} \tag{7.13}$$

where a value of 5 for m as is given as typical for a bipolar npn-transistor [Ben99]. The value $m = 5$ leads to agreement with measurements, if for α the value at typical forward operation conditions is used where both β and α are high. However, static avalanche breakdown effects appear at low current where the current amplification is low, since a high share of the current recombines in the base layer. If this is considered, $m = 2.2$ as suggested in Sect. 3.3 is a better choice.

An approximation for a transistor with typical structure can be found also for V_{CE0}, [Ben99]

$$V_{CE0} = K_1 \cdot w_C \tag{7.15}$$

where w_C is the width of the low-doped collector zone, as defined in Fig. 7.3, and K_1 is given as 10^5 V/cm.

The difference between reverse blocking voltage and reverse current of a transistor structure with open base and the reverse blocking voltage of a pure pn-junction is very important for devices with several pn-junctions. It is also very fundamental in the practical application of the bipolar transistor, since for a voltage applied at open base, the device goes into breakdown at a much lower voltage and it is potentially destroyed. On the other hand, if a negative voltage is applied to the base with respect to the emitter, both pn-junctions of the transistor are biased in reverse direction. The leakage current of both junctions is extracted as base current, and no more interaction of both pn-junctions occurs. For this case, the blocking capability between collector and emitter is approximately the same as between collector and base. A similar effect occurs if base and collector terminal are connected by a short circuit. In practical application, a small negative voltage is applied at the base of the transistor if a high reverse voltage emerges at the collector. Usually a negative voltage at the base is applied for turning off the collector current, and it is maintained continuously during the blocking mode.

7.5 Current Gain of the Bipolar Transistor

According to definition (7.1), it holds that

$$\alpha = \frac{j_C - j_{CB0}}{j_E} \tag{7.16}$$

Equation (7.16) is multiplied in nominator and denominator by the electron current j_{nB} which is injected by the emitter into the base

$$\alpha = \frac{j_{nB}}{j_E} \cdot \frac{j_C - j_{CB0}}{j_{nB}} = \gamma \cdot \alpha_T \tag{7.17}$$

The first term on the right-hand side of Eq. (7.17) corresponds to the emitter efficiency γ, which was already introduced in Chap. 3.4 with Eq. (3.99) and which is given for an n-emitter by

$$\gamma = \frac{j_{nB}}{j_{nE} + j_{pE}} \tag{7.18}$$

For an n-emitter, this definition describes the share of the electron current j_{nB}, injected into the base, related to the total emitter current.

The second term in (7.17) is denoted as transport factor α_T

$$\alpha_T = \frac{j_C - j_{CB0}}{j_{nB}} \tag{7.19}$$

For an npn-transistor, this corresponds to the share of the electron current injected by the emitter which arrives at the collector. For $j_C = j_{CB0}$ and therefore $\alpha_T = 0$, only the leakage current arrives at the emitter. For an npn-transistor with high current gain, γ as well as α_T shall be close to unity, so that α exhibits a value close to one.

Now the emitter efficiency γ shall be investigated in more detail. Recombination at the pn-junction between emitter and base shall be neglected; this is feasible at current densities higher than 1 mA/cm^2. The electron currents on both sides of this pn-junction are therefore assumed to be equal, $j_{nE} = j_{nB}$. Then it follows that

$$\gamma = \frac{j_{nB}}{j_{nE} + j_{pE}} = \frac{1}{1 + \frac{j_{pE}}{j_{nB}}} \tag{7.20}$$

The minority carrier current j_{pE} penetrating into the emitter can be expressed by

$$j_{pE} = q \cdot \frac{D_p}{L_p \cdot N_E} \tag{7.21}$$

and for the electron current j_{nB} penetrating into the base at the condition of low injection it holds

$$j_{nB} = q \cdot \frac{D_n}{L_n \cdot N_B} \tag{7.22}$$

Equations (7.21) and (7.22) are inserted in Eq. (7.20). For the case of low injection, that means that the flooding with free carriers is smaller than the doping of the base N_B, the emitter efficiency γ can now be expressed by

$$\gamma = \frac{1}{1 + \frac{D_p}{D_n} \cdot \frac{N_B}{N_E} \cdot \frac{w_B}{L_p}} \tag{7.23}$$

In this equation L_n, the diffusion length of the electrons in the base, was replaced by the width of the base w_B, since w_B is always smaller than L_n for the high carrier lifetime typical in bipolar transistors. Further L_p in the emitter will be small, after all w_B and L_p will be in the same order or magnitude. Therefore, the dominating term in (7.23) is the quotient N_B/N_E. To achieve γ close to one, the doping of the emitter N_E must be much higher than the doping of the base N_B. Equation (7.23) gives a first approximation for these relations, which are very important for the design of a transistor.

Equation (7.23) is valid for the case of low injection. Furthermore, bandgap narrowing was not considered in Eq. (7.23); that means a not too high doping of the n^+-emitter was presumed. Additionally, the description of the emitter efficiency in Eq. (7.23) does not contain the current dependency of the emitter efficiency. Based on Sect. 3.4, the emitter efficiency can be investigated in more detail. The n-emitter is characterized by the emitter parameter h_n. Auger recombination and bandgap narrowing determine an n-emitter with high doping; in analogy to (3.108) it holds that

$$h_n = e^{\Delta E_g / kT} \cdot \sqrt{D_p \cdot c_{A,n}} \tag{7.23a}$$

With the Auger coefficient $C_{A,p}$ = 2.8 × 10^{-31} cm^6/s and with the mobility μ_p of 79 cm^2/Vs (estimated for a doping of 1 × 10^{19}cm^{-3}) and using bandgap narrowing according to Slotboom and DeGraaf as described in Eq. (2.25) one obtains

$$h_n \approx 2 \times 10^{-14} \mathrm{cm}^4/s$$

The experience of the authors with design and fabrication of bipolar transistors has shown that the parameter h_n is in the range between 1 × 10^{-14} and 2 × 10^{-14} cm^4/s up to a doping of 5 × 10^{19} cm^{-3}, and it only increases at higher doping as results from (7.23a) with bandgap narrowing according to (2.25). A high value for h_n indicates a reduced emitter efficiency.

The emitter efficiency can be expressed in analogy to (3.100) as

$$\gamma = 1 - q \cdot h_n \frac{p_L^2}{j} \tag{7.24}$$

To estimate γ, the density of the free carriers p_L at the junction between emitter and p-base must be known. In a power transistor, shown again schematically in Fig. 7.6 (top), a low-doped collector layer with a thickness w_C follows adjacent to the p-base. This layer is responsible for taking over the space charge in the forward blocking mode of the transistor. It must be flooded with free carriers in on-state to achieve a low-voltage drop in the conduction mode. The effective base width increases from w_B to w_B + w_C (Fig. 7.6), referred as Kirk-effect [Kir62].

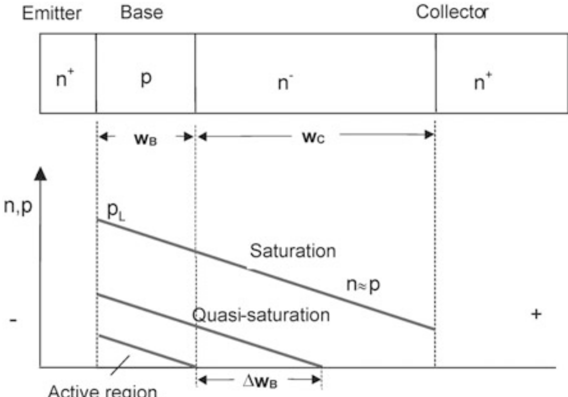


Fig. 7.6 Density of the plasma of free carriers in a bipolar transistor for varied base current at constant collector current

The transistor is in the mode of saturation at high base current and low voltage V_C. This point is marked in the I–V characteristic in Fig. 7.4 with the number 1. The holes injected in the base diffuse also into the low-doped collector region. A conductivity-modulated zone is built up in which $n \approx p$ holds. The shape of the carrier density at point 1 in the I–V characteristic is shown in Fig. 7.6 with the graph for the mode of saturation. The p-base as well as the low-doped collector layer are flooded with free charge carriers. The current transport from the collector to the emitter can be considered as solely carried by electrons, since the diffusion of holes is oriented in the direction against the electric field, whereas the diffusion of electrons is supported by the electric field.

The density of free carriers decreases from the emitter towards the collector. For the distribution across base and low-doped collector layer, Eq. (5.25) still applies. This equation leads to a sagging carrier distribution. The grade of deviation from a linear shape is determined by the charge carrier lifetime. It turns out to be more pronounced for a lower carrier lifetime. In high-quality bipolar transistors, the carrier lifetime is high and the loss by recombination is low, therefore this sagging is neglected in Fig. 7.6. This is a good approximation for $L_n > 2 \cdot (w_B + w_C)$.

Therefore, with the assumed simplified relation

$$\frac{dp}{dx} = -\frac{p_L}{w_B + w_C} \tag{7.25}$$

which applies for the point of transition from the saturation mode to the quasi-saturation mode, see Fig. 7.6, and with $j_p = 0$, $j = j_C$, the result obtained for the collector current is based on Eq. (5.21).

$$j_C = \frac{\mu_n + \mu_p}{\mu_p} \cdot q \cdot D_A \cdot \frac{p_L}{w_B + w_C} \tag{7.26}$$

Using Eq. (5.23) and the Einstein relations (2.44),

$$j_C = 2 \cdot q \cdot D_n \cdot \frac{p_L}{w_B + w_C} \tag{7.27}$$

is obtained. Solved for the density of free carriers at the emitter-base junction

$$p_L = \frac{j_C \cdot (w_B + w_C)}{2D_n \cdot q} \tag{7.28}$$

follows. Equation (7.28) inserted in (7.24) now allows the estimation of the emitter efficiency. For a transistor with w_C = 50 µm for example and w_B = 10 µm, a value for p_L of 2 × 10^{16} cm^{-3} results for a current density of 30 A/cm^2. The emitter efficiency is γ = 0.96. For 10 A/cm^2, a value of γ = 0.99 is obtained. The emitter efficiency depends strongly on the current density. With increasing collector current, the emitter efficiency and thereby the current gain decreases. This is observed for all power transistors (see Fig. 7.7 for high I_C).

The transport factor α_T was the second factor for the current gain according to Eq. (7.17), it can be written for $L_n > 2 \cdot w_{eff}$ [Sze81, Ben99]

$$\alpha_T = 1 - \frac{w_{eff}^2}{2 \cdot L_n^2} \tag{7.29}$$

where L_n represents the diffusion length in the base, which is combined with the carrier lifetime according to Eq. (3.48). w_{eff} is the effective width of the base, which may be shortened compared to w_B at increased voltage, see Fig. 7.8. Equation (7.29) is usually derived for the low-injection condition, for details see [Ben99]. To attain α_T close to 1, w_B must be chosen small and L_n must be the maximum possible. A high current gain requires a short base and a carrier lifetime in the base as high as possible.

The current gain of the bipolar transistor depends on the current and the temperature. For the current gain β, this is shown in Fig. 7.7.

For very small current, β is small; the current injected into the base recombines to a large extent in the base layer. For increasing current, β reaches a maximum; for further increasing current, at the condition of high injection, it decreases again. This

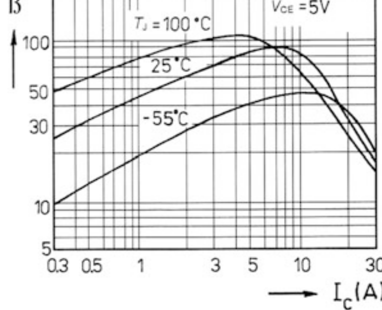


Fig. 7.7 Current gain β as function of collector current and temperature. Data by M. Otsuka, Development of High Power Transistors for Power Use, Toshiba 1975. Reprint from [Ben99] with permission of John Wiley & Sons, Inc

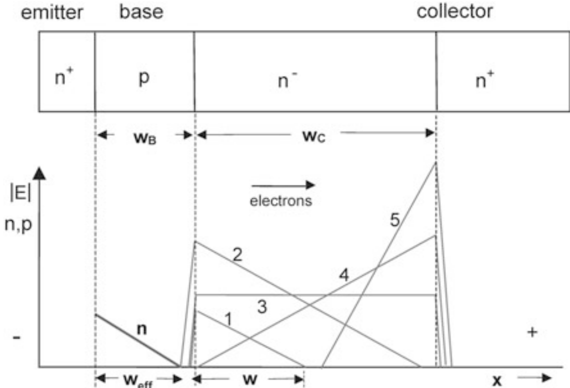


Fig. 7.8 Electric field in the active mode of the bipolar transistor. 1 → 2: increasing voltage V_{CE}. 2 → 3, 4, 5: constant voltage, increasing current I_C

is attributed to the emitter efficiency, which decreases as described by Eqs. (7.24) and (7.28). Additionally, the temperature dependency must be taken into consideration. β increases with temperature for small and medium current, because the carrier lifetime increases with temperature. For high current, the decrease of β with temperature is shifted to lower currents. The rated current of a bipolar transistor is typically in the range where β und α are already declining.

7.6 Base Widening, Field Redistribution and Second Breakdown

If the base current is reduced while the collector current remains constant, the transistor transits into the mode of quasi-saturation; in the I–V characteristic in Fig. 7.4, this is the shift from point 1 to point 2. Now the low-doped collector layer is flooded with plasma of free carriers only up to Δw_B in Fig. 7.6. Only electrons carry the current in the region w_C–Δw_B which is free of plasma, and because of the low doping, a significant resistive voltage drop is created. It can be described in analogy to Eq. (6.8) by

$$\Delta V_{CE} = \frac{j_C \cdot (w_C - \Delta w_B)}{q \cdot \mu_n \cdot N_D} \tag{7.30}$$

If the collector current is kept constant and the base current is decreased, then Δw_B decreases as shown in Fig. 7.6. For $\Delta w_B = 0$, only the base w_B is flooded with free carriers; the transistor has now reached the active region. At this point, $\Delta V_{CE} = 13.4$ V results for the example of a 600-V-transistor with $w_C = 60$ µm, background doping $N_D = 1 \times 10^{14}$ cm^{-3}, $j_C = 50$ A/cm^2 and $\mu_n = 1400$ cm^2/Vs.

Now the pn-junction between base and collector has become free of carriers. If now the voltage is increased, a space charge will build up. For a moderate voltage this is shown in Fig. 7.8 with line 1, for increased voltage line 2. The flooded zone

is pushed back further into the base by the amount that the space charge penetrates the higher doped base layer. The base width is shortened to the effective base width w_{eff}, which leads to a slight increase in the current gain. In literature, this is described as Early effect (named after its discoverer James M. Early, 1922–2004) [Ear52]. Power transistors are usually not operated in the active region. However, they pass through the active region at switching instants.

For a voltage in the active region close to the voltage V_{CE0}, the electric field is shown in Fig. 7.8, line 2. The space charge has build up over nearly the whole layer w_C for this case.

The collector current now flows through the space charge as electron current. The condition of a high electric field applies, and the electrons travel with the drift velocity $v_d \approx v_{sat}$ over almost the whole distance. For the electron density holds

$$n = \frac{j}{q \cdot v_{sat}} \tag{7.31}$$

The shape of line 2 in Fig. 7.8 is given only as long as the density of electrons travelling through the n$^-$-layer is small compared to the background doping N_D. The negatively charged electrons compensate the background doping, and according to Poisson's equation it is

$$\frac{dE}{dx} = \frac{q}{\varepsilon}(N_D - n) \tag{7.32}$$

If the collector current increases by increased base current, the condition will be reached that the density of electrons flowing through the space charge is equal to the background doping:

$$\frac{j}{q \cdot v_{sat}} = N_D \tag{7.33}$$

For this case, $dE/dx = 0$ applies and an almost rectangular electric field with the shape of line 3 in Fig. 7.8 exists. For a bipolar transistor with the background doping $N_D = 1 \times 10^{14}$ cm^{-3}, Eq. (7.33) is fulfilled at a current density of approximately 160 A/cm^2, using $v_{sat} = 10^7$ cm/s.

If the collector is current further increased by an increased base current, then $n > N_D$ holds, and the gradient of the electric field changes its sign, see line 4 in Fig. 7.8 [Hwa70]. The electric field is redistributed and the field maximum shifts from the pn-junction to the nn$^+$-junction. A further increase in the current leads to a further increased field strength (line 5), and finally avalanche breakdown at the nn$^+$-junction will occur.

This is the onset of second breakdown. This effect was explained first by Phil Hower [How70]. Second breakdown is destructive: holes, generated by avalanche at the nn$^+$-junction, are accelerated in the layer w_C. The front side of the transistor is in the active mode, the arriving holes act as an additional base current and even

more electrons are generated by the emitter etc. A positive feedback loop is established. Such a mechanism is usually destructive.

To avoid this critical condition, a safe operating area (SOA) is defined for the transistor for the case of positively biased base – FBSOA, forward-biased SOA – and for the case of negatively biased base – RBSOA, reverse biased SOA. The turn-off is a particularly critical condition in this respect. During turn-off with inductive load, the voltage must increase first before the current can decrease. The transistor passes through the active region of the I–V characteristic. The specification of an RBSOA sets a limit for the voltage against which the transistor can be turned off. Figure 7.9 shows an example.

The turn-off process is also critical because of the following reason: the current below an emitter finger is extracted from both edges toward the center. At turn-off with an inductive load, a small area in the center of the emitter finger finally remains, which carries the total current. Thus, the current density is increased and the mechanism according to Eqs. (7.31) – (7.33) is triggered at lower currents. Some design measures are possible to increase the SOA. The width of the emitter fingers can be reduced and the pitch of the structure can be diminished. Also special emitter structures were tested, which maintain the current flow at the edge of the emitter, e.g. the ring emitter structure [Mil85]. These structures showed an extended RBSOA and a considerably improved stability against second breakdown.

Finally, a shallow gradient of the diffusion profile at the nn$^+$-junction, as shown in Fig. 7.3, is a countermeasure against the formation of a field peak at the nn$^+$-junction. For a slightly increasing n-doping towards the collector, the electric field can penetrate significantly into the n collector layer and a higher voltage can be sustained before the conditions for avalanche breakdown are encountered. Usable transistors in the range of 1000–1400 V could be manufactured with a diffusion profile similar to Fig. 7.3.

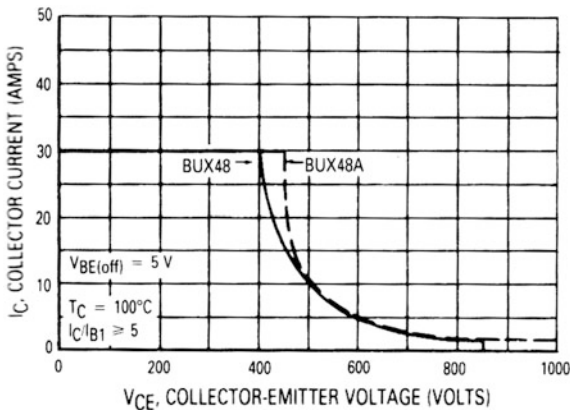


Fig. 7.9 RBSOA of the Motorola transistor BUX 48

7.7 Limits of the Silicon Bipolar Transistor

If the transistor is designed for higher voltage, the low-doped collector layer w_C must be widened. Since the function of the bipolar transistor is based on diffusion of holes into this low-doped collector region, the current gain decreases with increasing w_C. The Motorola transistor in Fig. 7.4 exhibits $\beta = 10$ only up to a collector current of 10 A. The high base current requires a high effort and generates significant losses in the base drive units.

The required base current could be reduced down to acceptable values by the introduction of double- and triple-stepped Darlington transistors [Whe76]. Darlington transistors with blocking voltages of 1200–1400 V became available with up to 100 A of controllable current per single die. With Darlington transistors, high-switching frequencies are no longer possible. But the requirements for variable speed motor drives with switching frequencies in the range of 5 kHz could be fulfilled by Darlington transistors.

An extension to higher voltages is not possible with silicon. Meanwhile, a field-controlled device, the IGBT, was found for the motor drive applications. IGBTs are much easier to control and cause lower losses in the driver. Therefore, bipolar power transistors have been widely taken from the market of power devices and replaced by IGBTs. But the knowledge of the effects in bipolar transistors is essential for a deeper understanding of the effects in more complex power devices.

7.8 SiC Bipolar Transistor

With SiC, a bipolar transistor with a much thinner collector layer and therefore a drastically reduced w_C is possible. Figure 6.8, which determines the necessary width of the low doped region, is also valid for dimensioning a SiC bipolar transistor. According to the small w_C, an acceptable current gain can be achieved even for transistors with blocking voltages above 1000 V. To obtain a high current gain, state-of-the-art SiC epitaxial growth and surface passivation are important. SiC bipolar transistors can be fabricated with very low on-state voltage drop, if ohmic contacts with low contact resistivity are formed [Lee07].

Figure 7.10 shows the measurement of a SiC "large area" bipolar transistor, fabricated by TranSiC AB [Dom09]. A value for β of 35 was achieved for a BJT with open-base breakdown voltage V_{CE0} of 2.3 kV. It should be noted that the characteristic is nearly ohmic in the saturation mode, and an "on-resistance" of 0.03 Ω can be observed, which is about 0.45 Ω mm^2 for this device with 15 mm^2 active area.

The SiC transistor establishes the opportunity to operate a high-voltage device with a voltage drop clearly below 1 V at room temperature. Further, SiC has the advantage of a possible higher doping of the region w_C. As to be see in Fig. 3.19, the maximum doping concentration for a given blocking voltage is for SiC two

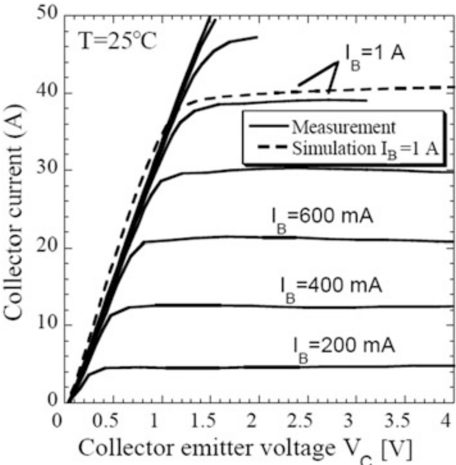


Fig. 7.10 I–V characteristic of a SiC bipolar transistor. Active area 15 mm^2, breakdown voltage V_{CE0} = 2.3 kV

decades higher than for Si. Therefore, the effect of second breakdown according to Eqs. (7.31)–(7.33) is to be expected only at a very high current density which is outside the range of possible operation. For 1200 V SiC BJT, successful operation at V_C = 1100 V and a current density up to 2990 A/cm^2 has been shown by measurement in [Gao07]. In contrast to the Si BJT, the SiC BJT is not restricted by the second breakdown in the application-relevant area.

Additionally, high operation temperatures are possible with SiC, but then a reduced current gain and an increased on-resistance must be taken into account. The progress in SiC technology had revitalized the interest in bipolar transistors again. However, in power device application the SiC bipolar transistor competes with the SiC MOSFET.

References

[Bar49] Bardeen, J., Brattain, W.H.: Physical principles involved in transistor action. Phys. Rev. **75**, 1208–1225 (1949)

[Ben99] Benda, V., Govar, J., Grant, D.A.: Power Semiconductor Devices. Wiley, New York (1999)

[Dom09] Domeij, M., Zaring, C., Konstantinov, A.O., Nawaz, M., Svedberg, J.O., Gumaelius, K., Keri, I., Lindgren, A., Hammarlund, B., Östling, M., Reimark, M.: 2.2 kV SiC BJTs with low V_{CESAT} fast switching and short-circuit capability. In: Proceedings of the 13th International Conference on Silicon Carbide and Related Materials, Nuremberg (2009)

[Ear52] Early, J.M.: Effects of space-charge layer widening in junction transistors. Proc. IRE **40** (11), 1401–1406 (1952)

[Gao07] Gao, Y.: Analysis and optimization of 1200 V silicon carbide bipolar junction transistor. Ph.D. Thesis, Raleigh, North Carolina (2007)

[How70] Hower, P.L., Reddi, K.: Avalanche injection and second breakdown in transistors. IEEE Trans. Electron Devices, **17**, 320 (1970)

[Hwa86] Hwang, K., Navon, D.H., Tang, T.W., Hower, P.L.: Second breakdown prediction by two-dimensional numerical analysis of BJT turnoff. IEEE Trans. Electron Devices. **33** (7), 1067–1072 (1986)

[Kir62] Kirk, C.T.: A theory of transistor cut-off frequency (fT) fall-off at high current density. IEEE Trans. Electron Devices, **23**, 164 (1962)

[Lee07] Lee, H.S., Domeij, M., Zetterling, C.M., Östling, M., Heinze, B., Lutz, J.: Influence of the base contact on the electrical characteristics of SiC BJTs. In: Proceedings ISPSD, Jeju, Korea (2007)

[Mil57] Miller, S.L.: Ionization rates for holes and electrons in silicon. Phys. Rev. **105**, 1246–1249 (1957)

[Mil85] Miller, G., Porst, A., Strack, H.: An advanced high voltage bipolar power transistor with extended RBSOA using 5 μm small emitter structures. 1985 Int. Electron Devices Meet. **31**, 142–145 (1985)

[Sho49] Shockley, W.: The theory of p-n junctions in semiconductors and p-n junction transistors. Bell Sys. Techn. J. **28**, 435–489 (1949)

[Sze81] Sze, S.M.: Physics of Semiconductor Devices. Wiley, New York (1981)

[Tan56] Tanenbaum, M., Thomas, D.E.: Diffused emitter and base silicon transistor. Bell Syst. Tech. J. **35**, 1–22 (1956)

[Whe76] Wheatley, C.F., Einthoven, W.G., On the proportioning of chip area for multistage darlington power transistors. IEEE Trans. Electron Devices. **23**, 870–878 (1976)

Chapter 8
Thyristors

The thyristor was the dominating switching device in power electronics for a long time. It was described already in 1956 [Mol56] and introduced to the market in the early 60s [Gen64]. The acronym SCR (Silicon Controlled Rectifier) was primarily used for a thyristor in early publications and is still occasionally in use today. In its basic structure, a thyristor can be fabricated without very fine structures and with low-cost photolithography equipment. The thyristor is still widely used in applications with low switching frequencies, such as controlled input rectifiers which are applied at the grid frequency of 50 or 60 Hz. A further actual application field of the thyristor is the power range that cannot be reached with other power devices - the range of very high blocking voltages and very high currents. For high voltage DC power transmission, thyristors with 8 kV blocking voltage and more than 5.6 kA rated current have been introduced in 2008 as a single device in the size of a 6-inch wafer [Prz09].

8.1 Structure and Mode of Function

Figure 8.1 shows the structure of a thyristor in a simplified drawing. The device consists of four layers forming three pn-junctions. The p-doped anode layer is located at the bottom, followed by the n-base, the p-base and finally the n^+-doped cathode layer.

The three pn-junctions formed by the four alternately doped layers are marked by diode symbols J_1, J_2 and J_3 in Fig. 8.1. If a voltage is applied in the forward blocking direction, the junctions J_1 and J_3 are biased in forward direction and the junction J_2 is biased in reverse direction as long as the device is in the forward blocking state. Thus, across J_2 a space-charge region with a high electric field will build up (Fig. 8.1c). It penetrates widely into the weakly doped n^--layer.

© Springer International Publishing AG 2018
J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8_8

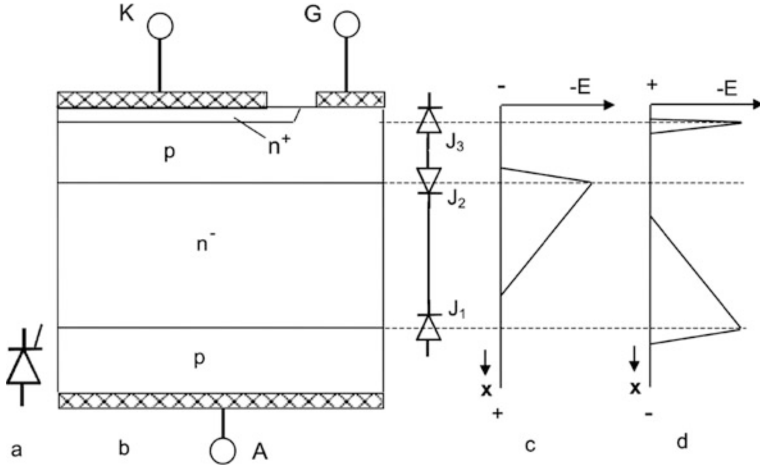


Fig. 8.1 Thyristor. **a** Symbol **b** pn-structure **c** shape of the electric field in the forward blocking mode **d** shape of the electric field in reverse blocking mode

If a voltage is applied in reverse blocking direction of the thyristor, the junction J_2 is forward biased; J_1 and J_3 are biased in reverse direction. Because of the high doping on both sides of the junction J_3, the avalanche breakdown voltage of it is usually relatively low ($\approx$ 20 V). The main part of the applied voltage is taken by the junction J_1, the shape of the electric field is shown in Fig. 8.1d. Since the same weakly doped n^--layer takes the electric field and since the upper and the lower p-layer are usually fabricated simultaneously from both sides in a single diffusion step, the blocking capability of the thyristor is nearly the same for both directions (if the npn-transistor is shorted, see Sect. 8.4). The thyristor is a symmetrically blocking device.

The thyristor can be divided in two partial transistors, a pnp-transistor and an npn-transistor with the common-base-circuit current gains α_1 and α_2, respectively (Fig. 8.2).

Then we obtain for the collector current I_{C1} of the pnp partial transistor according to Eq. (7.1)

$$I_{C1} = \alpha_1 \cdot I_{E1} + I_{p0} = \alpha_1 \cdot I_A + I_{p0}, \tag{8.1}$$

where I_{p0} is the diffusion leakage current from the middle weakly doped n^--layer. In the same way we obtain for the npn partial transistor

$$I_{C2} = \alpha_2 \cdot I_{E2} + I_{n0} = \alpha_2 \cdot I_K + I_{n0} \tag{8.2}$$

with I_{n0} as the diffusion leakage current in the p-base. The anode current I_A is the sum of both partial currents I_{C1} and I_{C2}:

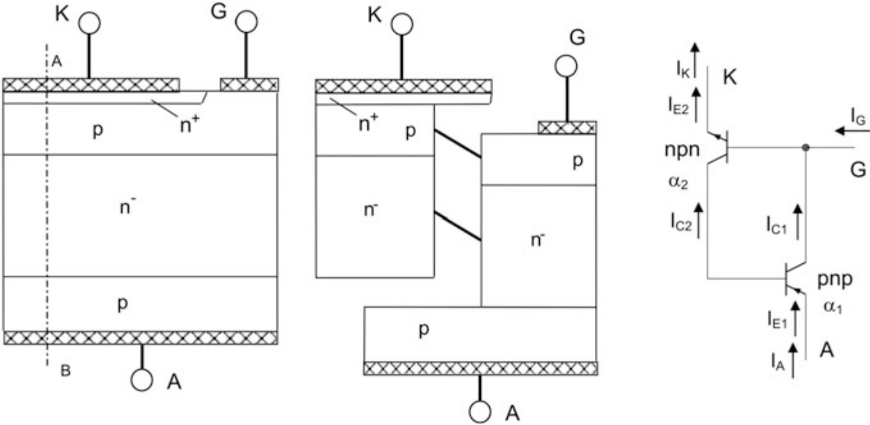


Fig. 8.2 Partition of the thyristor in two partial transistors, and equivalent circuit

$$I_A = I_{C1} + I_{C2} = \alpha_1 \cdot I_A + \alpha_2 \cdot I_K + I_{p0} + I_{n0} \tag{8.3}$$

From the balance of currents flowing into the device and out of the device it holds additionally

$$I_K = I_A + I_G \tag{8.4}$$

Equation (8.4) inserted into Eq. (8.3) leads to

$$I_A = \alpha_1 \cdot I_A + \alpha_2 \cdot I_A + \alpha_2 \cdot I_G + I_{p0} + I_{n0} \tag{8.5}$$

Equation (8.5) resolved for I_A results in an expression for the anode current

$$I_A = \frac{\alpha_2 \cdot I_G + I_{p0} + I_{n0}}{1 - (\alpha_1 + \alpha_2)} \tag{8.6}$$

which can be used as long avalanche multiplication can be neglected. From Eq. (8.6) one can see: I_A rises to infinity, when the denominator in Eq. (8.6) approaches zero. The current gains α_1 and α_2 are in turn dependent on the current. For very low currents they are close to zero, and they increase with current as shown for the bipolar transistor in Fig. 7.7. The trigger condition of the thyristor is therefore

$$\alpha_1 + \alpha_2 \geq 1 \tag{8.7}$$

If the triggering condition is fulfilled, the anode current has the tendency to increase infinitely, this is valid even if I_G = 0 in Eq. (8.6). The thyristor is in the forward conduction mode. In this mode, there is an internal positive feedback loop that is

established by the current amplification of the two partial transistors. Both transistors are in the saturation mode, this leads to a low forward voltage drop that is comparable to the voltage drop across a forward biased diode.

Equivalent to Eq. (8.7) is the condition $\beta_1 \beta_2 \geq 1$. At low current, both α and β grow with increasing current, see Fig. 7.7. Especially, said conditions are also fulfilled if the small-signal current gains $\alpha' = \Delta I_C/\Delta I_E$ or $\beta' = \Delta I_C/\Delta I_B$ are used [Ger79], which are larger than α and β at low current. In nowadays fabricated power thyristors, α_2 and β_2 are determined by the cathode emitter shorts and are zero at low gate current, and therefore the trigger function is mainly adjusted by the cathode shorts. Fore more details see Sect. 8.4.

An exemplary diffusion profile of a thyristor along the line A–B in Fig. 8.2 is shown in Fig. 8.3. The fabrication of a thyristor starts with a weakly doped n^--wafer. Usually, both p-layers are created simultaneously by diffusion: Predeposition with an acceptor dopant, e.g. aluminum, at both wafer surfaces and a subsequent high-temperature drive-in step. To create the deep pn-junctions J_1 and J_3, which are typical for high-voltage thyristors, aluminum is a suitable dopant because of its relatively fast diffusion in silicon. To adjust the required doping concentrations at the n^+p-junction J_3 and near the anode contact, additional p-diffusions are applied. Thus, the final doping profile in the p-layers can be approximated by superposition of several Gauss-type profiles. The junctions J_1 and J_2 both exhibit a very shallow gradient of the diffusion profile at the p-side: This facilitates the fabrication of a junction termination structure with beveled edges as shown in Figs. 4.17 and 4.18. The base of the pnp-transistor with thickness w_B and doping concentration N_D determines the blocking capability of the thyristor in both directions.

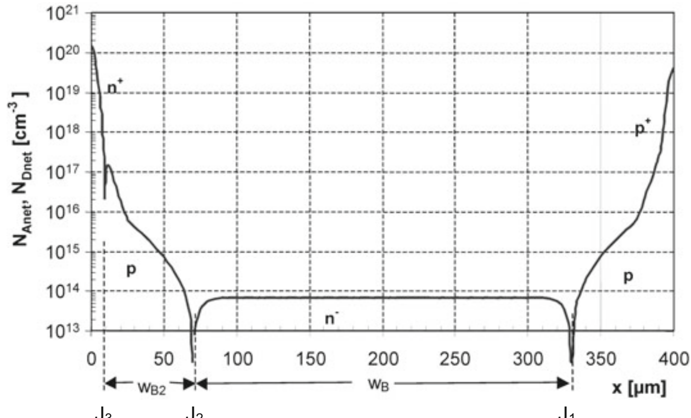


Fig. 8.3 Diffusion profile of a thyristor designed for 1600 V

8.2 I–V Characteristic of the Thyristor

Due to the symmetrical junctions J1 and J2, the blocking behavior of the thyristor is symmetrically in both directions. In forward direction two branches of the I-V characteristic exist: the forward blocking mode and the forward conduction mode. A simplified drawing of the I-V characteristic is shown in Fig. 8.4. In the forward blocking mode, the maximally allowed voltage V_{DRM} is defined at a leakage current I_{DD}. In reverse direction, the maximal voltage V_{RRM} is specified at a maximal allowed current $I_{RD,\ max}$.

The data sheet values for V_{DRM}, V_{RRM} etc. may differ considerably from the values measured for a real device, as mentioned already for the I-V characteristic of diodes. In reverse direction the blocking capability is limited by $V_{R(BD)}$. In forward direction the blocking capability is denoted by the breakover voltage V_{BO}. At a voltage higher than V_{BO} the device is triggered and switches to the forward conduction mode. This mode of triggering, breakover triggering, is usually avoided in power thyristors. Usually the thyristor is fired by the gate. At break-over triggering especially in large-area thyristors, the device may be locally overstressed by uncontrolled local current concentration, and destruction is possible.

In forward conduction mode, the voltage drop V_T is defined at a specified current I_T. The branch of the I-V characteristic for high current is similar to the forward characteristic of a power diode. The middle layer is flooded with free carriers, and similar current densities as in a power diode are possible. Again, as already discussed for the I-V characteristic of power diodes, the data sheet value V_{Tmax}, the maximally allowed forward voltage drop, is higher than the real value V_T of a device, since there is some unavoidable variation in the electrical characteristics of devices. Therefore, the manufacturer usually specifies values with a certain safety margin.

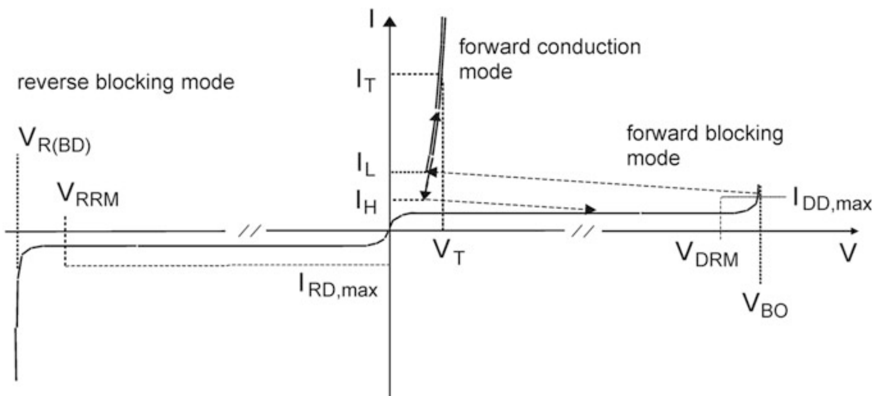


Fig. 8.4 Simplified I-V characteristic of the thyristor and some important thyristor parameters

Dedicated parameters of the forward characteristic are further:

The latching current I_L: The minimal anode current which must flow at the end of a 10 µs trigger pulse to safely switch the thyristor in conduction mode and to maintain the conduction mode safely when the gate signal returns to zero.

The holding current I_H: The minimal anode current necessary to maintain the thyristor in conduction mode without gate current and which ensures, that the conducting thyristor will not extinguish. A decrease of current below I_H can lead to a turn-off of the thyristor.

Since the device is not completely flooded with carriers at the initial phase of the turn-on process, $I_L > I_H$ always applies. The latching current is typically twice as large as the holding current.

8.3 Blocking Behavior of the Thyristor

Avalanche breakdown as limit for the blocking capability is already known from the paragraphs on power diodes and transistors. For a thyristors, there is a second limit of the blocking capability, the punch through effect: The space-charge region, which spreads across the n^--layer with increasing blocking voltage, may arrive at the adjacent layer of opposite doping.[1] Holes will be accelerated in the electric field and the blocking capability is no longer given.

For simplification a triangular electric-field shape across the n^--layer (Fig. 8.1c or d) is assumed in the following discussion. Furthermore, the penetration of the space-charge region into the p-layer of the blocking pn-junction shall be neglected. The avalanche breakdown voltage and its dependency on the background doping were already calculated in Sect. 3.3. It is given by Eq. (3.84) for of the triangular electric field shape. This dependency is drawn in Fig. 8.5 as line (1). It is equal to the dependency in Fig. 3.17, lower part. Additionally, the width of the space-charge region is given by Eq. (3.58). Solving Eq. (3.58) for the voltage and neglecting the small V_{bi} leads to

$$V_{PT} = \frac{1}{2}\frac{q \cdot N_D}{\varepsilon} w_B^2 \tag{8.8}$$

In this equation, the voltage was set to $V_r = V_{PT}$, the voltage at which the space-charge region reaches the region of opposite doping at the position $w = w_B$. For w_B = 250 and 450 µm, the calculated values of V_{PT} are drawn in Fig. 8.5 as line (2) and (3), respectively.

The optimal design parameters for the base width w_B and its doping concentration N_D for a thyristor with a blocking voltage somewhat above 1600 V can be

[1]Already for diodes, the term "punch-through" was used, however in a different meaning of the word. When in diodes the space-charge region extends through the entire n^--layer and penetrates into an n^+-layer, the blocking capability can further increase. See Sect. 5.3.

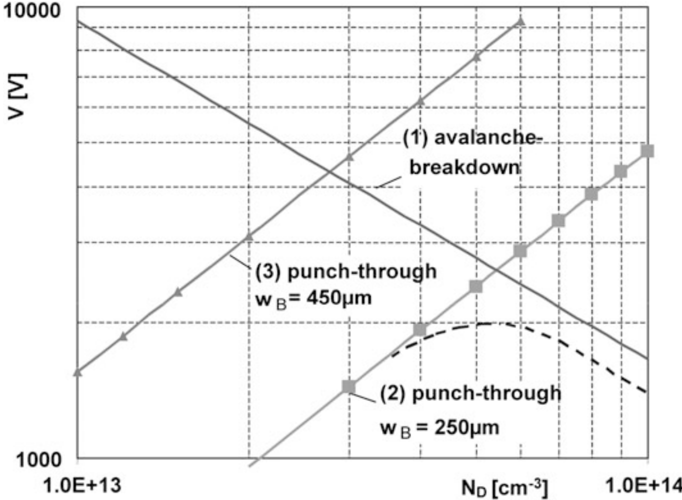


Fig. 8.5 Blocking capability of a thyristor: Avalanche breakdown voltage as a function of N_D, punch-through voltage for two different widths of the n$^-$-layer

estimated by considering the intersection point of lines (1) and (2) in Fig. 8.5. If the doping concentration is reduced below the concentration at the intersection point, the avalanche breakdown voltage will increase, but the space-charge region will reach the opposite p-layer at a voltage lower than the avalanche breakdown voltage and punch-through will limit the blocking capability.

In the following shall be investigated how close one can approach the limits given by avalanche breakdown and punch through. In reverse direction, we neglect the small voltage at J_3 (which is usually shorted, see Sect. 8.4). The behavior at the blocking junction J_1 is equivalent to that of a bipolar pnp-transistor in open base configuration [Her65]. According to Eq. (7.11), the blocking capability of this junction is lower than for a pn-diode. Avalanche breakdown sets on already if $M_p\,\alpha_1 = 1$ holds:

$$M_p = \frac{1}{\alpha_1}. \tag{8.9}$$

Only for $\alpha_1 = 0$, the avalanche breakdown voltage of the pn-junction reaches the value derived for diodes. Since $M_p \ll M_n$ applies (see Fig. 3.15), the effect is not as strong as in an npn-transistor. The onset of avalanche breakdown is reduced to lower voltage, as shown in Fig. 8.5 by the dotted line, the reduction of the breakdown voltage is most pronounced close to the intersection point of lines (1) and (2).

In forward direction the blocking junction is J_2. Its blocking capability is denoted by the breakover voltage V_{BO}. We can use the trigger condition (8.6), set $I_G = 0$ and

take into account the multiplication factors for the hole current in the pnp-transistor and for the electrons in the npn-transistor. The breakover voltage will be reached for

$$I_A = \frac{M_{SC} \cdot I_{SC} + M_p I_{p0} + M_n I_{n0}}{1 - \left(M_p \cdot \alpha_1 + M_n \cdot \alpha_2\right)}. \tag{8.10}$$

For α_1 and α_2 in (8.10), the small signal current gains must be used. The breakover voltage will be reached for $M_p \cdot \alpha_1 + M_n \cdot \alpha_2 = 1$. Since it holds $M_n \gg M_p$, the forward breakover voltage will be very sensitive to α_2. Only for $\alpha_2 = 0$ the blocking capability will be the same as in reverse direction.

Both current gains are dependent on temperature and increase with temperature for low currents. To ensure a blocking capability of the thyristor at higher temperatures, α_2 must be reduced for low currents; since symmetrical blocking capability is required it must be zero for low current. This can be achieved by the implementation of emitter shorts.

8.4 The Function of Emitter Shorts

The current gain of a transistor depends not only on current, but also on temperature. At low temperature the current gain is low; it increases with temperature (Chapter 7, Fig. 7.7). This has the consequence that the breakover condition in open base configuration, Eq. (8.10), will be reached at a lower voltage when the temperature is increased. For the thyristor, the breakover voltage V_{BO} will decrease strongly. This behavior is shown in Fig. 8.6 by the dotted line for a thyristor without emitter shorts.

By the implementation of emitter shorts on the cathode side (Fig. 8.7), a shunt parallel to the pn-junction between the base and the emitter of the npn-partial transistor is created [Chu70, Chk05, Ger79]. The current coming from the pnp-transistor flowing into the base of the npn-transistor is conducted via this shunt

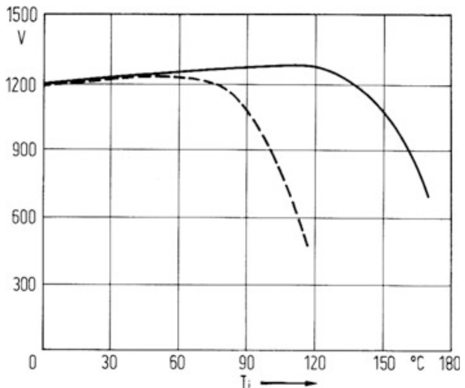


Fig. 8.6 Temperature dependency of the breakover voltage V_{BO}. Dotted line: Without emitter shorts. Full line: With emitter shorts. Figure taken from [Ger79]

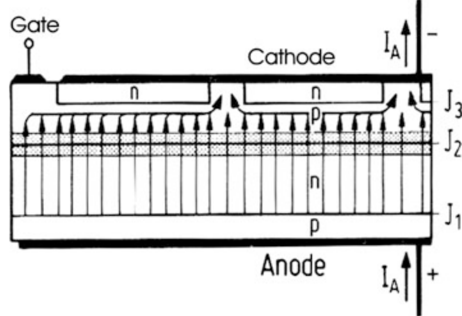


Fig. 8.7 Arrangement of emitter shorts at the cathode side of a thyristor. Figure according to [Ger79]

to the cathode contact. The shunt resistance is determined by the lateral distances between the shunts and the doping concentration of the p-base. If the current is high enough, the voltage drop across the short gets sufficiently high and a perceptible current gain of the npn partial transistor arises.

The cathode emitter shorts determine the effective α_2 and thus widely the dynamic and static characteristics of the thyristor. The dependency of the forward blocking capability on temperature for a thyristor with emitter shorts is drawn in Fig. 8.6 with the full line. With appropriate design of the emitter shorts, the thyristor will have the same blocking capability in forward and in reverse direction even at increased temperature.

Even if a thyristor is provided with emitter shorts, its blocking capability remains sensitive to temperature variations because of the temperature dependency of the current gain of the partial transistors. In most thyristors the maximal allowed operating temperature is restricted to 125 °C, in some special thyristors a little bit above. At increased temperature, thyristors show a significantly increased leakage current compared to diodes.

8.5 Modes to Trigger a Thyristor

A thyristor can be triggered

1. By a *gate current* I_G. This is the most common mode to trigger a thyristor. For technical applications, the following characteristics are given:
   I_{GT}, V_{GT}: Minimal current and minimal voltage that must be provided by a gate unit for securely triggering of the thyristor.
   I_{GD}, V_{GD}: Maximal current and maximal voltage at the gate, at which a thyristor will surely not trigger. To avoid unwanted triggering by disturbing signals, which could occur for example by electromagnetic crosstalk between cables and drive units, these thyristor parameters are very important.

2. By *exceeding the breakover voltage.* In usual power thyristors, this trigger mode is strictly avoided. However, special modifications of thyristor structures, e.g. SIDACs or SIDACtors [SID97], use breakover triggering to act as protection devices against too high voltages. They are connected in parallel to a device or an integrated circuit for protection. They trigger at a voltage higher than V_{BO} and protect other parts of the circuit from overvoltage. Their voltage range is limited to small and medium voltages in the range of some 10 V up to some 100 V. Further, pnpn-thyristor structures are used as electrostatic discharge (ESD) protection structures in integrated circuits, where the same principle of breakover triggering is applied.
3. By a *voltage pulse with a slope dv/dt above the critical* dv/dt_{cr} in forward direction. If such a voltage pulse occurs, the junction capacity of the pn-junction J_2 is charged. If the slope dv/dt is high enough, the generated displacement current may be sufficient to trigger the thyristor. *dv/dt* triggering is an unwanted triggering event. For the application of a thyristor, a maximal allowed dv/dt_{cr} is defined.
4. By a *light pulse* with photons that penetrate into the space-charge region across the junction J_2 [Chk05, Sil75, Sil76]. If the energy of the arriving photons is high enough, electrons transit from the valence band up to the conduction band. The generated electron-hole pairs are separated in the electric field immediately, the electrons flow to the anode, the holes to the cathode. The generated current has the same effect as a current supplied by the gate. If the light power is high enough, the triggering condition Eq. (8.7) can be fulfilled. Light triggering is preferably used in case of series connection of thyristors. This is the case, for example, in applications of high voltage direct current transmission (HVDC), where total voltages up to several 100 kV have to be controlled. The possibility to trigger a thyristor via a glass fiber cable is of high advantage because of the given electrical insulation with this type of signal cable.

The cathode emitter shorts (Sect. 8.4) determine widely the triggering of a thyristor [Sil75]. For activating the npn-transistor, the voltage drop below the emitter to the next emitter short $V = R_{p\text{-}base} \cdot I_G$ must approach the built-in voltage V_{bi} of the n^+p-junction between emitter and p-base, which is in the order of 0.7 V at 300 K. Since V_{bi} will decrease with temperature and $R_{p\text{-}base}$ will increase due to the reduced mobility of holes at elevated temperature, the trigger condition will be much earlier fulfilled at a high operation temperature of 125 °C.

Emitter shorts reduce the trigger sensitivity, increase the trigger current and increase the critical voltage slope dv/dt_{cr} [Sil75]. While the thyristor is most sensitive to dv/dt_{cr} is at 125 °C, it must also be ensured that the necessary trigger current I_{GT} at low temperature – room temperature or down to −40 °C – becomes not too high. The necessary compromise is especially difficult for a light triggered thyristor, where a low trigger power is required in the face of losses e.g. in the glass fiber cables, and still a sufficiently high dv/dt_{cr} must be maintained [Sil76].

8.6 Trigger Front Spreading

With injection of a gate current into a thyristor only the region close to the gate is switched into the conduction mode. The width of the conducting region at the first instant of the turn-on phase is only of the order of some fractions of millimeters. The situation immediately after triggering is illustrated in Fig. 8.8.

The triggered region spreads over the cathode area with a velocity v_z in the range of 50 – 100 µm/µs or 50 – 100 m/s, which is very small for electronic effects. Since there will be only a small lateral voltage drop in lateral direction, the spreading of the initial region of high carrier density is slow. It will take a time of typically 100 – 200 µs until the front of the primary carrier plasma in a thyristor spreads across a length of 1 cm. This slow spreading of the triggered region is a severe limitation for some thyristor applications. It limits the allowed current slope *di/dt* for the anode current.

The trigger front spreading velocity is approximately proportional to the square root of the current density

$$v_z \sim \sqrt{j} \tag{8.11}$$

Furthermore, v_z is reduced by emitter shorts: in the region of shorts it decreases strongly. Emitter shorts reduce v_z to approx. 30 µm/µs for a current density of 100 A/cm^2. Additionally, the carrier lifetime is of influence. If the carrier lifetime is reduced, for example by a gold diffusion, and if, as is necessary for a fast thyristor, strong emitter shorts are implemented, the trigger front spreading velocity v_z may be as low as 10 µm/µs equivalent to 10 m/s – a sprinter at the Olympic Games is running faster!

Thyristors with a higher blocking capability require a larger thickness w_B of the n$^-$-base, v_z decreases with increasing w_B of the thyristor. For a 4.5-kV thyristor with a high carrier lifetime, v_z will be in the range of 20 µm/µs. The dependence v_z on w_B is given by

$$v_z \sim \frac{L_A}{w_B}, \tag{8.12}$$

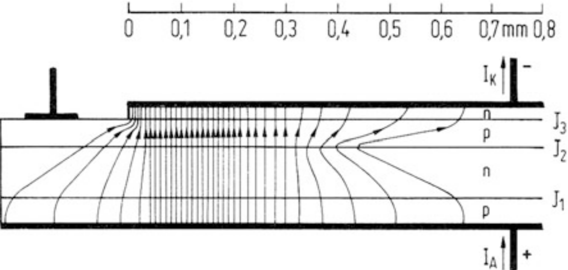


Fig. 8.8 Current distribution in the thyristor immediately after triggering. Figure taken from [Ger79]

where L_A is the ambipolar diffusion length given in Eq. (5.26). A low spreading velocity increases the danger of overstress of the thyristor by an increased local current density in the emitter region close to the gate contact. Therefore, measures to increase the di/dt capability are necessary.

8.7 Follow-up Triggering and Amplifying Gate

Thyristors for applications at line frequency are characterized by a high carrier lifetime and a high ambipolar diffusion length L_A. They are manufactured according to the basic structure described above. Their rated current and *di/dt* capability are of the order of 100 A and 150 A/µs, respectively. These properties are sufficient for applications at line frequency. For applications requiring a higher current capability, it would be possible to increase the gate area, this results in an increase of the initially triggered region. However, this strategy requires a higher gate trigger current I_{GT} and thus a higher effort for the gate drive unit. To avoid this drawback, follow-up triggering was introduced.

First, a pilot thyristor is triggered, which in turn triggers the main thyristor. Thus, the main part of the power necessary for triggering is not provided by the gate drive circuit, but by the main current.

The principle operation of a thyristor structure consisting of a pilot thyristor and a main thyristor is illustrated in Fig. 8.9a. Between the gate and the main cathode K, a small auxiliary cathode K' is arranged. A gate current will trigger the thyristor K'. Its cathode current flows via the resistor R and generates a voltage drop at the resistor. Thus, between K' and K a positive voltage builds up, it generates an electric field in the p-base and a lateral hole current to the n-emitter of the main cathode K, which triggers the main thyristor.

In an advanced configuration the resistor R is integrated into the thyristor (Fig. 8.9b). It is realized by the extension of the n-emitter layer and consists essentially of its lateral resistance in the region reaching from the pilot thyristor to the left border of cathode contact K. This structure was introduced as lateral-field emitter [Ger65] and allows much higher *di/dt* slopes at turn on.

In a thyristor with amplifying gate (AG) [Gen68], shown in Fig. 8.9c, the cathode metal of the auxiliary thyristor K' is connected to the p-base of the main thyristor. After turn-on of the auxiliary thyristor, the current flows across the lateral resistance of the p-base to the main cathode K. The total current of the cathode K' acts as gate current for the main thyristor K. In addition, the overlapping of the cathode contact K' and the p-base forms a cathode short and improves the di/dt capability.

The amplifying gate can be formed in miscellaneous geometrical shapes, for example stripes or other distributed structures spreading across the area of the main cathode. A structure of a large area thyristor is shown in Fig. 8.10. This thyristor is designed for applications of high voltage direct current transmission (HVDC) and has a blocking capability of 8000 V, a rated current of 3570 A (T_{case} = 60 °C), and a di/dt immunity of 300 A/µs.

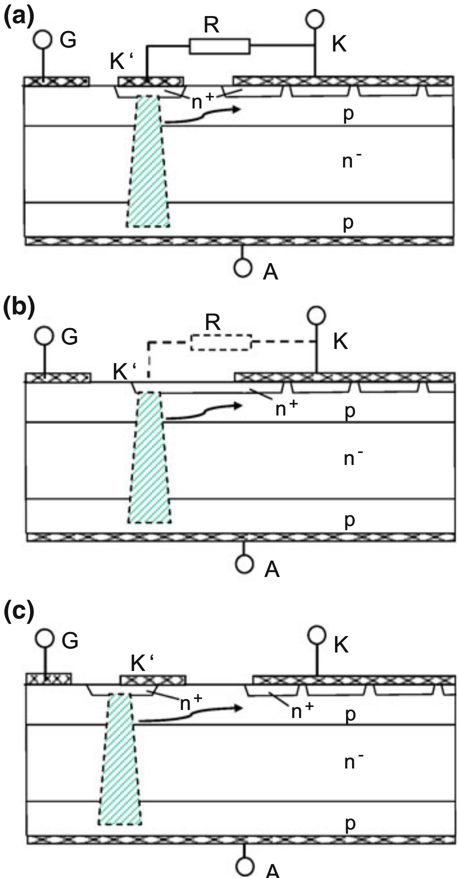


Fig. 8.9 Follow-up triggering of a thyristor. **a** Principle. **b** Lateral-field emitter, integration of the resistor in the n^+ emitter layer. **c** Amplifying gate

Fig. 8.10 Gate structure of a light triggered thyristor, diameter 119 mm. Manufacturer Infineon

The main thyristor area is triggered by a four-stage amplifying gate structure located in the center of the device. The three inner AGs are ring-shaped whereas the fourth AG is distributed about the main cathode. This amplifying gate design ensures triggering of the main thyristor by a wide initial trigger front.

The central structure of such a high-power thyristor can be very complex; especially protection functions can be integrated. Figure 8.11 shows details in the center of the structure in Fig. 8.10 [Nie01, Scu01]. The left edge in Fig. 8.11 corresponds to the chip center in Fig. 8.10. The fourth AG is distributed over the main cathode area to increase the region in which triggering of the main cathode area starts. Direct light-triggering of the thyristor is possible by irradiating the photosensitive area in the center of the device. The typical light power to turn on the thyristor is of the order of 40 mW.

An overvoltage protection is integrated by a breakover diode (BOD) in the center of the structure. Its avalanche current triggers the thyristor via the AG structure when an overvoltage is applied to the device. The voltage level V_{BOD} at which the overvoltage protection function is activated can be adjusted by the distance between the central p region with radius r_{BOD} and the concentric p ring with an inner radius r_p (Fig. 8.11).

The shunt resistance of the weakly doped p^- region below the n^+-emitter of the innermost AG is adjusted such that the *dv/dt* capability of this AG is lower than that of the other AGs and the main cathode area. By this way a reliable *dv/dt* protection function is integrated into the thyristor, because the device is turned on safely by the innermost AG when the anode-to-cathode voltage rises at a rate higher than the rated maximum *dv/dt* value. The resistor R_{23} between the second and the third AGs protects the two innermost AGs from being destroyed when the thyristor is turned on with a high current rate *di/dt*.

A thyristor with a structure in Fig. 8.11 needs high effort in design and fabrication. With the integrated self-protection functions, it includes already elements that support a reliable operation. Integration of these self-protection functions therefore reduces the number of electronic components of the total power electronic system.

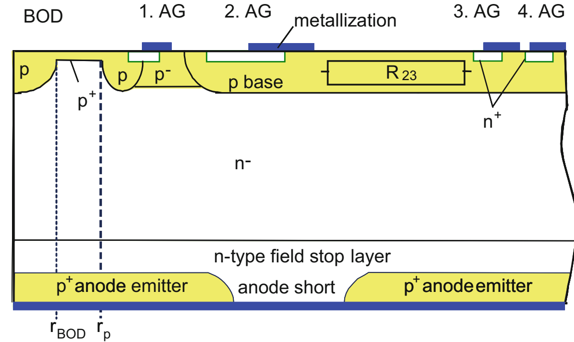


Fig. 8.11 Cross-section of the central amplifying gate structure of a modern high-power thyristor, consisting of four amplifying gates and several protection structures. Figure from [Nie07]

8.8 Thyristor Turn-off and Recovery Time

Only special configurations of the thyristor can be turned off via the gate, these are the Gate Turn Off (GTO) – thyristors which will be treated later. The usual turn-off of a thyristor happens by zero crossing of the anode current, which is given for applications operating in an AC circuit. In case of forward conduction, the base of the thyristor is flooded with free carriers, similar to the internal plasma at forward conduction of a diode. At commutation into the reverse direction, a reverse current will occur, because the stored charge must be removed. This process is similar to the turn-off process in a pin diode. The stored charge must be removed down to a very small rest charge, before a voltage in forward direction can be re-applied to the thyristor. The time which is necessary for this charge removal, and which must be minimally set as hold-off interval to avoid unwanted triggering of the thyristor, is denoted as the recovery time t_q. A thyristor is not able to withstand a forward voltage pulse with the rated blocking voltage or the rated maximum dv/dt_{cr} value until the charge-carrier plasma is almost completely removed from the n^--base.

Figure 8.12 shows the definition of the recovery time. The current slope di_T/dt of the anode current is determined by the external circuit. Equation (5.68) is valid similar as in diodes. The anode current crosses zero, and a reverse current peak occurs due to the removal of the stored charge.

In a typical thyristor first the junction J_3 is depleted of charge carriers when the current is turned-off. However, the blocking capability of this junction is usually only between 10 and 20 V, because of the relatively high doping concentration of the p-base in the range of 10^{17} cm^{-3}, and additionally it is provided with cathode shorts. Thus, the reverse current is increasing until the junction J_1 is depleted of carriers. Then, the thyristor starts to take over the reverse voltage V_R, and a short instant later the reverse recovery current reaches its maximum value.

After having reached the maximum, the reverse current decreases and a voltage peak V_{RM} is generated, similar to the behavior of a diode during the turn-off period.

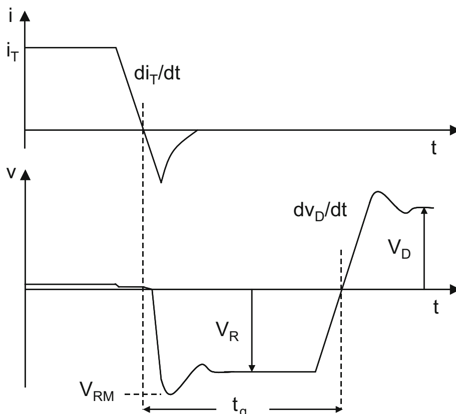


Fig. 8.12 Definition of the recovery time t_q of a thyristor

For thyristors, this period is less critical than for fast diodes. One reason for this is the large width w_B of the n^--layer typically used in thyristors. In most cases, the reverse current decays slowly and a tail current is observed. During this period, charge carriers are still stored in the region close to the junction J_2.

While the thyristor is in the reverse blocking mode, the polarity of the applied voltage changes. A voltage V_D with a defined slope dv_D/dt is applied in forward direction, the thyristor remains in the blocking mode (forward blocking mode) and does not switch into the on-state mode, if $t > t_q$ applies (Fig. 8.12). Otherwise, the control of the circuit is lost. The thyristor must be able to withstand the applied forward voltage, thus, the forward voltage V_D is allowed to be applied only after a certain time interval. This minimum time interval between zero crossing of the current i_T and zero crossing of an applied forward voltage V_D is denoted as turn-off time t_q.

The recovery time of a thyristor is much higher than the switching time of a diode. For the case of charge removal without applying a voltage V_R in reverse direction, and neglecting also V_D, one can estimate according to [Ger79]

$$t_q \approx 10 \cdot \tau, \tag{8.13}$$

where τ is the charge carrier lifetime in the n-base. However, t_q is specified at high operation temperature, whereas τ is typically measured at room temperature. In modern thyristors, t_q depends strongly on the cathode shorts. For turn-off with an applied reverse voltage V_R, Eq. (8.13) can only be used for an estimation of the upper limit value of t_q [Ger79]. If a voltage in reverse direction is applied, only in that part of the base, in which the space-charge region has build up, the stored carriers are removed efficiently by the electric field. The recovery time t_q depends on the application conditions:

- the forward current I_T: t_q increases with increasing forward current;
- the temperature: t_q increases with temperature because τ is increasing with temperature;
- the voltage slope dv/dt. This voltage slope must be smaller than the critical voltage slope dv/dt_{cr} in any case. The closer dv/dt approaches dv/dt_{cr}, the less rest charge is allowed, and the higher is t_q.

With a 100 A 1600 V thyristor, which is in its basic form triggered via a gate in the center, t_q is in the range of 200 μs. With a high-power 8-kV thyristor, t_q is in the range of 550 μs. In large-area thyristors, a forward voltage pulse appearing before t_q may turn on the thyristor somewhere in the main cathode area in an uncontrolled way and – in the worst case – may lead to a destruction of the thyristor. Special structures have been introduced to integrate a t_q protection function into the thyristor [Nie07].

The recovery time limits the maximal frequency range allowed in thyristor applications. With diffusion of gold t_q can be reduced, and also with a higher density of cathode shorts. Fast gold-diffused thyristors reached a t_q down to 10 – 20 μs. Since modern power devices with turn-off capability are meanwhile available,

the interest in fast thyristors has vanished. For the high voltage range > 3 kV, the development of fast thyristors was never successful.

8.9 The Triac

In a triac (triode AC switch) two thyristors are integrated in an anti-parallel configuration in a single device. The triac was introduced early [Gen65]. Figure 8.13 shows the structure.

For a triac, one can no longer distinguish between anode and cathode, therefore the notations "Main Terminal 1" and "Main Terminal 2", MT_1 and MT_2 are conventionally used.

The triac can be triggered via a common gate in both directions. Its I-V characteristic presents in both the first and the third quadrant a conducting branch and a blocking branch. The triac may replace two thyristors in an AC converter, but only within some limits.

The triac is described in more detail in [Ger79]. For the application, the main restriction is due to fact that at zero crossing of the current the triac must block the voltage in reverse direction. However, in conduction mode, the device is flooded with free charge carriers. If the commutation is executed with a too high *di/dt*, still a part of the stored charge remains after zero crossing of the current. If now a voltage with a too high *dv/dt* is applied, an unwanted re-triggering occurs. The device will not transit to the blocking block mode and the possibility to control the converter is lost.

Therefore the allowed current slopes *di/dt* and voltage slopes *dv/dt* are limited drastically for a triac—*di/dt* to some 10 A/µs and *dv/dt* to the order of 100 V/µs. This allows the use of triacs only for applications with comparatively small current and moderate voltage. In fact, in such applications they are often used. An example for triac application is the AC converter for controlling of a medium-power heater.

If currents higher than 50 A have to be controlled, two thyristors in anti-parallel configuration are commonly used instead of a triac.

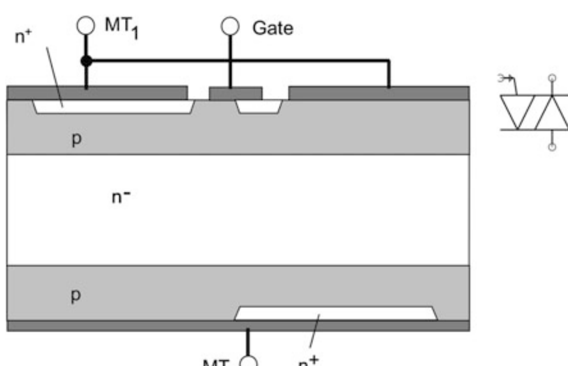


Fig. 8.13 Triac structure and symbol

8.10 The Gate Turn-off Thyristor (GTO)

To provide a thyristor with an active turn-off capability, several special measures are necessary. The Gate-Turn-Off (GTO) thyristor was introduced in the 1980s [Bec80]. In the voltage range above 1400 V it was superior to the bipolar power transistor, the competing device at that time. But with the introduction of the IGBT (see Chap. 10) and the capability to design the IGBT for a high voltage, the GTO thyristor was superseded by the IGBT, since a GTO thyristor requires a high negative gate current for turn-off and the effort for the drive unit is high. The GTO thyristor is nowadays used in the power range that is not reached by the IGBT. Up to 6 kA 6 kV GTO thyristors fabricated from one single 150-mm wafer are available [Nak95]. From the GTO-structure, the new device Gate Commutated Thyristor (GCT) was derived. It exhibits an improved ruggedness and an extended safe operation area.

Equation (8.6) was derived for the trigger condition of a thyristor and describes the dependency of the trigger condition on the current gains of both partial transistors. From this equation, also a turn-off condition can be derived. If in Eq. (8.6) the leakage currents of the partial transistors are neglected, we obtain:

$$I_A = \frac{-\alpha_2 \cdot I_G}{(\alpha_1 + \alpha_2) - 1} \tag{8.14}$$

For turn-off, a negative gate current—I_G is necessary. Similarly to the current gain β for the bipolar transistor, a turn-off gain β_{off} for the turn-off process of a GTO thyristor can be defined:

$$\beta_{off} = \frac{I_A}{-I_G} \tag{8.15}$$

With Eq. (8.14) it follows for the turn-off gain:

$$\beta_{off} = \frac{\alpha_2}{(\alpha_1 + \alpha_2) - 1} \tag{8.16}$$

A high turn-off gain requires on the one hand a high current gain α_2 of the npn partial transistor, on the other hand the denominator $(\alpha_1 + \alpha_2 - 1)$ must be small and ideally approaches to zero. In other words, the sum of the current-gain factors, $\alpha_1 + \alpha_2$, should be only slightly larger than 1. However, this leads to an increased trigger current I_{GT}, an increased latching current I_L, and finally to an increased forward voltage drop V_T of the thyristor. The demand for a high current gain is therefore in contradiction to the requirement for low conduction losses. GTO thyristors show typically a turn-off gain β_{off} between 3 and 5. Thus, to turn off a 3000-A GTO thyristor for example, the drive unit must supply a current of 1000 A.

In fact, Eq. (8.16) is of low value for the design of a GTO thyristor with high turn-off capability. Most important is the homogenous operation of the large number of segments in a high-current GTO thyristor [Shi99].

Further, Eqs. (8.14 - 8.16) hold only if the lateral voltage drop of the turn-off current below the emitter can be neglected [Wol66]. The observance of conditions (Eqs. (8.14 - 8.16)) alone is not sufficient to achieve a gate-turn-off capability of a thyristor. The GTO is thyristor distinguished from the conventional thyristor by its emitter structure, which consists of separated emitter fingers (Fig. 8.14). The width b of the fingers must be small, since the charge carriers below the emitter fingers must be extracted via the gate contact at turn-off. The width b is typically between 100 and 300 μm in modern GTO thyristors. The GTO thyristors consist of a large amount of emitter fingers. They are typically used to control very high currents, therefore a large device area is necessary and it is usual to fabricate a single GTO thyristor from a wafer.

Figure 8.15 shows such a GTO thyristor, fabricated from a 100-mm wafer. The gate contact is formed as a ring located between four inner and four outer rings with emitter fingers. The main reason for this arrangement is to ensure that the distance of the gate contact from the most distant fingers is small enough to allow an efficient extraction of the charge carriers, and that the voltage drop in the gate metallization is not too high.

During turn-off, the holes in the p-base are removed by the negative gate voltage and flow towards the gate contact. The charge carrier plasma which transports the anode current is first removed from the edge of the emitter finger, the residue of the plasma is located in the center of the finger (Fig. 8.16). The hole current must flow laterally underneath the emitter finger. Before the anode current finally is interrupted, a small region in the middle of the emitter fingers, or even of only very few single emitter fingers, takes the total anode current. This is the weak point of the GTO thyristor. To achieve a turn-off capability for high current, it is essential that the resistance of the p-base below the emitter finger is not too high.

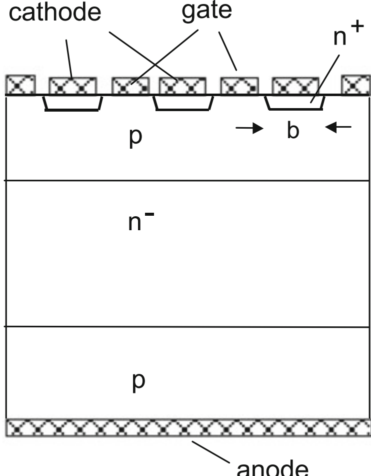


Fig. 8.14 Gate-Turn-Off (GTO) thyristor

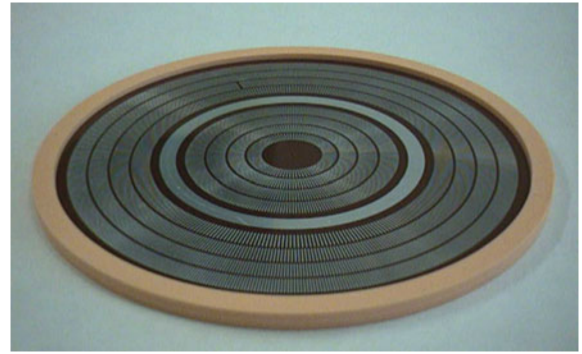


Fig. 8.15 Arrangement of the emitter fingers of a 4.5 kV GTO thyristor fabricated from a 100 mm silicon wafer; final device diameter 82 mm. Figure from Infineon

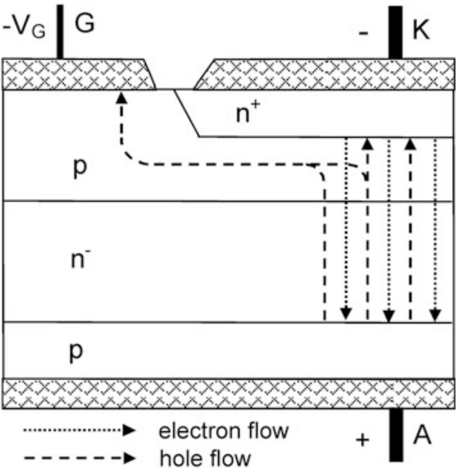


Fig. 8.16 Current flow in a single finger of the GTO thyristor at turn-off

The maximal current I_{Amax} that can be turned off by a GTO thyristor is determined by the breakdown voltage $V_{GK(BD)}$ of the n$^+$p-junction between the cathode and the gate and by the lateral resistance R_p of the p-base below an emitter finger:

$$I_{\mathrm{Amax}} = \beta_{off} \cdot \frac{V_{GK(BD)}}{R_p} \tag{8.17}$$

where

$$R_p \sim \rho\, b \tag{8.18}$$

and ρ is the specific resistivity of the p-base below the emitter finger. In a GTO thyristor with fingers of a width b = 300 µm, the resistivity ρ below the emitter must be four times smaller than in a conventional thyristor. This requires a sufficiently high p-base doping concentration N_A. Simultaneously, a sufficient blocking voltage $V_{GK(BD)}$ of the n$^+$p-junction between cathode and gate is required. This

blocking voltage is given by Eq. (3.84). However, the doping concentration that determines the breakdown voltage of the emitter-gate junction in the GTO thyristor is the doping concentration N_A of the p-base. Therefore, N_A must not be too high. Typical doping concentrations are of the order of 10^{17} cm^{-3}, resulting in breakdown voltages $V_{GK(BD)}$ of about 20 - 22 V. The applied gate voltage at turn-off is usually −15 V.

Values > 4 for β_{off} do not increase the turn-off capability significantly. Decisive for the GTO-design is the second term on the right side of Eq. (8.17).

With these measures the plasma can be efficiently extracted from the p-base of the GTO thyristor. However, the plasma of charge carriers still remains in the wide n$^-$-layer. Thus, additional measures are necessary to remove the charge carriers in this region. The first GTO thyristor generation utilized a gold diffusion to adjust a low carrier lifetime in the n$^-$-layer. However, the gold diffusion is very difficult to control with sufficient accuracy (see Sect. 4.9).

An effective improvement was the implementation of shorts at the anode side. The structure of a GTO thyristor with anode shorts is shown in Fig. 8.17. The hole current is extracted via the gate and the injection of electrons from the n$^+$-emitter is stopped. The electrons in the n-base are removed via the anode shorts by the high positive voltage at the anode side. The injection of the anode emitter is interrupted, and the charge carriers are removed effectively.

A GTO thyristor with anode shorts has no blocking capability in reverse direction. In most applications this is no disadvantage, since an inverse free-wheeling diode is connected in parallel to the GTO thyristor in the power circuit. In modern GTO thyristors, anode shorts and charge carrier lifetime control are combined. An implantation of protons or helium nuclei is preferably used for adjustment of the carrier lifetime. The region with high density of recombination centers

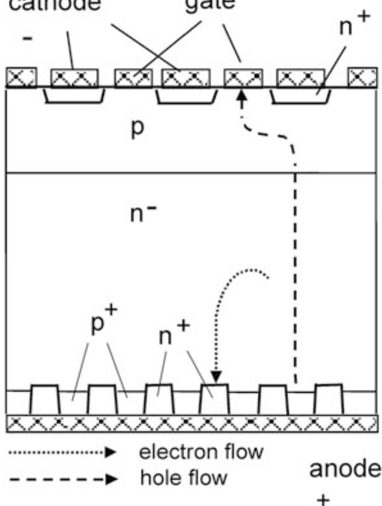


Fig. 8.17 GTO thyristor with emitter shorts at the anode side

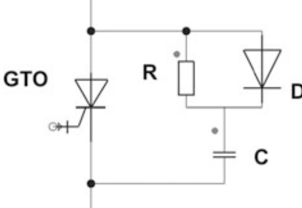


Fig. 8.18 RCD snubber circuit for a GTO thyristor

is located close to the p^+-anode layer, since at this position their effect on the stored charge is most effective.

Despite all these measures, the slope *dv/dt* of the voltage applied to a GTO thyristor at turn-off must be strongly restricted. This is done by an RCD circuit, often denoted as “snubber” (Fig. 8.18). The slope *dv/dt* of the increasing voltage is limited by the capacitor C.

Figure 8.19 finally shows the turn-off process of the GTO thyristor. The negative gate current increases up to the value I_{GRM}, just then the anode current begins to fall. The turn-off delay time t_{gs} is defined as time interval between the moments when the gate current I_G crosses zero and the anode current is dropped to 90% of starting anode current I_{T0}. The anode current then falls steeply within the fall time t_{gf}. During this period, a voltage peak V_{pk} occurs in the waveform of the anode voltage. The value of V_{pk} is determined by the parasitic inductance in the snubber circuit and by the forward recovery voltage peak V_{FRM} of the snubber diode D, the last term typically dominates. Just after V_{pk} the effect of the snubber starts. The slope *dv/dt* of the voltage is then limited by the capacitor C.

In a GTO thyristor, a tail current follows after the interval t_{gf}. This tail current is generated by the extraction of the stored charge in the part of the n-base close to the anode junction. Its duration is of the order of several microseconds, and it generates the main part of the switching losses during the turn-off phase. The implementation of effective anode shorts and the adjustment of the charge carrier lifetime reduce the tail current.

Even if the gate control unit is properly designed, two further factors are of disadvantage for the application of GTO thyristors:

1. The requirement for an RCD snubber. For a high voltage > 3 kV, the capacitor is very voluminous and expensive, especially under the additional constraint of low internal inductance.
2. As already shown, the charge removal below the emitter fingers starts at the finger edges. Before the anode current is decreasing, a narrow region in the centre of the finger remains which carries the total anode current. The larger the device, the more difficult it is to achieve a homogeneous operation of all fingers, and it may happen at the end of the turn-off period that a few of them or even a single finger has to carry the total anode current. This is the weak point of the GTO thyristor, because the respective finger may be destructed under such conditions.

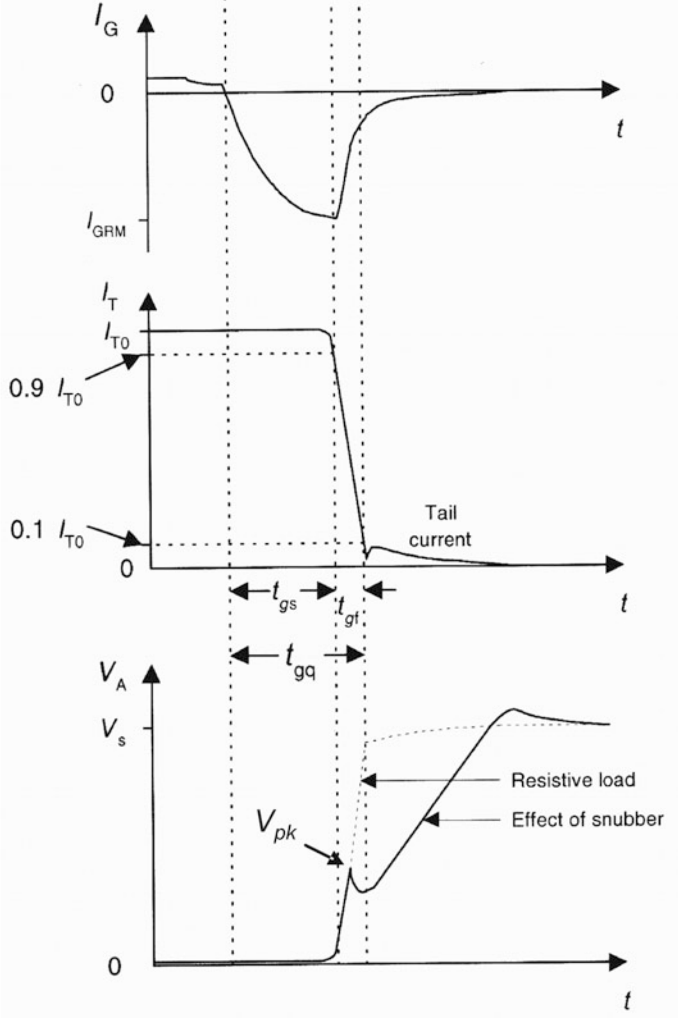


Fig. 8.19 Turn-off characteristic of a GTO thyristor. Figure taken from [Ben99]

8.11 The Gate Commutated Thyristor (GCT)

The operation principle of the GCT [Gru96] is to work with a drive unit that has the capability to transfer the total anode current within a very short time into the gate unit. The GCT is turned off with a turn-off gain $\beta_{off} = 1$.

The GCT consists of the semiconductor device, a gate connection with very low inductance realized by a printed circuit board PCB in coplanar design, and the drive unit assembled with low-inductive capacitors and low-resistive MOSFETs (Fig. 8.20). The drive unit must be capable to deliver the gate current in the range of

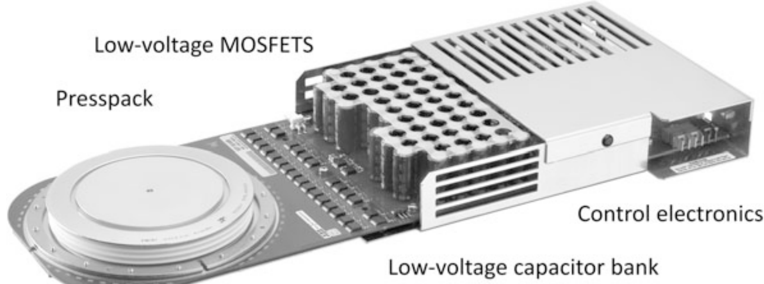


Fig. 8.20 GCT with gate drive unit. The gate wiring consists of wide metal layers on the PCB and is connected in form of a ring to the device. Fig. from [Vem15] © EPE ECCE Europe 2015 – European Conference on Power Electronics and Applications

the anode current within one microsecond. This is a severe challenge; especially the resistance R_G and the parasitic inductance L_G in the gate drive circuit must be extremely low. In the gate drive circuit, the differential equation holds [Lin06]:

$$L_G \cdot \frac{di_G}{dt} + R_G \cdot i_G = V_G \tag{8.19}$$

The solution of this equation is

$$i_G(t) = \frac{V_G}{R}\left(1 - \exp\left(-\frac{R_G}{L_G} \cdot t\right)\right) \tag{8.20}$$

The voltage V_G is limited by the avalanche breakdown voltage of the pn-junction between gate and cathode. The requirement is that the gate current must increase within a time t_{gs} up to the anode current I_A. Thus, the maximally allowed inductance L_G is given by

$$L_G = -\frac{R_G \cdot t_{gs}}{\ln\left(1 - \frac{R_G}{U_G} \cdot I_A\right)} \tag{8.21}$$

The maximally allowed inductance L_G is shown in Fig. 8.21 as a function of the gate resistance R_G for the conditions V_G = 20 V, anode current I_A = 4000 A and allowed t_{gs} = 1 µs. The resistance R_G and the parasitic inductance L_G must be kept extremely small to allow an operation as GCT. For R_G > 5 mOhm, no solution exists for the specified values of V_G, I_A and t_{gs}.

It must be considered that L_G as well as R_G consist not only of the external wiring. The resistance R_{on} of the used MOSFETs, the wiring, the resistance of the connections within the semiconductor housing and the resistance of the gate metallization of the semiconductor die contribute to R_G. The MOSFETs must have a blocking capability slightly higher than 20 V. In this voltage range very low-resistive

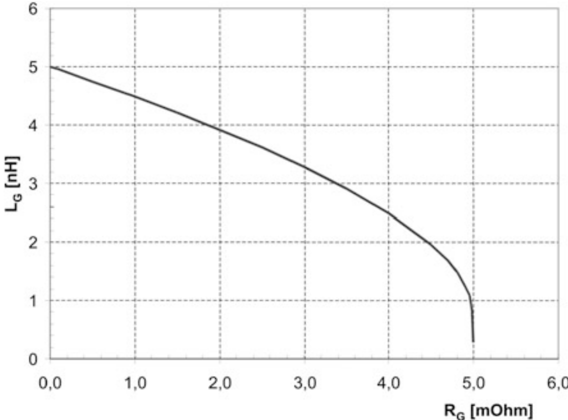


Fig. 8.21 Maximum allowed parasitic inductance L_G in the gate circuit of a 4000 A GCT. Figure according [Lin06]

Si-MOSFETs are available today. In this operation mode the npn partial transistor is turned off abruptly. The positive feedback loop in the thyristor is interrupted. The extraction of the charge carries below the emitter fingers still starts at the edges of the fingers, but the problem of narrow filaments during turn-off is solved to a great extent. The GCT can be operated without RCD-snubber.

There is no requirement for a reverse blocking capability of a GCT. Therefore, it is possible to implement an n-doped buffer layer in front of the anode region and to design the device with a moderate trapezoidal shape of the electric field (PT dimensioning). With this additional buffer layer, the thickness w_B of the n^--layer can be reduced. Consequently, conduction losses and turn-off losses are diminished.

A drive unit which must supply a current as high as the current to be controlled requires high effort without any doubt. But the power, which must be supplied by the driver unit, is not higher than the driver power for a GTO thyristor with the same current turn-off capability. In contrast, it is reduced. In a GCT the negative gate current increases rapidly and the high current flows only during a short time interval. The total charge which is extracted via the gate is even smaller than in a GTO thyristor, since in a GTO still new carriers are injected by the emitter junction J_3 during the increase of the negative gate current in the interval t_{gs} (Fig. 8.19), i.e. the n^+-emitter replenishes part of the charge that must be extracted by the gate driver. In a GCT, only the charge stored before turn-off must be removed. The GCT can be operated with even half of the drive power of a GTO [Lin06]. Although only few modifications of the silicon die are made in a GCT compared to a GTO, the CCT constitutes a significant progress. The main weaknesses of the GTO are solved to a large extend.

The Integrated Gate Commutated Thyristor (IGCT) contains a monolithically integrated diode, see Fig. 8.22. The carrier lifetime adjustment in the diode area is executed with particle irradiation, see Sect. 4.9. There must be no cumbersome compromises in diode performance [Kla97].

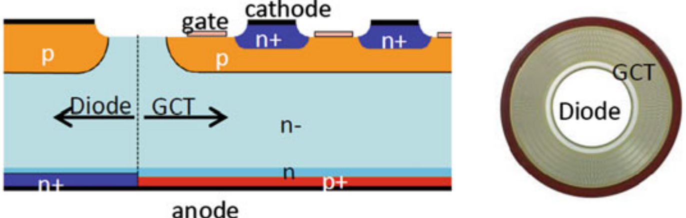


Fig. 8.22 IGCT with integrated Diode. Fig. adapted from [Kla97]

Figure 8.23 shows the IGCT-presspack internal construction. The center of the device (middle, top) is the diode area. It is contacted on bottom with a uniform molybdenum disk (bottom left) and on top side with 2 molybdenum disks for GCT and diode are. The top-side pressure camp, bottom middle, contains the ring-shaped gate with radial interconnections. They realize the low-inductive interconnection to the PCB.

A further progress for the GCT is the corrugated p-base, shown in Fig. 8.24. It is fabricated by an ion implantation through a mask covering mainly the n^+ segment. After the drive, the p-base has two different junction depths with an intermediate region with more shallow penetration depth. [Wik07]. An increased of the maximal current that can be turned-off was found, the increase is a factor of 1.4 at 25 °C and a factor of 1.1 at 125 °C.

A reason is that the position of highest current at turn-off, which is in the middle of the emitter finger (Fig. 8.16), is now different from the position of high electric field when the junction is subject to reverse voltage, as it is the case at turn-off. This

Fig. 8.23 IGCT Presspack internal construction. Reprinted with permission from ABB

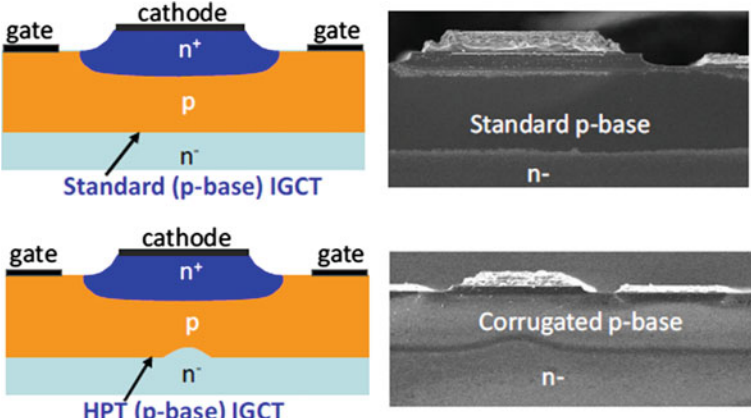


Fig. 8.24 High Power Technology (HPT) GCT with corrugated p-base design compared to a standard uniform p-base junction. Fig. from [Vem15] © EPE ECCE Europe 2015 – European Conference on Power Electronics and Applications

laterally modulated electric field helps to defocus the generated holes from the center of the emitter finger during dynamic avalanche [Wik07]. For further details on dynamic avalanche, see Chap. 12.

The GCT hat the potential of very low conduction losses. It is shown in [Rah13] that for 2.5 kV GCT (91 mm diameter) a forward voltage drop even below 1 V can be achieved at 1500 A, if the intended switching frequency is 125 Hz. This can be used for high-current applications where very high efficiency is required.

References

[Bec80] Becke, H.W., Misra, R.P.: Investigations of gate turn-off structures. Int. Electron Devices Meet. **26**, 649–653 (1980)

[Ben99] Benda, V., Govar, J., Grant, D.A.: Power Semiconductor Devices. Wiley, New York (1999)

[Chu70] Chu, C.K.: Geometry of thyristor cathode shunts. IEEE Trans. Electron Devices **17**(9), 687–690 (1970)

[Chk05] Chukaluri, E.K., Silber, D., Kellner-Werdehausen, U., Schneider, C., Niedernostheide, F.J., Schulze, H.J.: Recent developments of high-voltage light-triggered thyristors. In: Proceedings 36th Power Electronics Specialists Conference PESC '05 pp. 2049–2052 (2005)

[Gen64] Gentry, F.E., Gutzwiller, F.W., Holonyak, N., Von Zastrow, E.E.: Semiconductor Controlled Rectifiers: Principles and Applications of p-n-p-n Devices. Principle-Hall Inc, New York (1964)

[Gen65] Gentry, F.E., Scace, R.I., Flowers, J.K.: Bidirectional triode pnpn-switches. Proc. IEEE **53**, 355–369 (1965)

[Gen68] Gentry, F.E., Moyson, J.: The amplifying gate thyristor. IEEE Int. Electron Devices Meet. **14**, 110 (1968)

[Ger65] Gerlach, W.: Thyristor mit Querfeldemitter, Z. angew. Phys 17 pp. 396–400 (1965)

[Ger79] Gerlach, W.: Thyristoren. Springer, Berlin (1979)

[Gru96] Gruening, H., Odegard, B., Ress, J., Weber, A., Carroll, E., Eicher, S.: High-power hard-driven GTO module for 4SkV/3kA snubberless operation. In: Proceedings of the PCIM, pp. 169–183 (1996)

[Her65] Herlet, A.: The maximum blocking capability of silicon thyristors. Solid-State Electron **8**(8), 655–671 (1965)

[Kla97] Klaka, S., Linder, S., Frecker, M.: A family of reverse conducting gate commutated thyristors for medium voltage drive applications. In: Proceedings PCIM, Hong Kong (1997)

[Lin06] Linder, S.: Power Semiconductors. EPFL Press, Lausanne, Switzerland (2006)

[Mol56] Moll, J.L., Tannenbaum, M., Goldey, M., Holoniak, N.: p-n-p-n transistor switches. Proc. IRE **44**, 1174–1182 (1956)

[Nak95] Nakagawa, T., Tokunoh, F., Yamamoto, M., Koga, S.: A new high power low loss GTO. In: Proceedings of the 7th ISPSD, pp. 84–88 (1995)

[Nie01] Niedernostheide, F.J., Schulze, H.J., Kellner-Werdehausen, U.: Self protected high power thyristors. In: Proceedings of the PCIM, Nuremberg, 51–56 (2001)

[Nie07] Niedernostheide, F.J., Schulze, H.J., Felsl, H.P., Laska, T., Kellner-Werdehausen, U., Lutz, J.: Thyristors and IGBTs with integrated self-protection functions. IET J. Circuits, Devices Syst. **1**(5), 315–320 (2007)

[Prz09] Przybilla, J., Dorn, J., Barthelmess, R., Kellner-Werdehausen, U., Schulze, H.J., Niedernostheide, F.J.: Diodes and thyristor – Past, presence and future. In: Proceedings 13th European Conference on Power Electronics and Applications, EPE '09 (2009)

[Rad71] Raderecht, P.S.: A review of the 'shorted emitter' principle as applied to p-n-p-n silicon controlled rectifiers. Int. J. Electron. **31**(6), 541 (1971)

[Rah13] Rahimo, M., Arnold, M., Vemulapati, U., Stiasny, T.: Optimization of High Voltage IGCTs Towards 1 V On-State Losses, pp. 613–620. Nuremberg, Proc. PCIM Europe (2013)

[Scu01] Schulze, H.J., Niedernostheide, F.J., Kellner-Werdehausen, U.: Thyristor with Integrated Forward Recovery Protection. In: Proceedings of the ISPSD, Osaka, pp. 199–202 (2001)

[Shi99] Shimizu, Y., Kimura, S., Kozaka, H., Matsuura, N., Tanaka, T., Monma, N.: A study on maximum turn-off current of a high-power GTO. IEEE Trans. Electron Devices **46** (2), 413–419 (1999)

[SID97] SIDACtor Protection Thyristors http://www.littelfuse.com/products/sidactor-protection-thyristors.aspx called up Sept. 2017

[Sil75] Silber, D., Füllmann, M.: Improved gate concept for light activated power thyristors. Int. Electron Devices Meet. **21**, 371–374 (1975)

[Sil76] Silber, D., Winter, W., Füllmann, M.: Progress in light activated power thyristors. IEEE Trans. Electron Devices **23**(8), 899–904 (1976)

[Vem15] Vemulapati, U., Rahimo, M., Arnold, M., Wikström, T., Vobecky, J., Backlund, B., Stiasny, T.: Recent advancements in IGCT technologies for high power electronics applications. In: Proceedings EPE'15 ECCE-Europe (2015)

[Wik07] Wikström, T., Stiasny, T., Rahimo, M., Cottet, D., Streit, P.: The Corrugated P-Base IGCT – a New Benchmark for Large Area SQA Scaling, pp. 29–32. Jeju, Korea, Proc. ISPSD (2007)

[Wol66] Wolley, E.D.: Gate turn-off in pnpn devices. IEEE Trans. El. Devices ED-**13**, 590–597 (1966)

Chapter 9
MOS Transistors and Field Controlled Wide Bandgap Devices

9.1 Function Principle of the MOSFET

The MOSFET basic structure was investigated early [Hof63]. For the comprehension of the function of a MOSFET (Metal Oxide Semiconductor Field Effect Transistor), the surface of the semiconductor may be examined at first. The surface of a semiconductor is always a disturbance of the ideal lattice due to the lack of neighboring atoms. Therefore, a thin oxide will always be built up on the surface or other atoms and molecules are adsorbed. Thus, these surface layers are normally electrically charged.

A p-type semiconductor may be given as an example. Assumed is a positive charge on the surface (Fig. 9.1).

For a small positive charge we obtain

$$|qV_S| < E_i - E_F:$$

The hole-concentration/density on the surface is being reduced. The conduction band and the valence band are bent downwards. A depletion zone of the thickness h_d is built up.

A higher positive charge yields

$$|qV_S| > E_i - E_F:$$

In this case, the conduction and valence band are bent even stronger. In a small region at the surface the Fermi level is now closer to the conduction band than to the valence band. An inversion layer of the thickness h_i is being formed, in which electrons are the majority charge carriers.

Next to this is the depletion layer with the thickness h_d, which separates the inversion layer from the p-type area.

© Springer International Publishing AG 2018
J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8_9

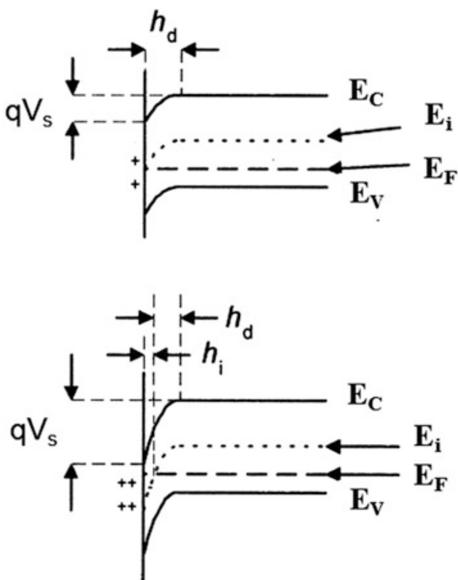


Fig. 9.1 Semiconductor surface. p-type semiconductor, positive charge on the surface

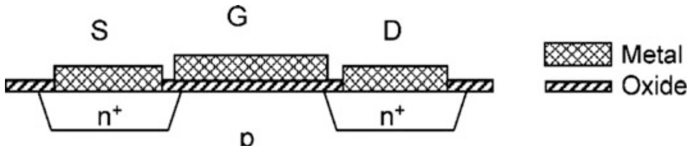


Fig. 9.2 Lateral n-channel MOSFET. Description: *S* Source, *G* Gate, *D* Drain

Having a negative charge on the p-type semiconductor, an accumulation layer of holes will be formed. The n-type semiconductor behaves similarly: With a positive charge at the surface, the accumulation layer is formed, with a negative surface charge a depletion zone is formed, with an increased negative charge the inversion layer.

Next, we assume a thin oxide film on the p-type semiconductor and apply a metallization on it. On this metal film a positive voltage is applied. Furthermore, two n^+-areas are added and bonded as source and drain region. We have the simplest case of a lateral MOS field effect transistor shown in Fig. 9.2 [Hof63].

A positive voltage has the same effect as the positive surface charge: When a sufficient positive voltage at the gate is applied, both n-areas are connected by the inversion layer. Due to the gate voltage $V_G > V_T$ a current can flow between the drain and the source.

Gate-Source Threshold-Voltage V_T (n-channel-MOSFET):
The threshold voltage *is the gate voltage, at which the generated electron concentration equals the concentration of the acceptors.*

It has to be distinguished:

n-channel MOSFET: An n-type channel is formed in a p-area.
p-channel MOSFET: A p-type channel is formed in a n-area.

On closer inspection, it has to be considered that the oxide contains positive charges on the boundary surface to the semiconductor. These charges are in the order of 5×10^9 to 1×10^{11} cm^{-2}. Moreover, the gate area of power-MOSFETs consists of a heavily n-doped poly silicon layer (see Figs. 9.4 and 9.5) and a potential difference already exists between gate and semiconductor due to the differing positions of the Fermi level in the n$^+$-doped poly silicon and in the p-type semiconductor (concerning the n-channel MOSFET). Both effects function in the same way as an external positive gate-voltage and result in a reduction of the threshold voltage V_T. In case of a low doping of the p-area and a high oxide charge of the n-channel MOSFET, V_T is negative, even without gate voltage a channel exists. The definition of the threshold voltage mentioned above remains valid.

It has to be distinguished:

Depletion type: $V_T < 0$. The device is normally on and does not block before a negative gate-source-voltage $V_G < V_T$ is applied.
Enhancement type: $V_T > 0$. An n-channel only develops when $V_G > V_T$ (normally off device).

Usually, MOSFETs of the enhancement type are used in power electronics, because of the normally-off feature that is required in voltage source converters. Almost always n-channel MOSFETs are used, being more advantageous since the mobility of the electrons is much higher than that of the holes (see Chap. 2). Typically, the threshold voltage of modern devices is adjusted between 2 and 4 V.

9.2 Structure of Power MOSFETs

The configuration shown in Fig. 9.2 will sustain little drain-source voltage. Thus, a structure such as in Fig. 9.3, which is named DMOS (D = double diffused), is used from 10 V upwards. The lowly doped n$^-$-area, the drain extension area, takes over the blocking voltage.

Lateral DMOS transistors are frequently used in power ICs and in monolithic integrated power semiconductor-circuits ("smart power"). But they have the disadvantage of having a low current-load capacity, because the n$^-$-area consumes a large part of the surface of the semiconductor. If real "power" has to be controlled, a

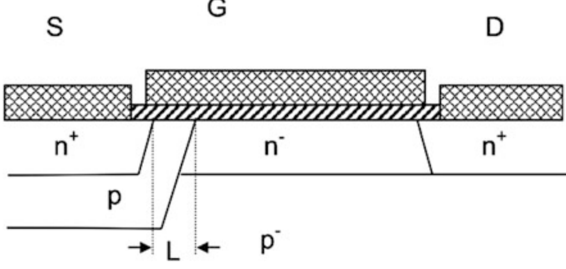


Fig. 9.3 Lateral DMOS

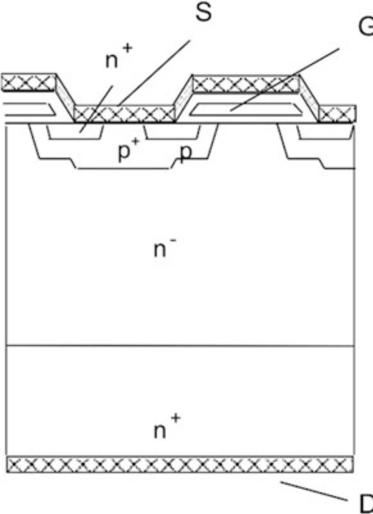


Fig. 9.4 Vertical DMOS transistor. The gate electrode is poly-silicon

vertical MOSFET is realized by arranging the area for the space charge vertically [Lid79], see Fig. 9.4. Consequently, the volume of the semiconductor is utilized and the surface can be used for the formation of the gate-source cell structure.

On the surface of the semiconductor the individual cells are formed, which consist of p-wells and n^+-source areas. A cross section of a cell can be seen in Fig. 9.4. The p-well is connected to the source metallization, so that the parasitic npn-transistor is shorted. In order that the short has very little resistance, the doping is increased at this spot by an additional p^+-implantation, followed by a diffusion step. At the edges of the well is the channel, which is covered with the thin gate oxide. Above the oxide the gate electrode is applied, usually consisting of a heavily doped n^+ poly-silicon. At one spot, mostly in the center of the chip, the gate electrode is brought to the surface to form a gate pad to which a bond wire can be attached.

Because the current has to flow through the inversion channel, many single cells are formed to gain a larger width of the channel. An example is given in Fig. 9.5. Here, the cells are of square form and arranged in a square pattern. The

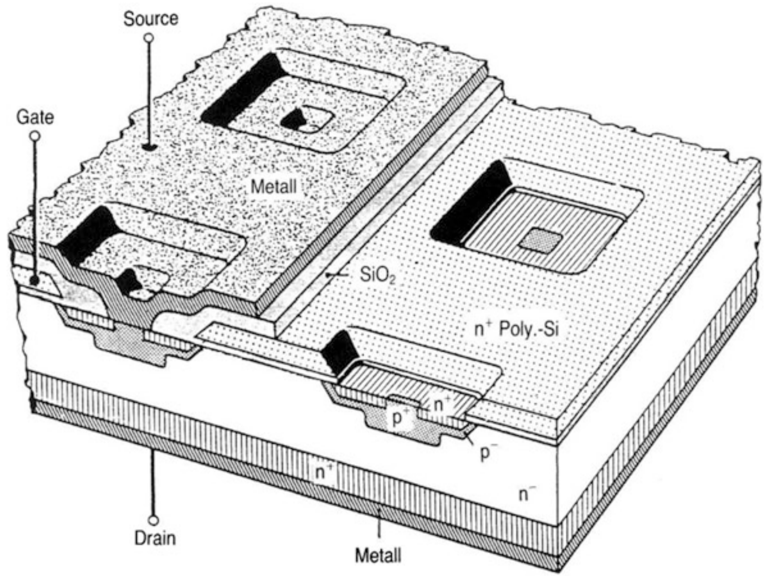


Fig. 9.5 Cell structure of a vertical DMOS. From [Ste92]

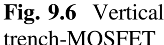
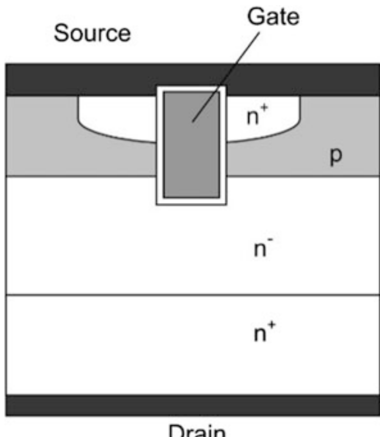


Fig. 9.6 Vertical trench-MOSFET



semiconductor surface area is used even more effectively with a hexagonal pattern, where the single cells are hexagonal, the so-called HEXFET structure [Col79].

The vertical DMOS (also called VDMOS) transistors are used in a wide field of applications, such as computer and electronic equipment power supplies with power factor correction front ends. Since the second half of the 1990s, with the introduction of the trench-MOS [Sod99] a further improvement has been made, in which the channel area is vertically arranged (Fig. 9.6). Due to this, a much smaller on-state resistance can be gained especially in the lower voltage range < 100 V.

9.3 Current-Voltage Characteristic of MOS-Transistors

Figure 9.7 shows the current-voltage characteristics of the MOSFET. The device is in the blocking state as long as a positive voltage V_D between drain and source is given and V_G is smaller than the threshold voltage V_T. The blocking voltage of the MOSFET is limited by the avalanche breakdown. Because the npn-transistor is suspended by a low resistance short, the blocking voltage of the MOSFET corresponds to the blocking voltage of the diode, which is formed by the p-well, the low doped base region and the n^+-layer.

A current carrying channel is built for $V_G > V_T$, resulting in the given current-voltage characteristics. Similar to the current gain of bipolar transistors a transconductance is defined here. For low voltages V_D, the current-voltage characteristics have the form of a straight line. For a defined gate voltage V_G the resistance $R_{DS(on)}$ (= R_{on}) is indicated.

The transition between the ohmic region and the pinch-off region is called the quasi saturation. This region is described by a parabolic curve.

In the reverse direction of the MOSFET a forward biased diode is present. As for a power diode, this diode forward characteristic is often approximated by a threshold voltage V_{F0} and a differential resistance.

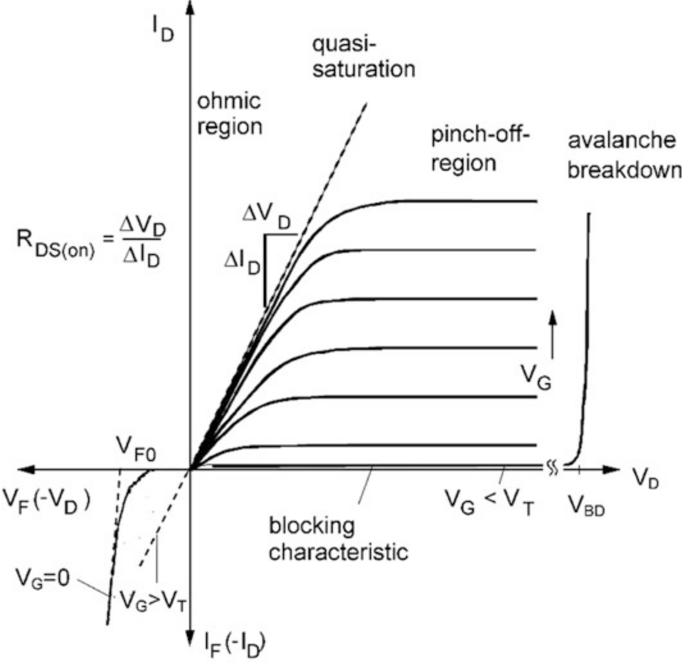


Fig. 9.7 Current-voltage characteristic of the MOSFET. Figure adapted from [Nic00]

9.4 Characteristics of the MOSFET Channel

Due to the oxide layer of the gate and the gate electrode a capacitor is built above the channel. It's area specific capacity can be described with

$$C_{ox} = \frac{\varepsilon_0 \cdot \varepsilon_r}{d_{ox}} \tag{9.1}$$

The oxide has the thickness d_{ox} (i.e. < 100 nm). The relative permittivity of the oxide is $\varepsilon_r = 3.9$ (for SiO_2). Having the gate voltage V_G higher than the threshold voltage V_T an inversion channel is created, as it is shown in Fig. 9.8a. As long as the voltage drop caused by the current in the channel can be neglected, the charge of the inversion channel yields

$$Q_s = C_{ox} \cdot (V_G - V_T) \tag{9.2}$$

The carriers, forming this charge, are available for the current transport in the inversion channel. As long as the pinch-off can be neglected in the channel, the resistance of the channel is

$$R_{ch} = \frac{L}{W \cdot \mu_n \cdot Q_s} = \frac{L}{W \cdot \mu_n \cdot C_{ox} \cdot (V_G - V_T)} = \frac{1}{\kappa \cdot (V_G - V_T)} \tag{9.3}$$

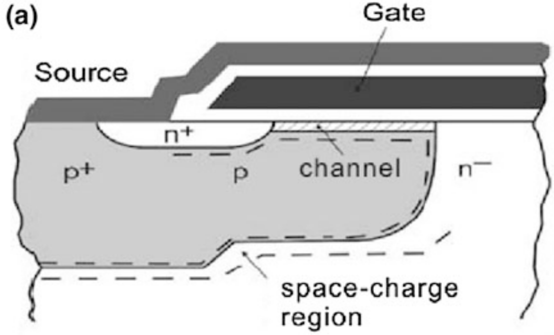
where L is the length of the channel (for example 2 µm, see Fig. 9.3) and W is the entire width of it. In Fig. 9.3, W is vertical to the plane of projection and corresponds to the circumference of the single cell multiplied by the number of cells (see Fig. 9.5). Having a high cell density, a large W is achieved, and thus a low channel resistance. W can amount to some 100 m per cm^2 of a chip surface in modern semiconductor devices. The parameters dependent on geometry can be summarized in

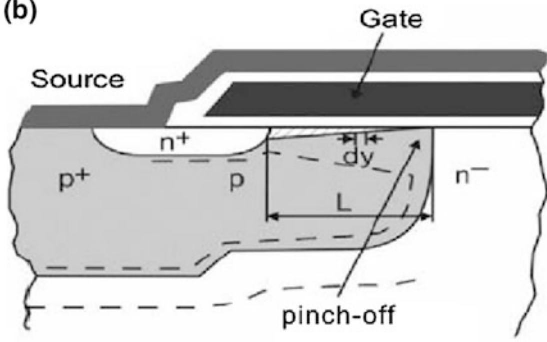
$$\kappa = \frac{W \cdot \mu_n \cdot C_{ox}}{L} \tag{9.4}$$

Equation (9.3) is valid for the ohmic region in Fig. 9.7; that is for the area in which the voltage drop across the channel can be neglected with regard to its influence on Q_S.

As can be seen in (9.3), the channel resistance is affected by the carrier mobility. In Sect. 2.6 it has been shown, that the mobility μ_p only amounts to approximately one third of μ_n. It is for this reason that, whenever it is possible, n-channel MOSFETs are used in power electronics.

With increasing current a voltage drop $V(y)$ develops across the channel. The channel narrows, see Fig. 9.8b. Along the length of the channel y a charge $Q(y)$ will exist. For an element dR of the resistance R_{CH}, with (9.3) one obtains





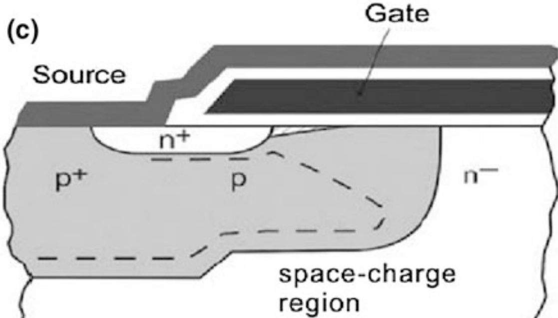


Fig. 9.8 MOSFET channel. **a** Ohmic region, $V_D \ll V_G - V_T$, **b** pinch-off, $V_D = V_G - V_T$, **c** channel length shortening $V_D \gg V_G - V_T$

$$dR = \frac{dy}{W \cdot \mu_n \cdot Q(y)} \tag{9.5}$$

with

$$Q(y) = C_{ox} \cdot (V_G - V_T - V(y)) \tag{9.6}$$

In a segment dR the voltage drop is

$$dV = I_D \cdot dR \tag{9.7}$$

Inserting (9.6) and (9.5) in (9.7) yields

$$I_D = W \cdot \mu_n \cdot C_{ox} \cdot (V_G - V_T - V(y)) \cdot \frac{dV}{dy} \tag{9.8}$$

The voltage V_D drops between the boundaries $y = 0$ and $y = L$:

$$\int_0^L I_D \cdot dy = W \cdot \mu_n \cdot C_{ox} \cdot \int_0^{V_D} (V_G - V_T - V(y)) \cdot dV \tag{9.9}$$

Integration leads to the following $I_D(V_G, V_D)$ characteristic:

$$I_D = \kappa \cdot \left((V_G - V_T) \cdot V_D - \frac{1}{2} V_D^2 \right) \tag{9.10}$$

for $V_D \leq V_G - V_T$. The characteristic corresponds to the parabolic section (quasi saturation) in Fig. 9.7. For small V_D it verges into

$$I_D = \kappa \cdot (V_G - V_T) \cdot V_D \tag{9.11}$$

and corresponds to the ohmic region, as already indicated in (9.3). The passing into the pinch-off region results from Eq. (9.10) for $dI_D/dV_D = 0$. Afterwards the channel is pinched off for

$$V_D = V_G - V_T \tag{9.12}$$

For a larger V_D, inserting (9.12) in (9.10) yields the characteristics in the pinch-off region. In this region, the current remains almost constant even for increased voltage V_D

$$I_{Dsat} = \frac{\kappa}{2} \cdot (V_G - V_T)^2 \tag{9.13}$$

The transconductance is defined by

$$g_{fs} = \left. \frac{\Delta I_D}{\Delta V_G} \right|_{V_D = const} \tag{9.14}$$

By differentiating (9.13) we obtain

$$g_{fs} = \kappa \cdot (V_G - V_T) \tag{9.15}$$

According to (9.13), the current I_{Dsat} is independent of V_D. However, in reality the electric field penetrates into the p-zone, when V_D is strongly increased (see Fig. 9.8c) and the channel becomes shortened. This shortening of the channel-length involves a slight ascent of the current at high voltages.

The current-voltage characteristics (9.10) can be found in numerous textbooks. However, comparing it with practically realized power devices, this equation is not very satisfactory. In the derivation it was not considered that a depletion zone is formed below the channel. This zone widens while the channel is narrowing, as it is indicated in Fig. 9.8c. A derivation of the current-voltage characteristics in consideration of the space charge can be found in [Gra89, Sze81]. This leads to

$$I_D = \kappa \cdot \left((V_G - V_T) \cdot V_D - \frac{1}{2} \left(1 + \frac{C_D}{C_{ox}} \right) V_D^2 \right) \tag{9.16}$$

for $V_D \leq V_G - V_T$. It contains the area specific capacity of the space charge region

$$C_D = \sqrt{\frac{\varepsilon_0 \cdot \varepsilon_r \cdot q \cdot N_A}{2 \cdot \Delta V_T}} \tag{9.17}$$

as it has been derived for the treatment of the pn-junction with the Eq. (3.109). The voltage ΔV_T in Eq. (9.17) corresponds to the voltage, which is necessary to develop a space charge region in the p-doped well with doping concentration N_A:

$$\Delta V_T = 2 \cdot \frac{k \cdot T}{q} \cdot \ln\left(\frac{N_A}{n_i} \right) \tag{9.18}$$

ΔV_T is necessary to fulfill the condition of strong inversion. ΔV_T amounts to approximately 0.81 V (resulting from a typical doping of the p-well of 1×10^{17} cm^{-3}). Considering the space charge this way, little changes at a very small voltages V_D. The approximation for the ohmic region in Eq. (9.11) remains the same. However, I_{Dsat} and g_{fs} vary. Equation (9.16) is effective as long as the voltage drop across the channel is smaller than ΔV_T that is $V_D < \Delta V_T$.

Apart from that, the reduced mobility in the channel has to be considered. Even without a lateral electric field, the mobility is already reduced compared to the values indicated in Fig. 2.12 of Chap. 2. The reason is the influence of the semiconductor surface. If a voltage $V(y)$ is formed above the channel, then a significant electric field develops in lateral direction. Equation (2.38) has to be consulted for the velocity of the electrons. For the electron mobility it yields

$$\mu_{nCH} = \frac{\mu_{e0}}{1 + \theta \cdot (V_G - V_T)} \tag{9.19}$$

In [Gra89] a suitable agreement with experiments is achieved, when the values of 600 $cm^2V^{-1}s^{-1}$ and 0.02 V^{-1} are used for μ_{e0} and θ.

The threshold voltage for the MOSFET is expressed with [Sze81]

$$V_T = V_{p,S} - \frac{Q_{ox}}{C_{ox}} + \Delta V_T + \frac{\sqrt{2qN_A\varepsilon_0\varepsilon_r\Delta V_T}}{C_{ox}} \tag{9.20}$$

with the potential difference n^+-poly/p-Si

$$V_{p,S} = \frac{kT}{q} \ln \frac{N_{D,poly} \cdot N_A}{n_i^2} \tag{9.21}$$

and the charge due to oxide traps Q_{ox}. The threshold voltage is temperature dependent and decreases with temperature primarily due to the strong temperature dependency of n_i in (9.18) and (9.21).

Measurements of the threshold voltage are in most cases executed with shorted gate and drain and applying a small current I_D. The found temperature dependency with this method is stronger than predicted by Eq. (9.20). Main reason is that at low currents the condition of strong inversion is not fulfilled. If for sufficient high drain voltages ($V_D > 5$ V) V_T is extracted at the intercept of the extrapolated tangent of $\sqrt{I_D}$ versus V_G at $V_{GS} \gg V_T$ [Lee82], a reasonable agreement with (9.20) is found.

9.5 The Ohmic Region

For the MOSFET's ohmic resistance not only the channel resistance has to be considered. Indeed, the resistance of the low doped middle region already dominates in devices with a blocking voltage of 50 V upwards. Because this layer is grown by epitaxy for vertical MOSFETs, the designation R_{epi} is commonly used. Figure 9.9 describes the structure of the MOSFET with a given path of the charge carriers (electrons) and with different parts of the resistance

$$R_{on} = R_{S^*} + R_{n+} + R_{ch} + R_a + R_{epi} + R_s \tag{9.22}$$

For MOSFETs with blocking voltages < 50 V, effort is made to reduce the channel resistance. The near-surface parts are reduced by increased cell density (larger W, see Eq. (9.1)). Most progress is achieved with the trench-cell (Fig. 9.5), where additionally the resistance R_a is eliminated.

In Table 9.1, the shares of the particular parts of the resistance are specified for a 30 V vertical MOSFET with planar cells together with a corresponding 600 V MOSFET.

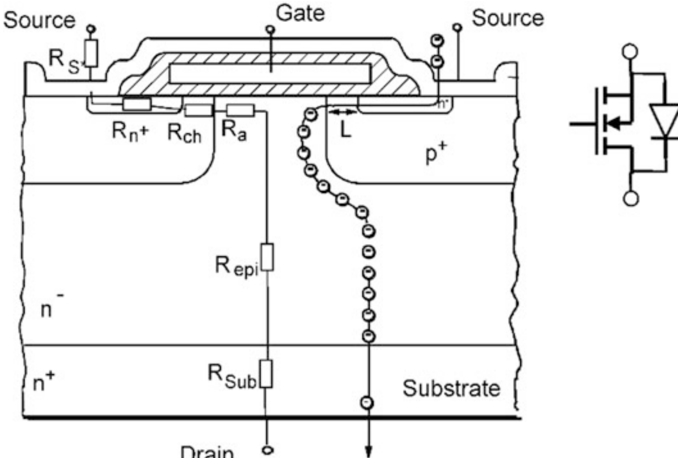


Fig. 9.9 Current path and resistances in a MOSFET. Figure inspired by [Lor99]

Table 9.1 Resistance R_{on} for MOSFETs with different blocking voltages

		V_{DS} = 30 V	V_{DS} = 600 V
R_S*	Package	7%	0.5%
R_{n^+}	Source layer	6%	0.5%
R_{CH}	Channel	28%	1.5%
R_a	Accumulation layer	23%	0.5%
R_{epi}	n^--layer	29%	96.5%
R_{Sub}	Substrate	7%	0.5%

Values from [Lor99]

The resistance R_{epi} of the low doped region is identical with the voltage drop across the low doped base region of a unipolar device, which has been given in Chap. 6 for Schottky diodes (Eq. 6.8):

$$R_{epi} = \frac{w_B}{q \cdot \mu_n \cdot N_D \cdot A} \tag{9.23}$$

If the device is designed for higher voltages, w_B has to be chosen larger as well as N_D smaller. Having a conventional MOSFET, the resistance can be calculated according to the shown approach in dependence of the voltage for which the device is designed. The lowest resistance can be obtained for a light PT design, as indicated in Eq. (6.14):

$$R_{epi,\min} = 0.88 \cdot \frac{2 \cdot B^{\frac{1}{2}} \cdot V_{BD}^{\frac{5}{2}}}{\mu_n \cdot \varepsilon \cdot A} \tag{9.24}$$

Thus, the resistance increases more than with the square of the blocking voltage, namely with $V_{BD}^{2.5}$. Equations (9.23) and (9.24) or comparable equations (see Chap. 6) are handled as "unipolar limit" in the literature. Meanwhile, this limit has been broken by the principle of compensation structures.

9.6 Compensation Structures in Modern MOSFETs

The compensation principle for power MOSFETs has been introduced in commercially available products in 1998 with the 600-V CoolMOS™ technology [Deb98]. The basic principle behind the drastic $R_{on} \cdot A$ reduction compared to conventional power MOSFETs is the compensation of n-drift region donors by acceptors located in p-columns (also known as superjunction). Figure 9.10 shows the structure of a superjunction MOSFET compared to a conventional MOSFET. In the middle layer, p-columns are arranged. Their p-doping is adjusted to the value necessary for compensation of the n-regions. The compensating acceptors are located in lateral proximity to the drift region donors.

The result is a low effective doping in the entire voltage-sustaining region. An almost rectangular shape of the electric field is obtained, as can be seen in the lower part of Fig. 9.10. For this field shape, the highest voltage can be absorbed at a given

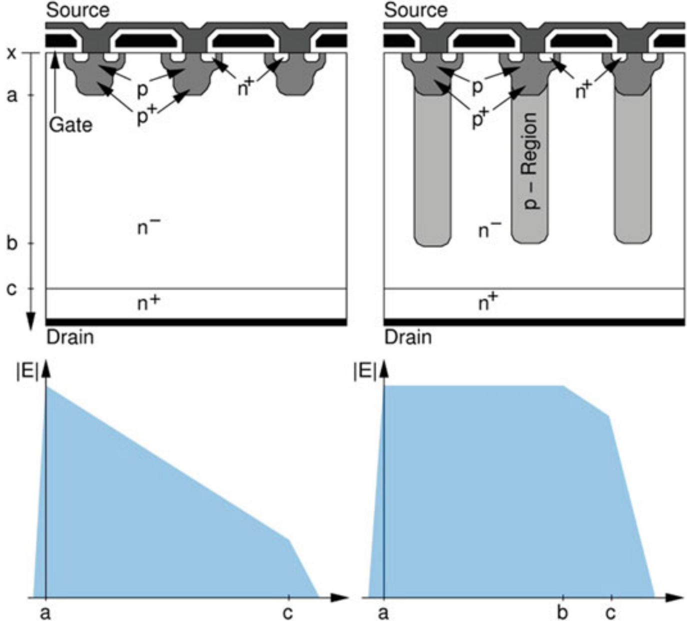


Fig. 9.10 Standard MOSFET and superjunction MOSFET

thickness. The doping of the n-layer can be lifted in so far, as it is technologically possible to compensate it through an equally large p-doping. In this process it has to be considered that the area of the n- respectively n^--region is decreased.

By means of the compensation principle, the coupling of the blocking voltage and the doping is neutralized and a degree of freedom for the adjustment of the n-doping is obtained. Since, according to (9.23), the n-doping determines the resistance in unipolar devices, the resistance can be lowered drastically.

In case of the rectangular shape of the electric field, the avalanche breakdown can be calculated by using the ionization integral (3.71), with the approach proposed by Shields and Fulop with $n = 7$ (see the sections about the PT diode, Eq. (5.12)), resulting in

$$w_B = B^{\frac{1}{6}} \cdot V_{BD}^{\frac{7}{6}} \tag{9.25}$$

with $B = 2.1 \times 10^{-35} \text{cm}^6 \text{V}^{-7}$ as derived in Eq. (3.83). Inserting (9.25) into (9.23) yields

$$R_{epi} = \frac{2 \cdot B^{\frac{1}{6}} \cdot V_{BD}^{\frac{7}{6}}}{q \cdot \mu_n \cdot N_D \cdot A} \tag{9.26}$$

The factor 2 in the numerator of (9.26) is obtained due to the simplified account that the width of the p-columns is equal to the width of the n-zones. Only the n-areas contribute to the conduction, hence only half of the area is available.

Figure 9.11 compares the relation between R_{epi} and the blocking voltage for the conventional design (9.24) and for the superjunction device (9.26). Here, the doping $N_A = N_D = 2 \times 10^{15}\,\text{cm}^{-3}$ is chosen for the superjunction device. Furthermore, half of the total area is assumed to carry the electron current.

The following consequences can be derived for this very simplified case.

Under blocking conditions the space charge laterally penetrates into the n- and p-region. This is shown in Fig. 9.12. In Fig. 9.12 it is assumed that p- and n-region

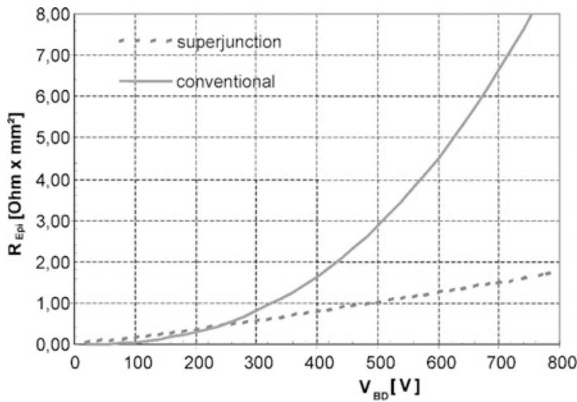


Fig. 9.11 Resistance R_{epi} as a function of the blocking voltage for the conventional MOSFET and for the superjunction device

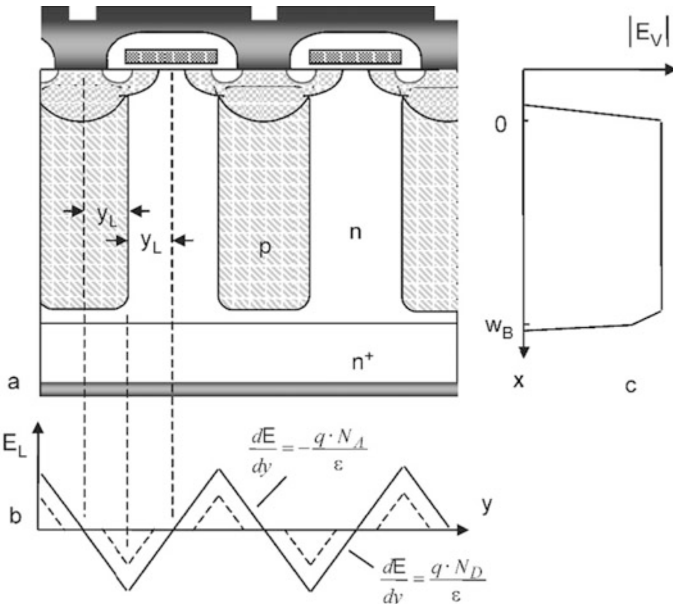


Fig. 9.12 Superjunction MOSFET. **a** Simplified structure **b** electric field in lateral direction in the region of the columns **c** electric field in vertical direction

are doped equally; the p-region has the doping N_A, the n-column the doping N_D with $N_A = N_D$. Moreover, both columns shall feature the same width, which constitutes $2 \cdot y_L$ respectively in Fig. 9.12. When a voltage is applied in reverse direction, at first the space charge region penetrates only laterally into the columns. For low voltages, the dashed line in Fig. 9.12b indicates the magnitude of the electric field along a section in lateral direction in the region of the columns. With increasing voltage, the space charges will eventually meet in the center of the respective columns, as shown by the solid line in Fig. 9.12b. Now all acceptors and donors are ionized.

With further increase of the voltage the zigzag line in Fig. 9.12b is lifted. This results in a structure similar to a corrugated iron roof. In vertical direction the electric field as shown in Fig. 9.12c is obtained.

In lateral direction, the respective p- and n-regions have to be penetrated entirely by the field. The expansion of the electric field at the avalanche breakdown into an n-region with the doping N_D has been given in Eq. (3.85). The indicated width w equals half the n-region y_L. A doping $N_D = 2 \times 10^{15}\ \mathrm{cm}^{-3}$ yields $y_L = 11\ \mu\mathrm{m}$. The width of the p- as well as the n-regions has to be smaller than $2 \cdot y_L$, otherwise breakdown occurs in lateral direction.

Therefore, the doping in Eq. (9.26) is connected to the width of the columns; a higher doping N_D demands a smaller y_L. Equation (3.85) inverted for N_D and inserted in Eq. (9.26) yields

$$R_{epi} = \frac{2 \cdot 2^{-\frac{3}{7}} B^{\frac{13}{42}} \cdot y_L^{\frac{8}{7}} \cdot V_{BD}^{\frac{7}{6}}}{\varepsilon \cdot \mu_n \cdot A} \tag{9.27}$$

Analogous considerations are leading to a similar result in [Zin01]. Equation (9.27) shows that the resistance can be reduced even more as shown in Fig. 9.11. This requires a yet smaller y_L. Note that the line for the superjunction in Fig. 9.11 can be shifted to a lower value for R_{epi} for an increased doping and a finer pattern. However, to realize this in the vertical structure with a depth $w_B \gg y_L$ is a major technological challenge.

A more precise consideration, also including peaks of the electric field at the source and drain sided border of the space charge region, can be found in [Che01]. There, different arrangements of the columns are analyzed as well. Seen from the top of a device, the considerations of Figs. 9.10 and 9.12 would result in a stripe pattern of the p- and n-regions. However, a hexagonal arrangement could also be suitable.

The requirement for precise lateral n- and p-dose compensation limits the n-drift region doping. The higher the deviation in the charge balance, the more loss in blocking capability occurs. This effect increases with higher doping N_D and smaller y_L. The process window for deviation from charge balance gets narrower [Kon06]. Process technology finally limits the possibility to reduce R_{on} in this type of compensation power MOSFETs.

For breakdown voltages below 200 V, field-plate or oxide-bypassed MOSFETs are an excellent alternative [Lia01, Sie06c]. The device comprises a deep trench penetrating most of the n-drift region. An isolated field plate provides mobile charges required to compensate the drift region donors under blocking conditions as shown in Fig. 9.13. A voltage source dynamically provides electrons on the field plate and therefore precise lateral drift region compensation is ensured under all operating conditions.

The field plate isolation has to withstand the full source-drain blocking voltage of the device at the trench bottom; therefore oxide layers with thickness in the micron range have to be fabricated carefully with a special focus on avoiding thinning at the bottom trench corners and preventing generation of stress-induced

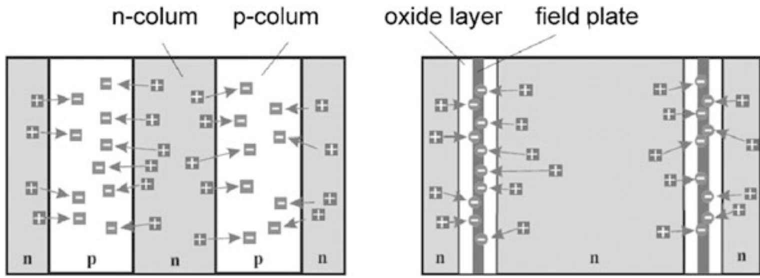


Fig. 9.13 Comparison of charge compensation by superjunction and by field-plates

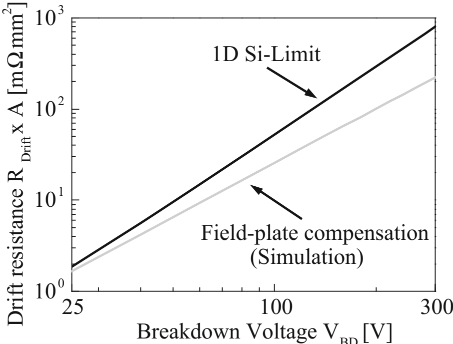


Fig. 9.14 Drift region resistance in dependency of breakdown voltage for field-plate devices, compared to the "silicon limit"

defects. In contrast to standard trench MOS structures that exhibit a linearly decreasing electric field with a maximum at the body/drift region pn-junction, the field-plate principle leads to a more constant field distribution. While in superjunction devices an almost homogeneous vertical field distribution is realized, the shape of the electric field of a field-plate device shows two peaks, one at the body/drift region pn-junction and the larger one at the bottom of the field-plate trench [Che05]. The necessary drift region length for a given breakdown voltage is reduced and the drift region doping can be increased, leading to a significantly reduced on-state resistance.

As in case of superjunction devices, the on-state resistance is reduced below the "Silicon Limit". Depending on the feasible device geometry, either superjunction or field-plate devices are advantageous in terms of device performance. In the voltage range between 30 and 100 V, the field plate compensation is superior to the superjunction compensation [Che05]. Figure 9.14 depicts the drift region resistance R_{epi} as a function of blocking voltage. A comparison is given between 2D Simulations [Paw08] for field plate compensation structures and the "silicon limit" given by Eq. (9.24). Due to the small device dimensions of the field-plate trench devices, the consequence is a larger doping density and thus a smaller drift region resistance [Paw08]. The simulation results in Fig. 9.14 agree well with experimental data.

9.7 Temperature Dependency of MOSFET Characteristics

The threshold voltage V_T given in (9.20) decreases with temperature. On the other hand, the mobility μ_n decreases with temperature, and it is included in all main factors that contribute to $R_{(on)}$ in Eq. (9.22). R_{on} increases with temperature. Since, for example, μ_n (125 °C) is approx. 0.5 · μ_n (25 °C), R_{on} doubles.

In the transfer characteristics is shown in Fig. 9.15 [Wil17]. At low V_G the saturation current I_{Dsat} increases with T due to the dominating influence of decreasing V_T. At high V_G, I_{Dsat} decreases due to the dominating mobility decrease. The intersection point is denoted as temperature compensation point TCP.

The temperature coefficient $ß_T$ is given by

$$ß_T = \frac{I_D(T_2) - I_D(T_1)}{\Delta T} \tag{9.28}$$

at $V_G = const$. Usually power MOSFETs are not designed to be operated in the pinch-off region, also denoted as linear region, see Fig. 9.7. If done, then operation below the TCP may lead to thermal instability. For small gate voltages, the drain current increases with temperature due the temperature dependency of the threshold voltage. If the device is operated within this region, a hot spot will develop and a thermal runaway occurs as shown in Fig. 9.16. At larger gate voltages, the drain current decreases with increasing temperature because the carrier mobility reduces at higher temperature. The larger the channel width, the more pronounced is the region of instability.

Figure 9.16 gives an example and shows temperature transients of a power MOSFET device at different voltage bias while keeping the pulse power at a the same value [Spi02]. Whether the device is stable or destroyed depends entirely on the bias conditions, i.e., the drain voltage, the drain current, and the duration of the pulse, not simply on the average power. For example, as shown in Fig. 9.16, a 90-W pulse power application realized as 6 A with $V_D = 15$ V is stable but at $V_D = 30$ V and $I_D = 3$ A, an operation point below the TCP, the device goes into destructive thermal runaway. Using the temperature coefficient of current $ß_T = \Delta I_D/\Delta T$ the operation points can be distinguished. Depending on the bias, the device remains stable of a hot spot grows.

At higher cell densities and large channel width W, as given in modern trench MOSFETs, the safe operating region for linear circuit mode is reduced [Cha16], whereby even relatively short durations of current saturation during start-up or shutdown may lead to device failure [Wil17].

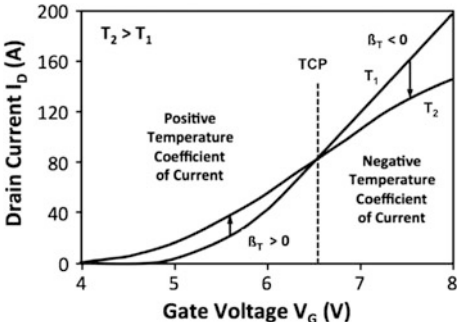


Fig. 9.15 Transfer characteristics at $V_D = const.$, $V_D > V_G - V_T$, at two different temperatures. Figure from R. Siemieniec, Infineon

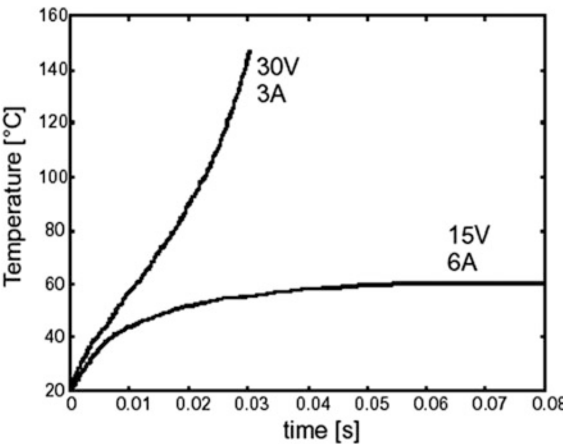


Fig. 9.16 Temperature transients of a power MOSFET at constant pulse power. Figure from P. Spirito, UniNa, redrawn from [Spi02]

9.8 Switching Properties of the MOSFET

Starting from the transition time of the charge carriers through the channel

$$\tau_t = \frac{L}{v_d} \tag{9.29}$$

with $v_d = \mu_n \cdot E$ and $E = V_{ch}/L$ yields

$$\tau_t = \frac{L^2}{\mu_n \cdot V_{ch}} \tag{9.30}$$

For instance, with $d = 2\ \mu\text{m}$, $V_{ch} = 1\ \text{V}$ and $\mu_n = 500\,\text{cm}^2\text{V}^{-1}\text{s}^{-1}$, we obtain the transition time $\tau_t \approx 80\,\text{ps}$. It corresponds to a transition frequency of

$$f_t \approx 12.5\,\text{GHz}$$

In practice, this is not accomplishable for a power MOSFET, since parasitic capacitances exist, leading to time-constants that determine the limiting frequency:

$$f_{co} = \frac{1}{2\pi \cdot C_{iss} \cdot R_G} \tag{9.31}$$

with $C_{iss} = C_{GS} + C_{GD}$ and $R_G = R_{G\text{int}} + R_{Gext}$. C_{iss} as well as the recommended gate resistance R_{ext} can be found in the data sheets. The internal gate resistance has to be asked for from the manufacturer.

Example: IXYS XFH 67 N10

$$C_{iss} = 4500\,\text{pF}$$
$$R_{ext} = 2\Omega, R_{\text{int}} \approx 1\,\Omega\,(\text{assumed})$$
$$\Rightarrow f_{co} = 12\,\text{MHz}$$

Figure 9.17 shows the structure of the MOSFET in which the parasitic capacitances are indicated. On the right, the equivalent circuit diagram of the MOSFET with its parasitic capacitances is shown. The inverse diode and some of the resistances are illustrated as well, only R_{CH} and R_{epi} are charted.

The turn-on and turn-off behavior shall be dealt with now under the condition of an inductive load, as an inductive load is usually existent in practice. The circuit corresponds to the one in Fig. 5.19. Figure 9.18 shows the turn-on waveform of the MOSFET with an inductive load. The characteristic quantities for the turn-on are:

t_d: *Turn-on delay time*
time until V_G reaches the threshold voltage V_T

$$t_d \sim R_G \cdot (C_{GS} + C_{GD})$$

t_{ri}: *Rise time*

During this time the current increases

$$t_{ri} \sim R_G \cdot (C_{GS} + C_{GD})$$

Due to the freewheeling diode for the inductive load, the reverse current peak I_{RRM} is added, see Figs. 5.20 and 5.21. During this time the voltage V_D remains virtually unaltered.

t_{fv}: *Voltage fall time*

Now the freewheeling diode takes over voltage and the voltage across the MOSFET drops. The capacitance C_{GD} (Miller capacitance) is being charged.

$$t_{fv} \sim R_G \cdot C_{GD}$$

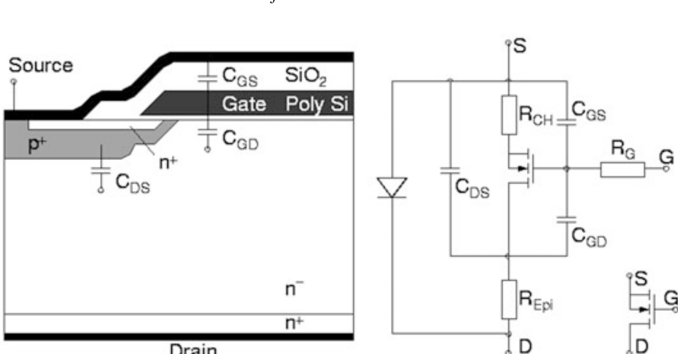


Fig. 9.17 MOSFET with parasitic capacitances, structure and electrical equivalent circuit according to [Mic03] © 2003 Springer

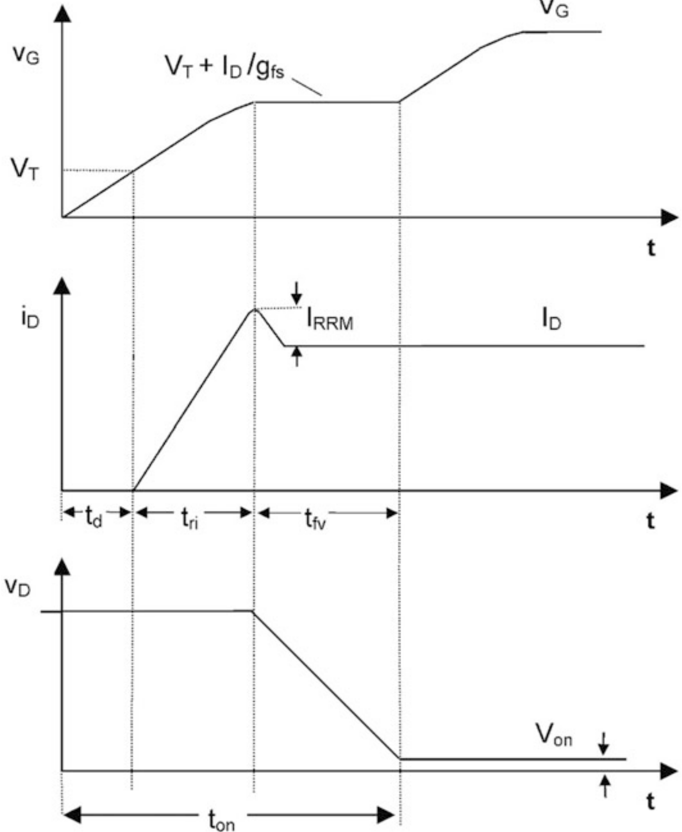


Fig. 9.18 Turn-on waveforms of the MOSFET with inductive load

In this phase, V_G remains at the value of the Miller plateau

$$V_G = V_T + I_D/g_{fs}$$

The voltage V_D falls to the value of the forward voltage

$$V_{on} = R_{on} \cdot I_D$$

The entire turn-on time t_{on} amounts to

$$t_{on} = t_d + t_{ri} + t_{fv}$$

Figure 9.19 shows the turn-off behavior for an inductive load. Characteristic quantities of the turn-off process are:

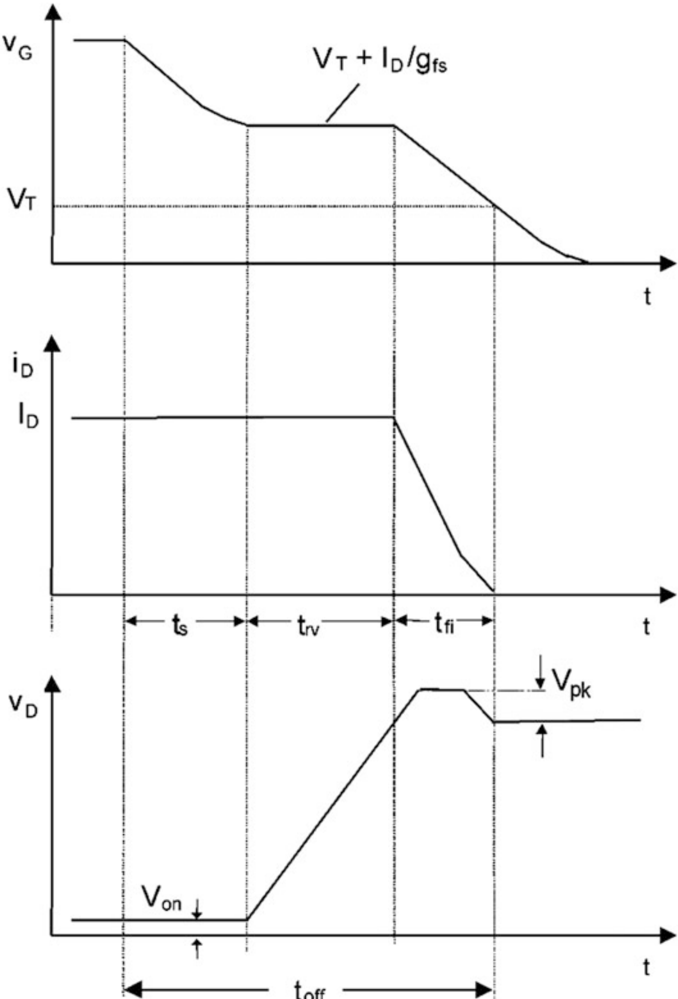


Fig. 9.19 Turn-off waveform of the MOSFET with inductive load

t_s: Storage time
In the driver the voltage signal is reset to zero or a negative value. However, the gate has to be discharged to the value at which the gate voltage corresponds to the value at which the on-state current I_D equals the saturation current, which means

$$V_G = V_T + I_D/g_{fs}$$

The capacitances C_{GS} and C_{GD}, which are in parallel to the channel, have to be discharged (see Fig. 9.15), For the storage time holds:

$$t_s \sim R_G \cdot (C_{GS} + C_{GD})$$

t_{rv}: *Voltage rise time*

The voltage increases to the value given by the circuit. The current remains constant at the initial value. The gate voltage persists at the Miller plateau. The Miller capacitance C_{GD} has to be discharged and, therefore

$$t_{rv} \sim R_G \cdot C_{GD}$$

t_{fi}: *Current fall time*

The gate capacitance $C_{GS} + C_{GD}$ is discharged and the current decreases. The current becomes zero (or, more exact, it attains the value of the off-state leakage current), when V_{GS} has fallen to V_T.

$$t_{fi} \sim R_G \cdot (C_{GS} + C_{GD})$$

In this phase a spike V_{pk} is added to the applied voltage. This spike consists of

- the inductive voltage, which is generated by the current slope *di/dt* at the parasitic inductance L_{par}, L_{par} is indicated in Fig. 5.18.
- the turn-on dynamic forward voltage spike V_{FRM} of the diode

Consequently

$$V_{pk} = \left| L_{par} \cdot \frac{di}{dt} \right| + V_{FRM}$$

The entire turn-off time is

$$t_{off} = t_s + t_{rv} + t_{fi}$$

The switching edges of the turn-on and turn-off can be controlled by means of the gate resistance under the conditions described. With a smaller R_G the switching time can be reduced and, thus, the switching losses in the device may be lowered as well.

From the switching times a frequency limit can be derived

$$f_{\max} = \frac{1}{t_{on} + t_{off}} \tag{9.32}$$

Taking the data sheet values of the above mentioned MOSFET IXYS IXFH 67 N10 and adding all switching times typically 220 – 340 ns is obtained, which corresponds to a frequency of 3 – 4 MHz. This is considerably lower than f_{co}. But the example is not unproblematic, as the switching times in the data sheets are mostly specified for an ohmic load, which is rarely the case in practice.

9.9 Switching Losses of the MOSFET

The maximum attainable switching frequency of a power MOSFET depends on the switching losses. The energy loss per pulse can be calculated, like for other devices, by integrating the product $v(t) \cdot i(t)$ during turn-on and turn-off. During turn-on it can be calculated with

$$E_{on} = \int_{t_{on}} v_D(t) \cdot i_D(t) dt \tag{9.33}$$

In practice, the energy loss per pulse is determined from oscillograms. Modern oscilloscopes are able to calculate the product of current and voltage and to integrate it over the selected time. An example for an IGBT is given in Fig. 5.21. For an estimation, Fig. 9.18 can be used from which emanates

$$E_{on} = \frac{1}{2} \cdot V_D \cdot (I_D + I_{RRM}) \cdot t_{ri} + \frac{1}{2} \cdot V_D \left(I_D + \frac{2}{3} I_{RRM} \right) \cdot t_{\mathrm{fv}} \tag{9.34}$$

assuming that the peak reverse current I_{RRM} caused by the diode decays linear during the time t_{fv}.

At turn-off, the energy loss is calculated with

$$E_{off} = \int_{t_{off}} v_D(t) \cdot i_D(t) dt \tag{9.35}$$

which can be estimated according to Fig. 9.19 with

$$E_{off} = \frac{1}{2} \cdot V_D \cdot I_D \cdot t_{rv} + \frac{1}{2} \cdot \left(V_D + V_{pk} \right) \cdot I_D \cdot t_{fi} \tag{9.36}$$

The total switching losses follow from

$$P_{on} + P_{off} = f \cdot \left(E_{on} + E_{off}\right) \tag{9.37}$$

Conduction losses and blocking losses add to the switching losses. For power MOSFETs the off-state leakage current is in the order of few µA, so that the blocking losses may be neglected. The conduction losses cannot be ignored. Defining the duty-cycle d as the ratio of the interval in which the MOSFET conducts versus the switching period, the conduction losses can be calculated according to

$$P_{cond} = d \cdot V_{on} \cdot I_D = d \cdot R_{on} \cdot I_D^2 \tag{9.38}$$

For the total losses one obtains

$$P_V = P_{cond} + P_{on} + P_{off} = d \cdot R_{on} \cdot I_D^2 + f \cdot \left(E_{on} + E_{off}\right) \tag{9.39}$$

These losses have to be conducted away as heat flux through the case of the device. The maximum allowable losses are determined by the cooling conditions, the acceptable temperature difference and the thermal resistance. Details are provided in Chap. 11.

For the MOSFET IXYS IXFH 67 N10 used as an example, it can be estimated from the data sheet values of the thermal resistance that switching frequencies up to 300 kHz can be realized. Clearly, the MOSFET, as it is a unipolar device, is the fastest Si power semiconductor switch available.

The potential switching frequency depends, on the one hand, on the thermal parameters and considerably on the other devices in the circuit as well. The entire circuit has to be optimized accordingly. From (9.34) and (9.36) arises that the switching losses depend on the switching times. By a reduction of the switching times due to smaller gate resistances R_G, switching losses can be lowered. On the other hand, the steepness of the slopes is limited in practice

- by motor windings which are not to be stressed with too high *dv*/*dt*
- even more by freewheeling diodes, which are required in inductive circuits. Inappropriate freewheeling diodes, in the presence of increased *di*/*dt*, lead to a snappy switching behavior, voltage spikes and oscillations.

9.10 Safe Operating Area of the MOSFET

Between source and drain, the structure of the MOSFET contains a parasitic bipolar npn-transistor in parallel to the MOS channel, as shown in Fig. 9.20. This parasitic npn-transistor could lead to many problems:

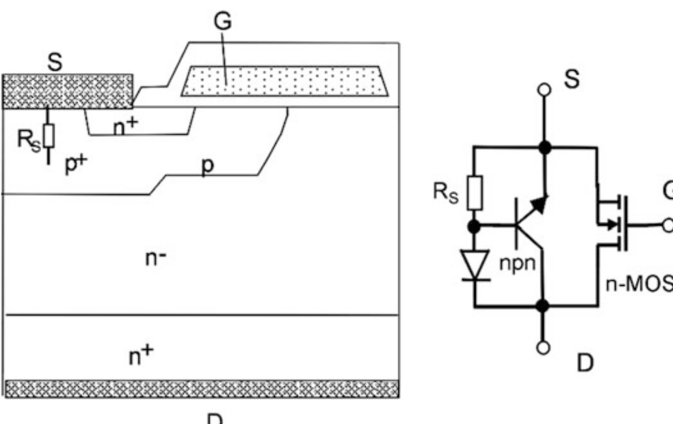


Fig. 9.20 MOSFET with its equivalent circuit, containing the parasitic npn-transistor and the parasitic diode

- the blocking voltage would be decreased by such an open base transistor
- when applying a voltage with a high *dv*/*dt*, a displacement current could be generated due to the charging of the depletion layer of the base-collector junction. This current triggers the transistor,
- and finally, the safe operating area (SOA) of a transistor is limited by the second breakdown effect.

Therefore, the base-emitter junction of the npn-transistor has to be shorted by a low resistance R_S. This resistance is chosen preferably small by increasing the doping in this region with an additional p^+ ion implantation (p^+ doping) and by choosing the length of the n^+ source region as small as possible, as the photolithographic process allows it.

In today's MOSFETs the parasitic transistor is effectively made inoperative. Thus, the safe operating area is no longer limited by the second breakdown. The safe operating area of today's MOSFETs is rectangular, as it is shown in Fig. 9.21. It is only limited by the blocking voltage and the occurring losses. The SOA curves for pulse times above 10 µs in Fig. 9.21 are limited due to the maximum power losses, at which the junction temperature remains below 150 °C.

9.11 The Inverse Diode of the MOSFET

Because of the contact between the p-well and the source metallization, a pin-type diode structure is built by the p-well, the n^- region and the n^+-substrate, as shown in Fig. 9.20. Consequently, for the application in a bridge topology in a voltage source converter circuit, a freewheeling diode is intrinsically present. The characteristics of

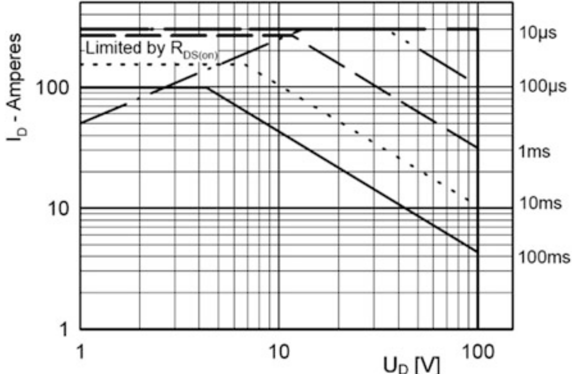


Fig. 9.21 Safe operating area (SOA) of a MOSFET, example IXYS IXFH 67 N10

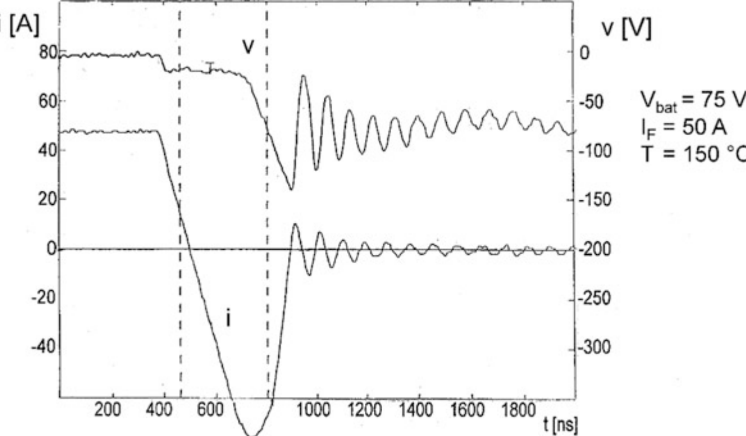


Fig. 9.22 Switching behavior of the inverse diode of a 200 V MOSFET with high-frequency LC-oscillations

this diode corresponds to the current-voltage characteristic of the MOSFET in the 3rd quadrant (see Fig. 9.7 for $V_G = 0$). However, the turn-off behavior of this intrinsic diode is relatively poor, compared to optimized pin-type or Schottky barrier diodes for the same blocking voltage. Figure 9.22 shows a snappy turn-off event of the inverse diode in a 200 V-MOSFET.

The MOSFET manufacturing technology usually leads to a high carrier lifetime. Therefore, a high stored charge and a high peak reverse current of the diode occurs in conventional MOSFETs. This is an impediment for many applications.

A carrier lifetime adjustment is applicable to reduce the stored charge. It has to be carried out in a separate production step. As a first approximation, the insertion of recombination centers in the n^--region does not affect the properties of the

MOSFET, because the MOSFET is a unipolar device. During the on-state of the MOSFET, carrier recombination cannot take place. Hence, the resistance R_{on} should remain unaffected. However, secondary effects have to be taken into account. Recombination centers, which are inserted for the reduction of the carrier lifetime, can retroact on the effective doping. Thus, the utilization of gold is ruled out, because gold, acting as acceptor, will compensate the base doping and increase the resistance R_{on}. This effect does not occur when platinum or electron irradiation is used. With electron irradiation it has to be considered that it affects the charge in the gate oxide. Electron irradiation reduces the threshold voltage V_T. By means of adequate annealing, the threshold voltage can partially be restored.

MOSFETs with platinum diffusion or electron irradiation, which are used for the reduction of the stored charge of the inverse diode, are known as "FREDFET" (Fast Recovery Diode Field Effect Transistor). Here, the stored charge of the inverse diode is reduced. The reverse recovery behavior is improved slightly, so that these diodes can be used in circuits with low parasitic inductance.

Primarily, the reverse recovery behavior is problematic. Though the area of the p-region is considerably smaller – a measure, which is also used with an MPS diode to achieve soft recovery behavior – fundamental requirements on the MOSFET are contradictory to what is necessary to achieve soft recovery turn-off behavior:

- To keep R_{on} as low as possible, the base of the MOSFET has to be designed as thin as possible.
- To obtain an effective short-circuit R_S, the p^+-doping is chosen as high as possible.

Both measures result in a snappy switching behavior of the diode, and they limit the possibility for both the MOSFET as well as the diode to be optimized.

In many hard-switching applications the inverse diodes are unusable, and they are often called parasitic diodes. By inserting a Schottky diode in series to the MOSFET and in reverse direction to the inverse diode, they can be suspended, and an optimized soft recovery diode can be connected in parallel. However, further losses occur due to the threshold voltage of the additional junction.

When a part of the current in the diode freewheeling mode flows via the n^+-source layer, the performance of the diode can be improved., as shown in [Zen00]. This effect is more pronounced in a Trench MOSFET [Dol04]. The effect is shown in Fig. 9.23. As already mentioned in the discussion of the threshold voltage at the beginning of the MOSFET chapter, a potential difference between gate and semiconductor exists due to the different positions of the Fermi level in the heavily n^+-doped poly silicon gate and in the p-type semiconductor, which acts as a small positive gate voltage. Further, if the current flow is in reverse direction, a potential drop builds up and the point D^* in Fig. 9.23 becomes negatively biased against the source. If $V_G = 0$ (=source potential), then this leads to a positive potential of V_G compared to D^*. This is of same effect as a positive voltage between gate and p-body, and an inversion channel can be formed, respectively the threshold voltage is dynamically reduced, as expressed by [Dol04]. Although the outside gate potential is set to zero, an electron-conducting channel becomes present. In devices

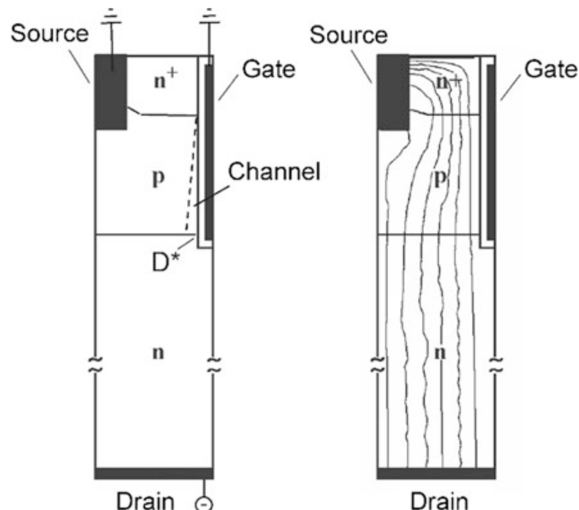


Fig. 9.23 Inverse diode in a trench MOSFET: potential and channel formation (left), simulated forward conduction current distribution (right)

with high channel density, as in Trench MOSFETs, it can become the dominant current component. Under these bias conditions, more than 90% of the total current is confined to the channel region [Dol04]. An example is given in Fig. 9.23.

The effect is somewhat similar to Fig. 9.8, where a positive potential at D* to source leaded to the channel pinch-off, now the potential at D* is of opposite sign.

The occurrence of an electron current is of positive effect to the reverse diode characteristic. In its on-state, significant conduction takes place at a voltage below the conventional diode turn-on voltage at ca. 0.7 V, reducing conduction losses. Additionally, since this current is primarily due to majority carriers, this effect will positively impact the diode recovery behavior, since less stored charge will need to be removed during the reverse recovery process. Only a fraction of the total current is conducted by the pure diode part in the structure, i.e. the body-epilayer junction, as shown in Fig. 9.23.

Therefore, only the fraction of the total current which is conducted in parallel by the pure diode part causes an excess carrier concentration in the base-region of the pin-diode and contributes to the reverse-recovery charge. The ratio between the channel-conducted electron current part and the hole current injected by the body of the inverse diode depends on the device structure (thickness of gate-oxide, body doping). It also depends on the current density. The larger the current density, the larger becomes the share of injected holes and the larger will be the stored charge. With said measure, diodes of some modern Trench MOSFETs are significantly improved. Figure 9.24 shows a strong reduction of the stored charge of a new generation of trench MOSFETs (Device B) compared to an older design (Device A).

In applications with low blocking voltages and high currents (e.g. blocking voltage < 100 V and output current > 10 A), like switched power supplies, the technique of synchronous rectification is very common (see chapter MOS Controlled Diodes). pin-diodes or Schottky-barrier diodes used as free-wheeling

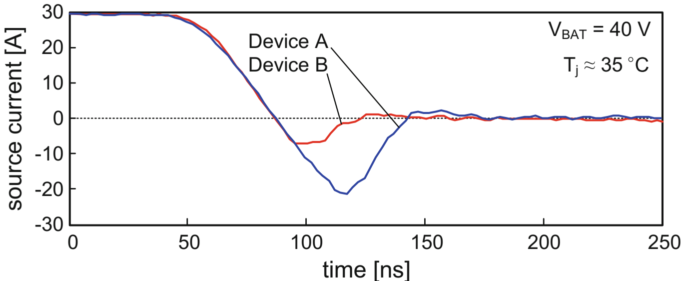


Fig. 9.24 Turn-on of MOSFETs in a half-bridge configuration with reverse recovery of the respectively inverse diodes. Both devices rated for 75 V and for $I_D > 100$ A. *Device A* IRF3808S planar technology, *Device B* IRFS3207 trench technology. $di/dt = 800$ A/µs

devices are replaced by MOSFETs, which are utilized for inverse conduction and allow a reduction of conducting losses due to the significantly lower voltage drop across the device.

When the channel of the synchronous rectifier is open, its intrinsic diode is bypassed. In the I-V-characteristic in Fig. 9.7 this mode is given by the dotted line in the 3rd quadrant for $V_G > V_T$. In this case, the voltage drop across the device is lower compared to the case of a closed channel, particularly when low currents are applied.

Figure 9.25 shows the circuit of a buck converter, which is a typical application of low-voltage power MOSFET. To avoid shoot-through currents, which would occur if the low-side MOSFET gate still has an on-signal while the high side MOSFET turns on, the gate control scheme for the switches must contain so-called interlock or dead times during the current commutation, wherein both channels are closed. In the example shown in Fig. 9.25, the low-side MOSFET must be turned-off before the high-side MOSFET is turned on. The same holds for the high-side switch. Thus, the intrinsic diode of the synchronous rectifier is conducting during these necessary dead times.

A control scheme for this type of operation is given in Fig. 9.26. During the diode conducting time, the gate of the low side MOSFET has the signal "on", and

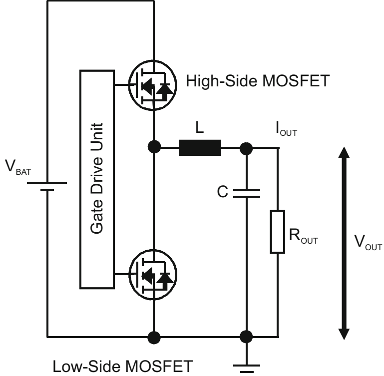


Fig. 9.25 Basic schematic of a buck converter with MOSFETs and their intrinsic diodes

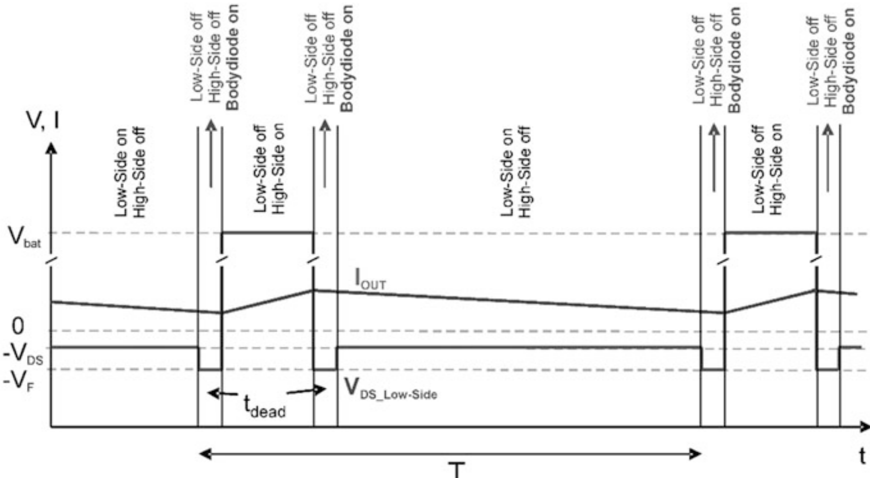


Fig. 9.26 Timing scheme for the buck converter as shown in Fig. 9.25

the voltage drop across the device is $-V_{DS}$, a value below the junction voltage of the intrinsic diode. During the dead times, the voltage drop is $-V_F$, a value which is in its absolute value above the junction voltage of the intrinsic diode.

Besides the effort of precise gate control, the turn-off behavior of the inverse diode becomes again one of the main drawbacks in this kind of application. In practice, the dead times cannot be reduced below a certain value, thus, the losses related to body diode conduction and body diode turn-off are responsible for an increasingly significant part of the total losses with increased switching frequencies.

The losses caused by the inverse diode of the MOSFET make up the largest share of the total losses in case of devices with low breakdown voltages. Inserting a Schottky-barrier diode in parallel to the synchronous rectifier is a way to reduce or to prevent the conduction of its body diode during the dead times. While the effect of this measure is limited by parasitic inductances of the circuit when discrete devices are used [Pol07], it shows significant improvement of the switching behavior if the Schottky barrier diode is integrated monolithically into the MOSFET die [She90, Cal04, Bel05].

9.12 SiC Field Effect Devices

9.12.1 SiC JFETs

SiC unipolar devices allow a very thin drift zone and a higher doping of the base, they attain a much lower resistance R_{epi} compared to Si, see Fig. 6.9. Therefore, research and development in SiC MOSFETs is done since several years. However,

mostly due to problems manufacturing a stable gate oxide, the first stable field effect device from SiC was the Junction Field Effect Transistor (JFET). The structure of a SiC JFET half cell [Mit99], which is in this configuration fabricated by Infineon, is shown in Fig. 9.25. With $V_G = 0$, there is a path from source to drain, and the structure is conducting. For the blocking mode, a negative voltage must be applied at the gate, this builds a space charge between gate and source, and the channel is pinched-off (Fig. 9.27).

The I-V characteristic of a JFET is shown in Fig. 9.28. For $V_G = 0$ V the JFET is in the forward conducting mode. With $V_G < 0$, the channel is narrowing, and for $V_G = -20$ V this sample is in the forward blocking mode. The cut-off voltage V_{CO} of this specific sample was at $V_G = -17.5$ V. To ensure reverse blocking of the JFET, a gate voltage below the cut-off voltage must be applied, however the negative gate voltage is limited by the breakdown voltage $V_{BD}(GS)$ of the gate-source diode, which occurs typically at -25 V. Due to tolerances in the manufacturing process, there is a spread of the values of V_{CO} as well as $V_{BD}(GS)$, this spread should be small. To ensure blocking capability, the JFET needs a negative voltage between V_{CO} and $V_{BD}(GS)$.

A "normally on" device is unwanted for voltage source converter applications. However, this can be solved by a MOSFET in series to the JFET in the cascode configuration, as shown in Fig. 9.29.

If a voltage $V_G > V_T$ is applied to the MOSFET, it is in the conduction mode. The normally-on JFET in series is conducting, and the cascode configuration is in the conduction mode. If the MOSFET is turned off, a voltage up to its blocking capability is built up. This voltage, which has a negative polarity to the source of the JFET, is applied to the JFET gate. If it is lower than the pinch-off voltage of the JFET, the JFET is turned off. The JFET is in the blocking mode. The behavior of the cascode configuration is similar as that of a MOSFET and it can be operated widely similarly as known for MOSFET.

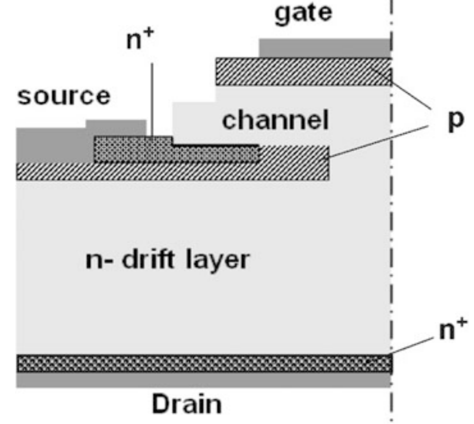


Fig. 9.27 SiC JFET half cell

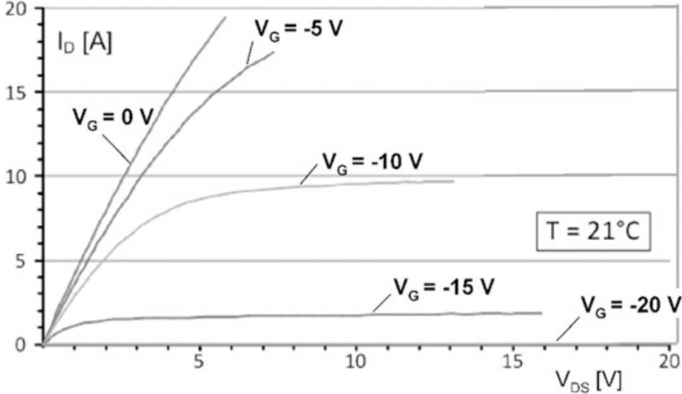


Fig. 9.28 Forward I-V characteristic of a JFET

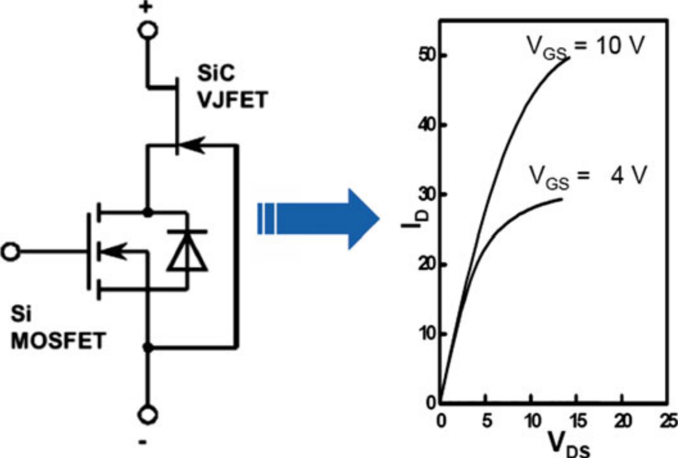


Fig. 9.29 JFET and low-voltage MOSFET in the cascade mode. Configuration (left), forward I-V characteristic of the cascode with V_{GS} (=V_G) at the Si-MOSFET gate (right)

In the conduction mode, the R_{on} of the MOSFET is in series to the R_{on} of the JFET. However, MOSFETs in the 30 V range can meanwhile be designed with extremely low on-resistances. They are available with R_{on} down to 0.1 mΩcm^2. Nevertheless, packaging an additional device requires additional effort.

Critical for the configuration in Fig. 9.29 is the control of the switching slopes at turn-on and turn-off. Especially, oscillations may occur during the cascode switching processes. They are caused by the capacitive elements of the cascode arrangements, which form resonant circuits with unavoidable inductive elements in the complete circuit. In particular, the turn-off process was found to be critical

[Sie12]. As an alternative, a so-called "direct-driven JFET" was proposed. Here separate drivers for the MOSFET and the JFET are present. The additional driver can be integrated into a single driver IC.

9.12.2 SiC MOSFETs

The SiC MOSFET is a very attractive device, and much progress in research and development has been achieved over the past decade. From several manufacturers, commercial released devices are available.

SiC MOSFETs are of n-channel type and have basically a similar structure as the vertical DMOS transistor (see Fig. 9.4). However, a problem of the SiC MOSFET is the channel conductivity, since the effective mobility of electrons in the channel is low. Whereas channel mobilities in the range of 5–10 cm^2/Vs were typical for first SiC MOSFETs, channel mobilities up to 13 $cm^2V^{-1}s^{-1}$ have also been reported [Ryu06]. It must be considered that also in Si MOSFETs the channel mobility is lower than the bulk mobility due to the influence of the surface (in the range of 500 $cm^2V^{-1}s^{-1}$, see Eq. 9.19), but the effect is much worse in SiC. This is due to a high density of electron traps in the SiC MOS interface, which results in the capture of channel electrons and in increased Coulomb scattering due to these captured electrons [Ima04]. The gate oxide is typically grown in an atmosphere containing NO or N_2O to reduce MOS interface state density.

According to (9.22), the total resistance R_{on} of the MOSFET has different contributions, where R_{ch} and the drift layer resistance R_{epi} are main contributors. For R_{ch}, Eq. (9.3) can be used with the parameters for SiC, and for R_{epi} Eq. (9.23) holds. For the dependency of R_{epi} on the blocking voltage, Eq. (6.16) derived for SiC Schottky diodes is suited.

R_{ch} was found to decrease with temperature for the first generations of SiC MOSFETs. The channel mobility was found to increase with temperature due to decreasing Coulomb scattering at surface states, however not to exceed 16 $cm^2V^{-1}s^{-1}$ at the highest temperature. While the channel mobility increases with temperature, the mobility in the drift region decreases keeping the on-state resistance almost independent of temperature [Rum09].

A measurement of R_{on} over time for two SiC MOSFETs available on the market is displayed in Fig. 9.30. It is seen that for more recent SiC MOSFETs R_{on} increases with temperature, however, from 25 to 125 °C only by 25%. For the Si-MOSFET, R_{on} usually doubles between 25 and 125 °C. This shows that there is a high contribution of R_{ch} to the total mobility.

In a more detailed treatment, the channel mobility μ_{ch} consists of different contributions and their influence can be combined with "Matthiessen's Rule" (developed from work by Augustus Matthiessen in 1864) which for the given situation can be written as [Uhn15]:

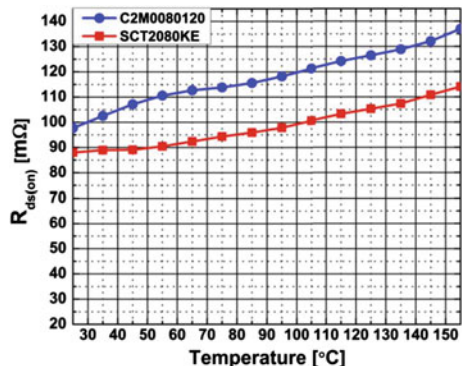


Fig. 9.30 On-state resistances versus temperature increase of the SiC MOSFETs SCT2080KE (manufactured by ROHM) and C2M0080120 (manufactured by Cree). Figure from [Muh16]

$$\frac{1}{\mu_{ch}} = \frac{1}{\mu_{bulk}} + \frac{1}{\mu_C} + \frac{1}{\mu_{sr}} + \frac{1}{\mu_{sp}} \tag{9.40}$$

Wherein μ_{bulk} is the bulk mobility, as expressed before as μ_n for the n-channel MOSFET, μ_C is the mobility contribution due to scattering at surface states, μ_{sr} due to surface roughness and μ_{sp} due to surface phonons.

The term with the lowest mobility is most strong in the result of Eq. (9.40). μ_C strongly increases with temperature, since the amount of charged traps decreases with temperature, and additional due to the increased Fermi-Potential more electrons are present shielding the scattering centers. Further, a strong dependency on the doping N_A at the position of the channel was found. For SiC MOSFETs with lower N_A, the contribution of μ_C is found to be smaller. For a lowly p-doping, channel mobilities up to 70 cm^{2}V^{-1}s^{-1} are reported, and a decrease of μ_{Ch} with temperature was found [Uhn15].

In modern SiC-MOSFETs, further the channels are designed as very short, and additional effort is done to improve the channel conductivity e.g. by an additional thin n-doped layer.

Prototypes of SiC MOSFETs reached remarkable low R_{on}, e.g. 5 mΩcm^2 [Miu06] and 1.4 mΩcm^2 [Nak11] for a 1200-V type. However, these values are still far above the theoretically possible low limits for SiC (see Fig. 6.9). Intensive work is on-going. The structures of SiC trench-MOSFETs are shown in Fig. 9.28.

In a trench-MOSFET (Fig. 9.31 left), the resistance R_a is eliminated, however, at applied high drain-source voltage the highest electric field occurs at the gate oxide at the trench bottom. In the double-trench MOSFET structure (Fig. 9.31 right) the position of highest field moves to the edge of the source trench whose position is deeper. Hence, the gate trench is released from high electric fields to some extend [Nak11], preventing overload or destruction of the oxide layer at the gate trench.

Important for the reliability of SiC MOSFETs is the gate oxide quality. The commonly used material for gate oxide in Si and SiC devices is SiO_2. In some first

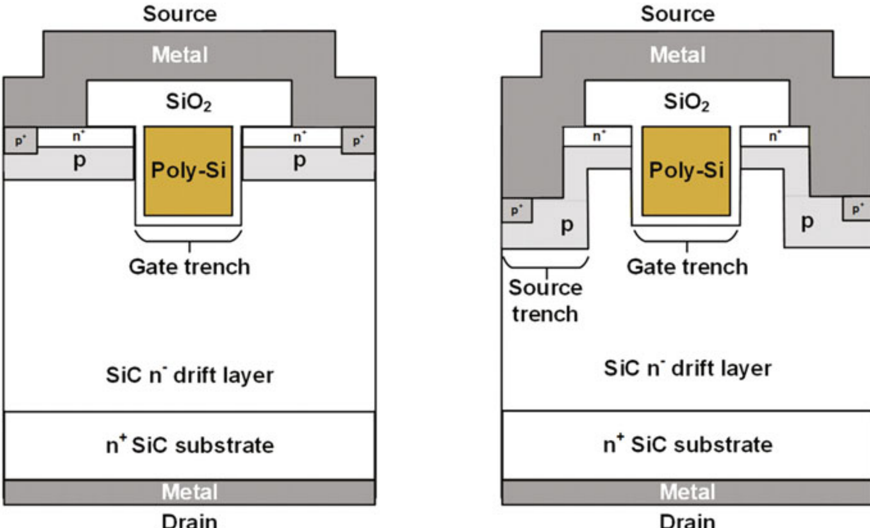


Fig. 9.31 Trench MOSFET and double-trench MOSFET structures. Figure according to [Nak11]

generations of SiC MOSFETs, the gate oxide is thinner in comparison to Si power MOSFETs and IGBTs due to the impact of oxide thickness on the channel resistance, see Eqs. (9.1) and (9.3). At a high electrical field, the Fowler-Nordheim tunneling which describes emission of electrons at surfaces in presence of high electric field [Fow28] contributes to dielectric breakdown [Scd96]. The Fowler-Nordheim tunneling current is higher in SiC than in Si [Gur08]. The tunneling current is exponentially dependent on the electric field in the dielectric and barrier height to carriers. This barrier height is primarily determined by band offsets between SiC and the dielectric. Since band offsets for SiC to SiO_2 are smaller than those with respect to Si, a lower reliability is expected for SiC MOS-based devices as compared to MOS devices for the same electric field oxide from Si [Sin04]. Further, in oxide layers on SiC exists a high density of interface and near-interface traps [Gur08].

It was shown that the effective lifetime of SiC MOSFETs will be determined not by oxide lifetime (so-called intrinsic failure) but by extrinsic failures which are caused by imperfections of the manufacturing process. Gate stress tests reported in [Bei16, Bei17] showed several extrinsic failures. Major differences between the investigated manufacturers were found; however, it is shown that SiC MOSFETs might reach IGBT-like gate oxide reliability and that they are on the way to become mature and reliable devices. It is possible to obtain SiC MOSFETs down to the same low ppm rate as Si MOSFETs or IGBTs by applying smart screening measures [Lut18]. The enabler for efficient gate oxide screening is a much thicker bulk oxide than what is typically needed to fulfill intrinsic life time targets.

A further challenge for the SiC MOSFET is the control of the threshold voltage V_T [Aga06]. In particular, V_T is found to shift with accumulated stress for the device. Possibly, a thick enough gate oxide and a suited screening process will also minimize this effect.

9.12.3 The SiC MOSFET Body Diode

Changing the polarity of V_{DS}, the inverse body diode of the SiC MOSFET is in the forward conduction state. However its threshold voltage is for devices of several manufacturers not at the expected threshold voltage of a pn-junction in SiC at about 2.7 V at room temperature, but at much lower values. Indeed, when $V_G = 0$ (=source potential) and a voltage drops across the body pn-junction, a positive potential of V_G appears at the p-body to n^{-}-SiC interface. This effect is the same as if a positive potential of gate to p-body is applied such that and an inversion channel can be formed. This effect is already known from Si-MOSFETs, see Sect 9.10 and Fig. 9.2. However, due to the higher built-in voltage of a pn-junction in SiC, it is much more apparent in SiC. An example of the forward characteristics of the inverse diode for different gate voltages is shown in Fig. 9.32. A negative voltage at

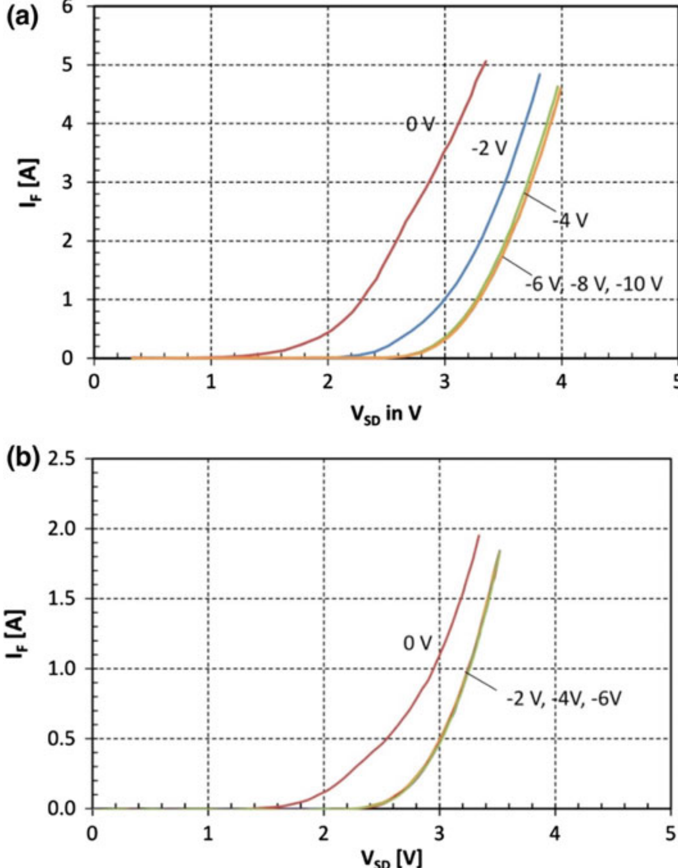


Fig. 9.32 Forward characteristics of the inverse diode of a SiC MOSFET for different gate voltages. **a**) 25 °C **b**) 150 °C

the gate of −6 V (RT) or −4 V (150 °C) must be applied before the channel is closed and the threshold voltage of the pn-junction becomes visible.

In first applications, a SiC Schottky diode was connected antiparallel to the MOSFET. More and more, the inverse MOSFET diode is used in the synchronous rectifier mode, compare chapter on diodes. In this mode, the inverse current is conducted by a significant part in unipolar mode. Short before turn-off, however, the channel must be closed to avoid phase-leg short circuit. Figure 9.32 shows that also in this mode the main part of the current is flowing across the channel, this is an advantage for the switching behavior, as explained in Sect. 9.11.

9.13 GaN Lateral Power Transistors

Vertical GaN Power devices are not available yet. All GaN devices used as power devices are lateral devices. Their function is based on the 2-dimensional electron gas, which connects source and drain and is switched by the gate.

For the lateral resistance with the length L and width W and a charge density per area G_{2DEG}, one can calculate (compare Eq. 9.3)

$$R_{on} = \frac{L}{W \cdot \mu_{2DEG} \cdot q \cdot G_{2DEG}} = \frac{L}{W \cdot \mu_{2DEG} \cdot Q_{2DEG}} \tag{9.41}$$

G_{2DEG} for AlGaN/GaN heterostructure can be above 10^{13} cm^{-2}, μ_{2DEG} can be up to 2000 cm^2/Vs. In a lateral device, the lateral electric field shape is rectangular in ideal case. Then holds for the breakdown voltage [Ued17]:

$$V_{BD} = \mathrm{E}_c \cdot L_{GD} \tag{9.42}$$

E_C is the critical field strength and L_{GD} the distance between gate and drain, see Fig. 9.33. For the length L in Eq. 9.41, the distance between gate and source L_{GS}

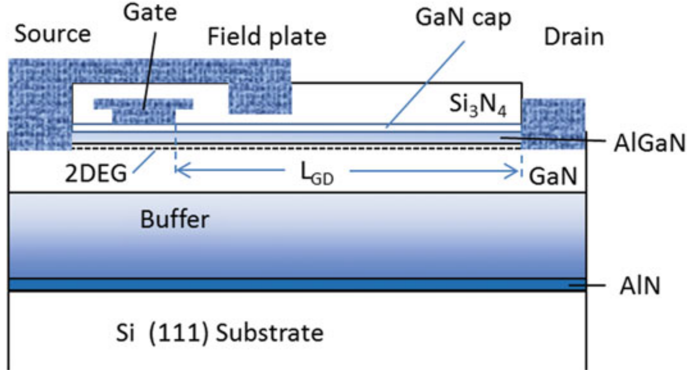


Fig. 9.33 GaN High-electron-mobility transistor (HEMT)

has to be added, $L = L_{GD} + L_{GS}$. Then from (9.41 with (9.42) it can be derived for R_{on} [Ued17]

$$R_{on} = \frac{L_{GS}}{W \cdot \mu_{2DEG} \cdot q \cdot G_{2DEG}} + \frac{V_{BD}}{W \cdot \mathsf{E}_c \cdot \mu_{2DEG} \cdot q \cdot G_{2DEG}} \tag{9.43}$$

For lateral devices the performance is often specified by resistance times gate width $R_{on} \cdot W$. For devices with blocking capability of 650 V, values as low as 27 Ωmm can be seen in recent publications [Miy15]. The resistance per area depends on details of the realized structure as well. High channel width W per area is desired. The structure is arranged in stripes, which shall be narrow. Due to the high critical electric field strength in GaN, the lateral extension L_{GD} is quite small, significant below 10 μm for a 650 V device. In contrast to lateral devices made in Silicon, GaN devices require a significantly smaller surface area for the space charge. The length L_{GS} is nowadays below 1 μm, e.g. 0.8 μm. Therefore, with GaN even in a lateral device a high current density in respect to the die area can be achieved. Comparing the resistance for the device area, today values of 0.54 Ωmm^2 [Miy15] and 0.66 Ωmm^2 [Hil15] have been demonstrated. These values are well beyond 1–2 Ωmm^2 which are achieved for nowadays with 650 V superjunction structures, as illustrated in Fig. 9.11.

Especially the gate capacities are much lower for GaN devices compared to Si. This leads to low switching losses and allows high switching frequencies. An often used criterion is the product $R_{on} \cdot Q_g$ with

$$Q_g = C_{iss} \cdot (V_{Gon} - V_{Goff}) = (C_{GS} + C_{GD}) \cdot (V_{Gon} - V_{Goff}) \tag{9.44}$$

Since the gate capacities are proportional to W, the product $R_{on} \cdot Q_g$ resp. $R_{on} \cdot C_{iss}$ becomes independent of W. With GaN, a $R_{on} \cdot Q_g$ is achieved about one decade smaller compared to the best Si MOSFETs in the 600 V voltage range.

Two main structures are developed today. The first, the high-electron-mobility transistor (HEMT), is initially a device for microwave application and was presented several decades ago with the $GaAs/Ga_{1-x}Al_xAs$ heterojunction [Mim80]. After the methods for deposition of GaN on sapphire substrates by MOCVD were developed, the production of AlGaN/GaN-based HEMTs was possible. The high polarization in the AlGaN/GaN heterostructure leads to a high charge density in the 2-dimensional electron gas (2DEG), see Sect. 4.11. Nowadays, HEMTs for power applications are fabricated on Si or SiC substrates. The basic structure of a HEMT as power device on a Si substrate is shown on example in Fig. 9.33.

It was shown by [Sai03] that 600 V blocking voltage could be achieved, and a current up to 850 A/cm^2 could be switched. In the blocking mode, the electric field drops laterally across the length L_{GD}, see Fig. 9.33. In [Sai03] L_{GD} amounts only 10 μm, in recent structures it can be even lower. The field plate on top in Fig. 9.33 expands the electric field from source in direction drain, the function is similar to a field plate in Fig. 4.25. Also multi-step field plates are in use [DeS16].

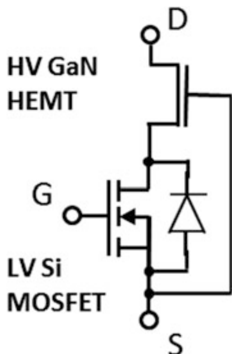


Fig. 9.34 GaN HEMT and Si MOSFET in cascode configuration

Meanwhile 600 V HEMTs are available on the market [Hon15]. To interrupt the conductivity of the 2-dimensional electron gas, a negative voltage at the gate must be applied in case of a depletion-mode HEMT. The device shows normally-on characteristics. To achieve normally-off characteristics, as preferred in typical power electronics voltage source converters, the HEMT device is combined with a low-voltage Si MOSFET in cascode configuration as shown in Fig. 9.34. The cascode is similar to the SiC JFET in cascode configuration shown in Fig. 9.27 and acts in the same way as explained in Sect. 9.12.1.

With this configuration, when a positive gate voltage is applied to the MOSFET, at the GaN HEMT the gate voltage drop equals the voltage drop between drain and source of the low-voltage Si MOSFET, which must be small to avoid losses. Hence, a relatively low positive gate voltage is applied at the gate of the HEMT.

For a long time, GaN devices suffered from the fact that they lacked current withstand capability at avalanche breakdown. In 2015, HEMT devices with avalanche capability have been reported [Liu15].

As a second main type, a normally-off 600 V Gate Injection Transistor (GIT) is available [Mor14, Ish15]. The basic structure is displayed in Fig. 9.35.

The GIT was first presented by [Uem07]. It features a p-AlGaN gate formed over the undoped AlGaN/GaN heterostructure, as shown in Fig. 9.35. The p-AlGaN lifts up the potential at the channel which enables normally-off operation. At the gate voltage of 0 V, the channel under the gate is fully depleted and no drain current is flowing. At the gate voltages between threshold voltage V_T and the forward built-in voltage V_{bi} of the pn junction, the 2D electron gas is recovered and the GIT is operated as a field effect transistor. Further increase of the gate voltage exceeding V_{bi} results in the hole injection from the p-AlGaN to the channel.

The injected holes accumulate the equal number of electrons that flow from the source to maintain charge neutrality at the channel. The accumulated electrons are moved by the drain bias with high mobility, while the injected holes stay around the gate, because the hole mobility is at least two orders of magnitude lower than that of the electrons. This kind of local conductivity modulation results in an increase of the drain current [Uem07]. In the $g_{fs} - V_G$ characteristics this occurs as a second peak, as shown in Fig. 9.36. However, since the holes do not contribute to current

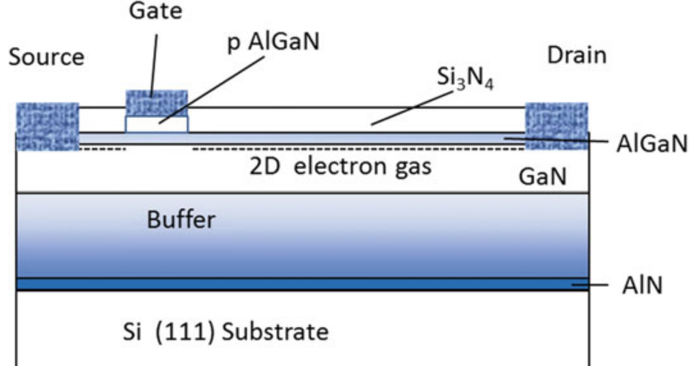


Fig. 9.35 Gate injection transistor (GIT) at $V_G = 0$ V

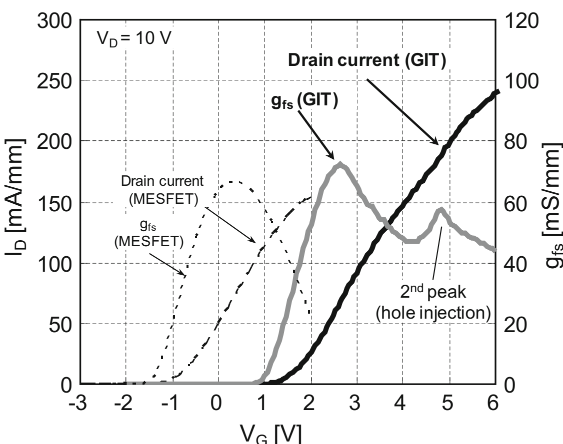


Fig. 9.36 $I_D - V_G$ and $g_m - V_G$ characteristics of **a**) a fabricated GIT and **b**) a MESFET (metal semiconductor field effect transistor). The GIT shows a peculiar transconductance characteristics with two peaks. Figure according to [Uem07]

transport, the GIT is not treated as bipolar device. The positive gate is usually limited to 5 V in order to keep the injected gate current below 10 μA/mm [Hil15].

The gate threshold voltage of a GIT is typically in the range of only 1–1.2 V. In applications with high electromagnetic noise, this low V_T could be a problem concerning unwanted turn-on, leading to short circuits in converter phase-legs. By introducing a negative gate voltage at the gate driver, GITs can be operated in converter bridges without shoot-through currents [Mor14].

Further concepts for normally-off (enhancement mode) GaN HEMTs are displayed in Fig. 9.37. Figure 9.37a shows the "recessed gate" structure, in which the thickness of the AlGaN layer is reduced. The threshold voltage becomes positive, is however typically below 1 V. In Fig. 9.37b an additional insulator layer is implemented, leading to a reduced gate leakage current. Different insulators are reported, oxides such as Al_2O_3 or HfO_2 and are possible [Hil15]. Figure 9.37c displays the GaN Metal Insulator Semiconductor Field Effect Transistor (MISFET).

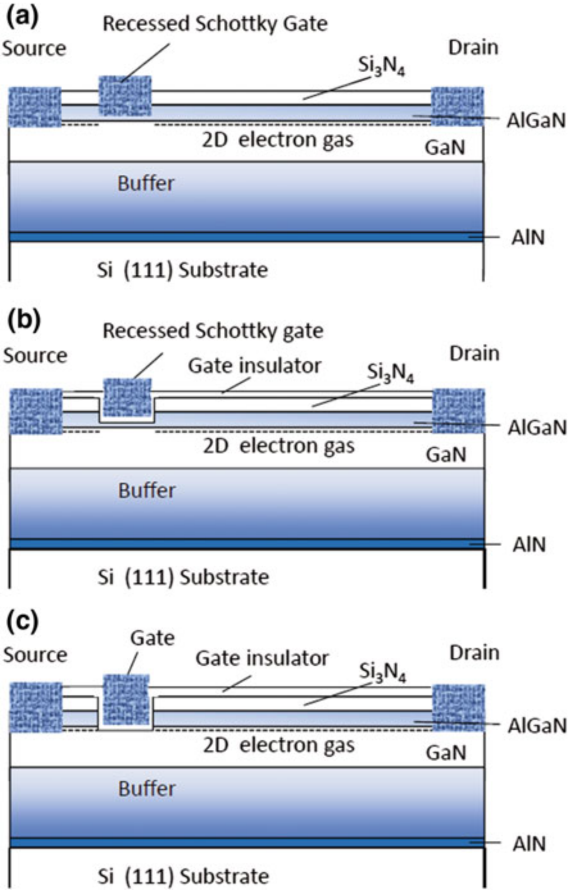


Fig. 9.37 Structures for enhancement mode GaN devices. **a** Recessed Schottky gate **b** recessed Schottky gate with additional gate insulator **c** GaN MISEFT. Figure inspired by [Wür15]

There is a full gate recess through AlGaN barrier. The insulator can consist of SiO_2 [Wür15] or a second thin layer of Silicon Nitride, in [Hua17] an additional interface protection layer of oxidized GaN is applied. At a positive voltage at the gate, an electron accumulation is created in the low conductive GaN layer below the gate, leading to a lateral conductivity.

A further method to achieve enhancement mode is the implementation of fluorine into the AlGaN layer below the gate, see Fig. 9.38. Since the group-VII element fluorine (symbol F) has the strongest electronegativity among all the chemical elements, a single F atom at the interstitial site tends to capture a free electron and becomes a negative fixed charge. These negative fixed charges deplete the two-dimensional electron gas (2DEG) in the channel at $V_G = 0$ V [Che17]. To restore the 2DEG a positive gate voltage is required. A threshold voltage up to +3 V is possible [Wür15]. The F^- incorporation can be combined with measures shown in Fig. 9.37, for example a second SiN_x thin film is applied as gate insulator

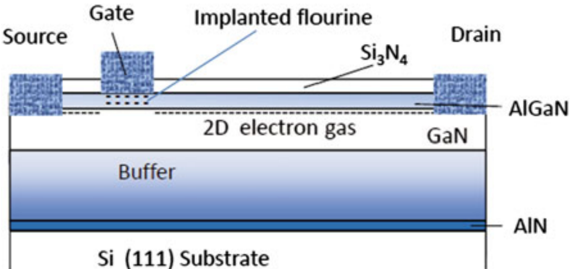


Fig. 9.38 GAN HEMT with negatively charged fluorine ions below the gate

to reduce the gate leakage current. With this combination, a threshold voltage of +3.6 V is achieved in [Che17].

The normally-off type is favored for power electronic applications. There are different approaches, a variety of different concepts, and no clear commitment to a specific technology. Measures shown in Figs. 9.35, 9.37 and 9.38 can also be combined. High threshold voltage levels (> +3 V) combined with low gate leakage current is desirable [Wür15]. Intensive work in research and development is done.

A problem of all lateral GaN devices is the so-called **current collapse**, this is the temporarily decrease of the 2DEG conductivity after applying blocking voltage. It leads to a decreased current capability I_{Dsat}. Power devices are usually not operated in the linear mode, however the effect of decreased 2DEG conductivity results also in the increase of the on-state resistance R_{on}. If desaturation occurs, the device might no longer be capable to carry the applied current, therefore the denomination as current collapse. More recently, the denomination "**dynamic R_{on}**" is often used for this phenomenon. The effect is shown in Fig. 9.39.

The dynamic R_{on} effect is reversible. It depends strongly on the value of the blocking voltage which is applied before turn-on, and on the time interval of negative voltage. For high switching frequency, the blocking time intervals are short and the effect is less exposed.

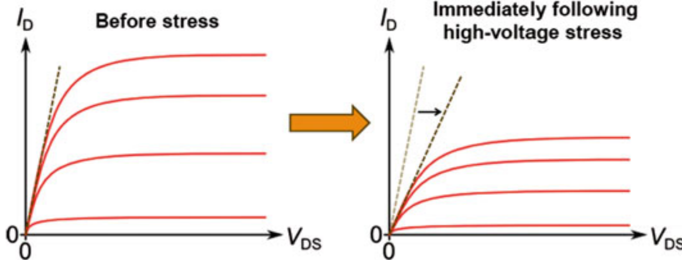


Fig. 9.39 Exemplifying drawing of the current collapse respectively dynamic R_{on} for a GaN device. Figure from S. Sque, NXP Semiconductors [Squ13]

Regarding its explanation, one has to consider that the Si-Substrate (Figs. 9.33, 9.35, 9.37, 9.38) is usually at the same potential as the source electrode. So the applied voltage lies in the same way across the vertical layers. The vertical blocking capability is specified in GaN-on-Si epitaxial wafers to 1000 V [EPI16]. A part of the leakage current will flow between substrate and drain.

There are two explanations:

- At off-state blocking, electrons from the gate are injected into trap states next to the gate. At on-state after stress, trapped electrons act like a negatively biased gate. The 2DEG is partially depleted leading to increased R_{on}. Time-dependent negative charges de-trap and the 2DEG current capability is restored.
- In blocking state, the Si-Substrate is usually at the same potential as the source electrode. Thus, the applied voltage also stands across the vertical layers. A part of the leakage current will flow between substrate and drain. Electrons are trapped in the bulk forming negative charge states. Trapped electrons partially deplete the 2DEG above. After the electrons de-trap, the 2DEG current capability is restored.

Lots of work is done for improvement. It seems that the main effect is the creation of negative charges by the leakage current from the substrate: An effective measure for the GIT is reported as the so-called Hybrid-Drain-embedded GIT, the HD-GIT. Its features an additionally p-GaN region formed in the vicinity of the drain, see Fig. 9.40 [Kan15]. This p-GaN region is electrically connected to the drain electrode by the interconnection metal layers. Injected holes from the p-GaN at the off-state effectively release the trapped electrons so that the current collapse is fully eliminated. In order to avoid the depletion of the 2DEG under the p-GaN region, a thicker i-AlGaN layer is employed, while at the gate a recessed gate structure is applied, compare Fig. 9.37a.

GaN device technology is still at a very early stage of development. The devices compete with Si superjunction devices in the range of 650 V and above. It is

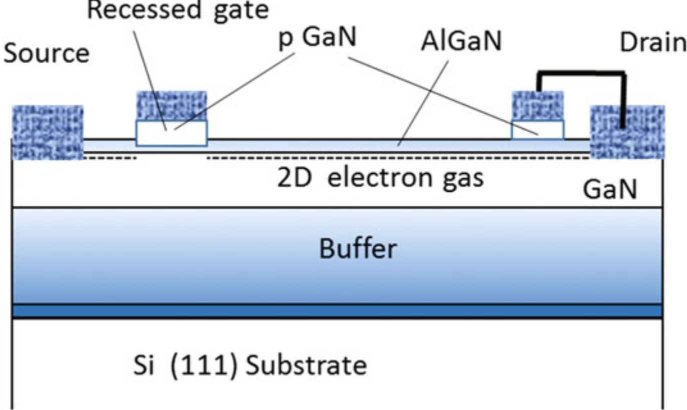


Fig. 9.40 Hybrid-Drain-embedded GIT (HD-GIT). A p-GaN at the drain and a recessed-gate structure are introduced in the HD-GIT. Figure following [Kan15]

expected that the main application field for GaN is the voltage region below 1000 V. The parasitic capacities C_{GS} and C_{GD}, summarized as C_{iss}, are much smaller compared to the superjunction devices [Hil15], enabling higher switching slopes and higher switching frequencies (above 1 MHz). Therefore these devices can lead to very compact and efficient power electronic converters. However, the reliability for industrial applications and electromobility applications is to be proven.

9.14 GaN Vertical Power Transistors

As a promising vertical device from GaN is the current aperture vertical electron transistor (CAVET) is described [Cho13]. Its structure is shown in Fig. 9.41.

The horizontal high-mobility electron channel at the interface of the AlGaN/GaN is located beyond the Gate electrode. A vertical GaN drift region is arranged through the aperture with the width L_{ap}. The buried current blocking layers, denoted as CBL in Fig. 9.41, consist of p-doped GaN. Since the structure sustains the voltage in vertical direction, a high blocking voltage can be achieved with a small chip area.

The development of vertical GaN devices, however, is in a very early stage. Realized devices reached 200 – 300 V, with high leakage current. The devices are of normally-on type and require a need negative gate voltage for blocking [Cho13]. Nevertheless, vertical GaN devices are promising as power devices due to the material characteristics of GaN.

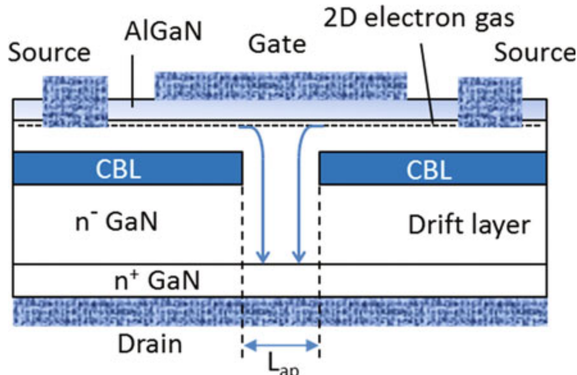


Fig. 9.41 Hybrid-current aperture vertical electron transistor (CAVET) fabricated with GaN. Figure following [Cho13]

9.15 Outlook

The MOSFET is a unipolar device. In forward direction, a "knee", i.e. threshold forward voltage does not occur. The MOSFET offers many advantages: its control is easy and requires only low power. The switching slopes are adjustable via gate resistances. At turn-off no tail current exists, in difference to IGBTs. The MOSFET has minor switching losses and high switching frequencies are possible. Moreover, it is short-circuit resistant and possesses a rectangular safe operating area.

Therefore, a MOSFET will always be the preferred device in applications whenever it is possible to apply it. MOSFETs can be paralleled without problems and the series connection is possible as well. Even high-power applications (50 kV, some kA), which require fast switching behavior and where costs are of minor importance, have been realized with parallel and series connection of MOSFETs.

A disadvantage of the Si-MOSFET, however, is the inverse body diode, which has insufficient properties.

When designed for higher blocking voltages, the resistance R_{on} increases strongly. The introduction of the superjunction principle, which does not show this strong correlation, represents an important development. Nowadays, superjunction MOSFETs are available for 600 and 900 V, even 1000 V devices are possible in principle. But the technological complexity is increases at higher voltage levels. Consequently, applications requiring higher voltages are still dominated by bipolar devices.

With superjunction devices in the range of 600 V, it is expected that future generations will achieve a further reduced R_{on} per area.

In the low voltage range (< 100 V), the trench technology leads to a reduction of the on-state resistance R_{on}. Further development in microelectronics will allow to produce finer structures and, thus, a higher cell density will be possible. This will further reduce the on-state losses. Furthermore, measures for the reduction of the capacitances are taken to reduce switching losses.

Field controlled devices made of SiC become available. Due to the possible very thin drift zone and the feasible higher doping of the base, they attain a much lower resistance R_{epi}, see Fig. 6.8. They extend the range of unipolar field controlled devices well above the range of 1000 V, several kV are possible. Also below 1000 V they will compete with Silicon devices, because of their potential to offer low R_{on}. SiC MOSFETs for 3.3 kV are announced, SiC MOSFETS of 10 kV and more are in development.

In the voltage range below 1000 V, lateral GaN transistors compete with Si MOSFETs. With GaN, low switching losses are possible, and a $R_{on} \cdot Q_g$ is achieved that is one decade lower compared to Si MOSFETs. Therefore GaN devices are very attractive for high-frequency applications. Many effort is done today to investigate the reliability for industrial applications and electromobility applications.

References

[Aga06] Agarwal, A., Ryu, S.H.: Status of SiC power devices and manufacturing issues. In: CS MANTECH Conference, pp. 215–218. Vancouver, Canada, 24–27 Apr 2006

[Bei16] Beier-Möbius, M., Lutz, J.: Breakdown of gate oxide of 1.2 kV SiC-MOSFETs under high temperature and high gate voltage. In: Proceedings of the PCIM Europe, Nuremberg (2016)

[Bei17] Beier-Möbius, M., Lutz, J.: Breakdown of gate oxide of SiC-MOSFETs and Si-IGBTs under high temperature and high gate voltage. In: Proceedings of the PCIM Europe, Nuremberg (2017)

[Bel05] Belverde, G., Magrì, A., Melito, M., Musumeci, S., Pagano, R., Raciti, A.: Efficiency improvement of synchronous buck converters by integration of Schottky diodes in low-voltage MOSFETs. In: Proceedings of the IEEE ISIE, pp. 429–434 (2005)

[Cal04] Calafut, D., Trench power MOSFET lowside switch with optimized integrated Schottky diode. In: Proceedings of the International Symposium on Power Semiconductor Devices & ICs, pp. 397–400 (2004)

[Cha16] Chang, M.-H., Rutter, P.: Optimizing the trade-off between the $R_{DS(on)}$ of power MOSFETs and linear mode perfomance by local modification of MOSFET gain. In: Proceedings of the 28th ISPSD, pp. 379–382. Prague, Czech Republic (2016)

[Che01] Chen, X.-B., Sin, J.K.O.: Optimisation of the specific on-resistance of the COOLMOS. IEEE Trans. Electron Device **48**(2), 344–348 (2001)

[Che05] Chen, Y., Liang, Y., Samudra, G.: Theoretical analyses of oxide-bypassed superjunction power metal oxide semiconductor field effect transistor devices. Jpn. J. Appl. Phys. **44**(2), 847–856 (2005)

[Che17] Chen, K.J: Fluorine-implanted enhancement-mode transistors. In: Meneghini, M., Meneghesso, G., Zanoni, E. (eds.) Power GaN Devices – Materials, Applications and Reliability. Springer, Switzerland (2017)

[Cho13] Chowdhury, S., Swenson, B.L., Wong, M.H., Mishra, U.K.: Current status and scope of gallium nitride-based vertical transistors for high-power electronics application. Semicond. Sci. Technol. **28**, 074014 (2013)

[Col79] Collins, H.W., Pelly, B.: HEXFET, a new power technology cuts on-resistance boosts rating. Electron. Des. **17**, 36 (1979)

[Deb98] Deboy, G., März, M., Stengl, J.P., Sack, H., Tihanyi, J., Weber, H.: A new generation of high voltage MOSFETs breaks the limit line of silicon. In: Proceedings of IEDM, pp. 683–685 (1998)

[DeS16] De Santi C.: Field- and time-dependent degradation of power gallium nitride (GaN) high electron mobility transistors (HEMTs). In: Tutorial ESREF, Halle (2016)

[Dol04] Dolny, G.M., Sapp, S., Elbanhaway, A., Wheatley, C.F.: The influence of body effect and threshold voltage reduction on trench MOSFET body diode characteristics. In: Proceedings ISPSD, pp. 217–220. Kitakyushu (2004)

[EPI16] EPIGAN Data sheet HV 650

[Fow28] Fowler, R.H., Nordheim, L.: Electron emission in intense electric fields. Proc. R. Soc. Lond. Ser. A **119**, 173 (1928)

[Gra89] Grant, D.A., Gowar, J.: Power MOSFETS – Theory and Application. Wiley, New York (1989)

[Gur08] Gurfinkel, M., et al.: Time-dependent dielectric breakdown of 4H-SiC/SiO_2 MOS capacitors. IEEE Trans. Device Mater. Reliab. **8**(4), 635–641 (2008)

[Hil15] Hilt, O., Bahat-Treidela, E., Knauer, A., Brunner, F., Zhytnytska, R., Würfl, J.: High-voltage normally OFF GaN power transistors on SiC and Si substrates. MRS Bull. **40**(05), 418–424 (2015)

[Hof63] Hofstein, S.R., Heiman, F.P.: The silicon insulated-gate field-effect transistor. Proc. IEEE **51**(9), 1190–1202 (1963)

[Hon15] Honea, J., Zhan Wang, Z., Wu, Y.: Design and implementation of a high-efficiency three-level inverter using GaN HEMTs. In: Proceedings of the PCIM Europe, pp. 486–492 (2015)

[Hua17] Hua, M., Zhang, Z., Qian, Q., Wei, J., Bao, Q., Tang, G., Chen, K.J.: High-performance fully-recessed enhancement-mode GaN MIS-FETs with crystalline oxide interlayer. In: Proceedings of the 29th ISPSD, pp. 89–92. Sapporo (2017)

[Ima04] Imaizumi, M., Tarui, Y.: 2 kV Breakdown voltage SiC MOSFET technology. Mitsubishi Electric R&D Progress Report, March 2004. http://global.mitsubishielectric.com/pdf/advance/vol105/08_RD1.pdf (2004)

[Ish15] Ishida, M., Ueda, T.: GaN-based gate injection transistors for power switching applications. In: Japan-EU Symposium on Power Electronics. Tokyo, 15–16 Dec 2015

[Kan15] Kaneko, S., Kuroda, M., Yanagihara, M., Ikoshi, A., Okita, H., Morita, T., Tanaka, K., Hikita, M., Uemoto, Y., Takahashi, S., Ueda, T.: Current-collapse-free operations up to 850 V by GaN-GIT utilizing hole injection from drain. In: Proceedings of the 27th ISPSD, Hong Kong (2015)

[Kon06] Kondekar, P.N., Oh, H., Kim, Y.B.: Study of the degradation of the breakdown voltage of a super-junction power MOSFET due to charge imbalance. J. Korean Phys. Soc. **48**(4), 624–630 (2006)

[Lee82] Lee, H.G., Oh, S.Y., Fuller, G.: A simple and accurate method to measure the threshold voltage of an enhancement-mode MOSFET. IEEE Trans. Electron Devices **29**(2), 346–348 (1982)

[Lia01] Liang, Y., Gan, K., Samudra, G.: Oxide-bypassed VDMOS (OBVDMOS). An alternative to superjunction high voltage MOS power devices. IEEE Electron Device Lett. **22**, 407–409 (2001)

[Lid79] Lidow, A., Herman, T., Collins, H.W.: Power MOSFET technology. In: 1979 International Electron Devices Meeting, vol. 25, pp. 79–83 (1979)

[Liu15] Liu, C., Salih, A., Padmanabhan B., Jeon, W., Moens, P., Tack, M., Debacker E.: Development of 650v cascode GaN technology. In: Proceedings of the PCIM Europe, pp. 994–1001 (2015)

[Lor99] Lorenz, L., März, M.: CoolMOS^{TM} – a new approach towards high efficiency power supplies. In: Proceedings of the 39th PCIM, pp. 25–33. Nuremberg (1999)

[Lut18] Lutz, J., Aichinger, T., Rupp, R.: Reliability evaluation. In: Suganuma, K. (ed.) Wide Bandgap Power Semiconductor Packaging: Materials, Components, and Reliability. Woodhead Publishing, Elsevier (2018) (in preparation)

[Mic03] Michel, M.: Leistungselektronik, 3rd edn. Springer, Berlin (2003)

[Mim80] Mimura, T., Hiyamizu, S., Fujii, T., Nanbu, K.: A new field-effect transistor with selectively doped GaAs/n-$Al_xGa_{1-x}As$ heterojunctions. Jpn. J. Appl. Phys. **19**(5), L225–L227 (1980)

[Mit99] Mitlehner, H., Bartsch, W., Dohnke, K.O., Friedrichs, P., Kaltschmidt, R., Weinert, U., Weis, B., Stephani, D.: Dynamic characteristics of high voltage 4H-SiC vertical JFETs. In: Proceedings of the 11th ISPSD, pp. 339–342 (1999)

[Miu06] Miura, N., et al.: Successful development of 1.2 kV 4H-SiC MOSFETs with the Very low on-resistance of 5 mΩcm^2. In: Proceedings of the 18th ISPSD, Naples, Italy (2006)

[Miy15] Miyamoto, H., et al.: Enhancement-mode GaN-on-Si MOS-FET using Au-free Si process and its operation in PFC system with high-efficiency. In: Proceedings of the 27th ISPSD, pp. 209–212. Honkong (2015)

[Mor14] Morita, T., Tanaka, K., Ujita, S., Ishida, M., Uemoto, Y., Ueda, T.: Recent progress in gate injection technology based GaN power devices. In: Proceedings of the ISPS, Prague, pp 34–37 (2014)

[Muh16] Muhsen, H.: Ph.D. thesis, Chemnitz University of Technolgy (2017)

[Nak11] Nakamura, T., Nakano, Y., Aketa, M., Nakamura, R., Mitani, S., Sakairi, H., Yokotsuji, Y.: High performance SiC trench devices with ultra-low ron. In: IEEE International Electron Devices Meeting (IEDM) (2011)

[Nic00] Nicolai, U., Reimann, T., Petzoldt, J., Lutz. J.: Application Manual Power Modules. Semikron, ISLE Verlag, Ilmenau (2000)

[Paw08] Pawel, I., Siemieniec, R., Born, M.: Theoretical evaluation of maximum doping concentration, breakdown voltage and on-state resistance of field-plate compensated devices. In: Proceedings of ISPS'08, Prague (2008)

[Pol07] Polenov, D., Lutz, J., Pröbstle, H., Brösse, A.: Influence of parasitic inductances on transient current sharing in parallel connected synchronous rectifiers and Schottky-Barrier diodes. IET Circ. Devices Syst. **1**(5), 387–394 (2007)

[Rum09] Rumyantsev, S., Shur, M., Levinshtein, M., Ivanov, P., Palmour, J., Agarwal, A., Hull, B., Ryu, S.H.: Channel mobility and on-resistance of vertical double implanted 4H-SiC MOSFETs at elevated temperatures. Semicond. Sci. Technol. **24**(7), 075011 (2009)

[Ryu06] Ryu, S.H., et al.: 10 kV, 5A 4H-SiC Power DMOSFET. In: Proceedings of the 18th ISPSD, Naples, Italy (2006)

[Sai03] Saito, W., Takada, Y., Kuraguchi, M., Tsuda, K., Omura, I., Ohashi, H.: High breakdown voltage AlGaN/GaN power-HEMT design and high current density switching behavior. IEEE Trans. Electron Devices **50**(12), 2528–2531 (2003)

[Scd96] Schlund, B., et al.: A new physic-based model for time-dependent-dielectric-breakdown. In: Proceedings of the International Reliability Physics Symposium, pp. 84–92 (1996)

[She90] Shenai, K., Baliga, B.J.: Monolithically integrated power MOSFET and Schottky diode with improved reverse recovery characteristics. IEEE Trans. Electron Devices **37**(3), 1167–1169 (1990)

[Sie06c] Siemieniec, R., Hirler, F., Schlögl, A., Rösch, M., Soufi-Amlashi, N., Ropohl, J., Hiller, U.: A new fast and rugged 100 V power MOSFET. In: Proceedings of EPE-PEMC, Portoroz, Slovenia (2006)

[Sie12] Siemieniec, R., Nöbauer, G., Domes, D.: Stability and performance analysis of a SiC-based cascode switch and an alternative solution. Microelectron. Reliab. **52**, 509–518 (2012)

[Sin04] Singh, R., Hefner, A.R.: Reliability of SiC MOS devices. Solid-State Electron. **48**, 1717–1720 (2004)

[Sod99] Sodhi, R., Malik, R., Asselanis, D., Kinzer, D.: High-density ultra-low R_{dson} 30 volt N-channel trench FETs for DC/DC converter applications. In: Proceedings of ISPSD'99, pp. 307–310 (1999)

[Spi02] Spirito, P., Breglio, G., d'Alessandro, V., Rinaldi, N.: Thermal instabilities in high current power MOS devices: experimental evidence, electro-thermal simulations and analytical modeling. In: Proceedings of the 23rd International Conference on Microelectronics (MIEL), pp. 23–30. Serbia, Europe (2002)

[Squ13] Sque, S.: High-voltage GaN-HEMT devices, simulation and modelling. In: Tutorial at ESSDERC, Bucharest (2013)

[Ste92] Stengl, J.P., Tihanyi, J.: Leistungs-MOSFET-Praxis. Pflaum-Verlag, München (1992)

[Sze81] Sze, S.M., Physics of semiconductor devices. Wiley, New York (1981)

[Ued17] Ueda, D.: Properties and advantages of gallium nitride. In: Meneghini, M., Meneghesso, G., Zanoni, E.: Power GaN Devices – Materials, Applications and Reliability. Springer, Switzerland (2017)

[Uem07] Uemoto, Y., et al.: Gate injection transistor (GIT) – a normally-off AlGaN/GaN power transistor using conductivity modulation. IEEE Trans. Electron Devices **54**(12), 3393–3399 (2007)

[Uhn15] Uhnevionak, B., Burenkov, S., Strenger, C., Ortiz, G., Bedel-Pereira, E., Mortet, V., Cristiano, F., Bauer, A.J., Pichler, P.: Comprehensive study of the electron scattering mechanisms in 4H-SiC MOSFETs. IEEE Trans. Electron Devices **62**(8), 2562–2570 (2015)

[Wil17] Williams, R.K., Darwish, M.N., Blanchard, R.A., Siemieniec, R., Rutter, P., Kawaguchi, Y.: The trench power MOSFET – part II: application specific VDMOS, LDMOS, packaging, and reliability. IEEE Trans. Electron Devices **64**(3), 692–712 (2017)

[Wür15] Würfl, J.: GaN power switching transistors: survey on device concepts and technology. In: Tutorial GaN Based Power Electronics, Organized by Oliver Häberlen, 45th European Solid-State Device Research Conference ESSDERC (2015)

[Zen00] Zeng, J., Wheatley, C.F., Stokes, R., Kocon, C., Benczkowski, S.: Optimization of the body-diode of power MOSFETs for high efficiency synchronous rectification. In: Proceedings of the ISPSD, pp. 145–148 (2000)

[Zin01] Zingg, R.P.: New benchmark for RESURF, SOI, and super-junction power devices. In: Proceedings of the ISPSD, pp. 343–346. Osaka (2001)

Chapter 10
IGBTs

10.1 Mode of Function

A lot of work was spent to combine bipolar devices with their superior current density with the possibility of voltage-control as given in MOSFETs. Early works tried to combine thyristor-related structures with MOS gate control. However, a transistor-based device won the race. Plummer and Scharf from Stanford University California [Plu80, Scf78] worked on planar MOS controlled TRIACs, they discussed an intermediate bipolar state as a step before the intended thyristor mode. They even mentioned a kink in the characteristics toward better conductivity and showed a measured IGBT-like I–V characteristics. However, the final aim was thyristor mode. The Insulated Gate Bipolar Transistor (IGBT) was invented in the United States of America by F. Wheatley and H. Becke from RCA Corporation in U.S. [Bec80]. The patent clearly points out "Although this creates a four layer device, the conductivities and geometries of the four semiconductor regions are manipulated so as not to form a regenerative thyristor". In 1982, the same year as said patent was published, the first experimental demonstration of a practical discrete vertical IGBT device was reported by Baliga et al. from GE [Bal82], and the advantage of a 10 times higher current density compared and MOSFET and two times compared to the bipolar transistor was shown. Also Russel et al. from RCA submitted a paper describing the IGBT in the end of 1982 [Rus83]. The IGBT was firstly referred to as COMFET (Conductivity Modulated FET) [Rog88, Rus83], also the term IGT (Insulated Gate Transistor) [Bal83] was used. A report on the first story of the IGBT is given in [She15]. A description of the technical details leading to a mature device is made in [Iwa17].

It started an intensive work of several groups on the new device. Nakagawa and Ohashi from Toshiba showed the first 1200 V device [Nak84]. About 10 years later, IGBTs were introduced in the market by manufacturers from Japan and Europe. In a short time, IGBTs won an increasing share of applications, and they

© Springer International Publishing AG 2018
J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8_10

replaced the formerly used bipolar power transistors, and nowadays even GTO-thyristors in the high power range.

In a first rough approach, the IGBT can be seen as a MOSFET, in which the n^+-layer at the drain side is replaced by a p-layer. Figure 10.1 shows the structure of the IGBT.

The notations collector and emitter have been taken from the bipolar transistor, the notations anode (for collector) and cathode (for emitter) also make sense.

If at the IGBT a positive voltage between collector C and emitter E is applied, the device is in the blocking mode. If now a voltage V_G higher than the threshold voltage V_T is applied between gate and emitter, an n-channel is created; electrons flow to the collector (Fig. 10.1). At the collector side pn-junction, a voltage in forward direction is generated, and holes from the p collector layer are injected into the lowly doped middle layer. The injected holes allow an increased charge carrier density; the increased carrier density lowers the resistance of the middle layer, and the conductivity of the middle layer is modulated. Like in a MOSFET, the turn-on and turn-off of the IGBT happen by the creation and removal of an n-channel by applying a gate voltage. Regarding threshold voltage and channel resistance, the same holds as described in Chap. 9 for the MOSFET.

Figures 10.1 and 10.2 show a pnpn-structure as in a thyristor, however the action of this thyristor is strictly avoided by a high-conductive emitter short R_S realized by a highly doped p^+-layer in the center of the cell [Nak84]. Figure 10.2b shows the equivalent circuit of the four-layer structure. With the pnp and npn partial transistor, a parasitic thyristor structure is visible. The emitter of the npn partial transistor and its base are shortened by the resistor R_S. With this, the current gain of the npn partial transistor is eliminated at low current. But at very high current, the npn transistor can be activated, and the parasitic thyristor can be triggered into the on-state mode

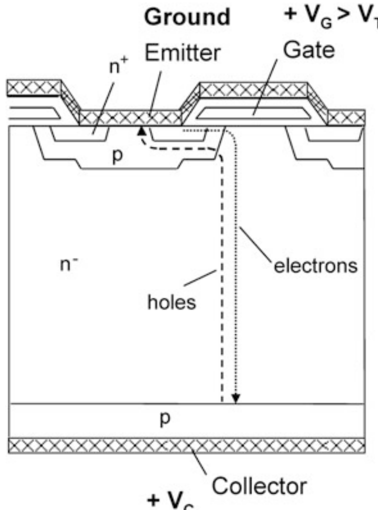


Fig. 10.1 IGBT in conduction mode at $V_G > V_T$, electron and hole current

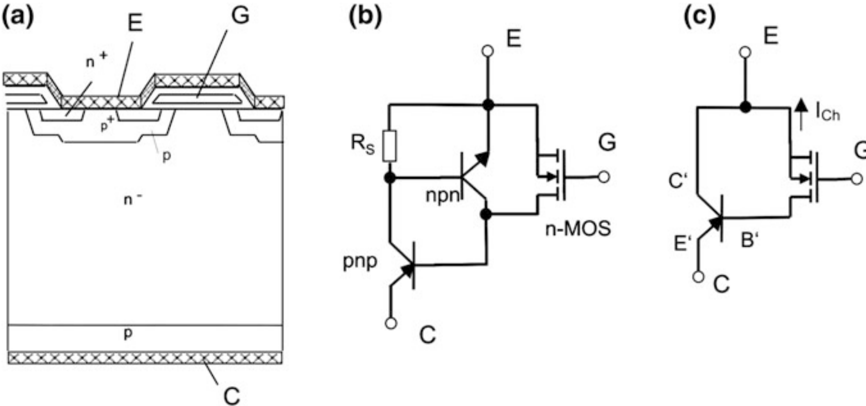


Fig. 10.2 IGBT **a** simplified structure **b** equivalent circuit with parasitic npn-transistor and resistor R_S **c** simplified equivalent circuit

with the internal feedback loop with the partial pnp-transistor. This effect is denoted as latch-up: the device can no longer be controlled by the MOS gate. Latching of the parasitic thyristor is a destructive effect for the IGBT.

For sufficiently low R_S, the npn partial transistor can be neglected, and the simplified equivalent circuit, as shown in Fig. 10.2c, is obtained. This is the most important equivalent circuit for understanding the IGBT. The terminals of the pnp-transistor are denoted with C', E' and B'. The collector C of the IGBT is the emitter E' of the pnp-transistor. Concerning the physics of the IGBT, it is an emitter.

The channel is created for a gate voltage V_G higher than the threshold voltage; in the base terminal of the pnp-transistor the channel current I_{CH} is flowing. For the current $I_{C'}$ at C' then holds $I_{C'} = \beta_{pnp} \cdot I_{CH}$, or, with the relation $\alpha = \beta/(\beta + 1)$ known from the sections on bipolar transistors

$$I_{C'} = \frac{\alpha_{pnp}}{1 - \alpha_{pnp}} \cdot I_{CH} \tag{10.1}$$

For the collector current of the IGBT holds

$$I_C = I_{C'} + I_{CH} = \frac{\alpha_{pnp}}{1 - \alpha_{pnp}} \cdot I_{CH} + I_{CH} = \frac{1}{1 - \alpha_{pnp}} \cdot I_{CH} \tag{10.2}$$

Therefore the collector current of the IGBT is always higher than its channel current. The saturation current of the IGBT will also be much higher than that of the MOSFET. With the parameter of the channel conductivity κ, defined for the MOSFET in Eq. (9.4), results for the IGBT

$$I_{Csat} = \frac{1}{1 - \alpha_{pnp}} \cdot \frac{\kappa}{2} \cdot (V_G - V_T)^2 \tag{10.3}$$

However, α_{pnp} must be adjusted not too high, and to meet all the different requirements, it should be adjusted very exactly. The current necessary for latch-up must be shifted to such a high value which will not occur in application. To achieve this, different measures are used in the design of the IGBT, which will be explained in the following.

10.2 The I–V Characteristic of the IGBT

The I–V characteristic of an IGBT in forward direction is shown in Fig. 10.3. The characteristic has some similarities to that of a MOSFET.

For a gate voltage V_G higher than the threshold voltage V_T, the channel is open. The IGBT differs from the MOSFET by the junction voltage of the additional pn-junction at the collector side. The IGBT is operated in the saturation region, as is usual for power devices and as it was already the case with the MOSFET and the bipolar transistor. The operation point is on the branch of the characteristic for V_G = 15 V. At this branch for operation at a given current I_C, the generated voltage drop V_C is read off.

Figure 10.4 compares the characteristic of an IGBT for V_G = 15 V with that of a bipolar transistor, also at a high drive level of I_B = 2.5 A. Both devices are specified for 600 V and have a comparable area. The threshold voltage caused by the reverse pn-junction is recognizable in the characteristics of the IGBT. At low current densities, the bipolar transistor has the inferior voltage drop because no threshold voltage occurs with it. However, at higher current densities, above 14 A, the forward voltage of the IGBT is much lower than that of the bipolar transistor. The shown IGBT is rated to a nominal current of 20 A, this current level is not reached with the used bipolar transistor even at a high base current.

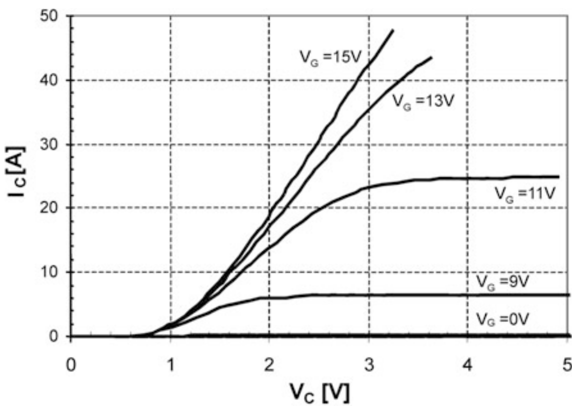


Fig. 10.3 I–V characteristic of a 20 A / 600 V IGBT

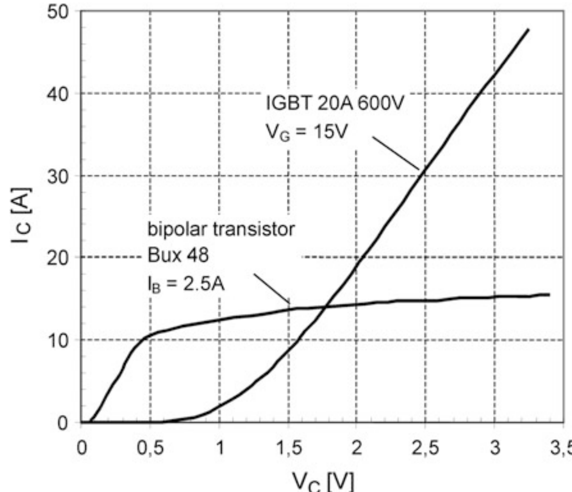


Fig. 10.4 Forward voltage-current characteristic of an IGBT in comparison to a bipolar transistor

A comparison of power devices of a higher voltage range would show the difference even more drastically. The IGBT especially can be designed also for higher voltages; it is not restricted like the MOSFET or bipolar transistor by physical mechanisms. Meanwhile, IGBTs have been fabricated for voltages up to 8 kV [Rah02], they have been commercially available in 2016 for voltages up to 6.5 kV.

10.3 The Switching Behavior of the IGBT

The determination of the switching behavior of the IGBT is done in a circuit according to Fig. 10.5 with inductive load. The time-constant of the load $\tau = L/R$ is chosen so high that courses of voltage and current can be assumed as constant before the switching instant.

The turn-on process of a field-controlled device was discussed already with the MOSFET. The relation between the rise time of the current, the fall time of the voltage, the internal capacities and the chosen gate resistors are the same as were discussed in the context with Fig. 9.18. The IGBT is used with a freewheeling pin-diode in almost all applications; at the turn-on process, it additionally has to take over the reverse current peak and the stored charge of the freewheeling diode. The processes are shown in Figs. 5.20 and 5.21 and are described in context with Eqs. (5.82) to (5.84). For the turn-on energy per pulse in the IGBT, it can be given under the same simplifications

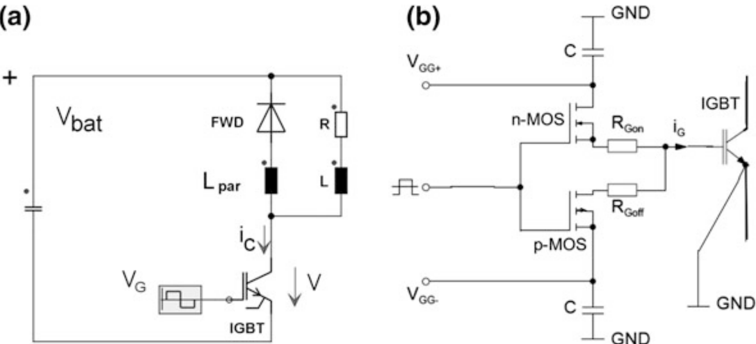


Fig. 10.5 Measurement of the IGBT switching behavior. **a** Power circuit **b** Simplified output stage of a gate drive circuit containing gate resistors R_{Gon} and R_{Goff}. Adapted from [Nic00]

$$E_{on} = \frac{1}{2} \cdot V_{bat} \cdot (I_C + I_{RRM}) \cdot t_{ri} + \frac{1}{2} \cdot V_{bat} \left(I_C + \frac{2}{3} I_{RRM} \right) \cdot t_{fv} \tag{10.4}$$

A more exact determination is done with the oscilloscope, see Eq. (9.33).

At turn-off of the IGBT, the positive gate voltages are put on zero or a negative value, and in the first time interval, processes appear like those described with the MOSFET in the context of Fig. 9.19. As long as the stored charge in the IGBT is not too high, similar relations between the rise time of the voltage, the internal capacities and the chosen gate resistance are valid. If the gate capacity is discharged abruptly, the channel current is interrupted. Usually a gate resistor is applied, and V_G is reduced to the value of the Miller plateau at $V_G = V_T + g_{fs} \cdot I_D$. While the voltage increases, the current in the IGBT in a circuit with inductive load flows on continuously until the voltage is higher than the applied battery voltage V_{bat}. During the voltage rise time, the channel current is reduced and the hole current flowing through the p-well is increased in the same amount. Carriers from the n-base are extracted by the hole current. The hole current is increased compared to the steady-state conduction mode. The problem of possible latch-up is most serious at the turn-off event. The requirement that latch-up must be avoided limits the maximal possible current which an IGBT can turn off.

The hole current leads to the removal of the charge carriers from the n-base, a space charge is built-up, and the device takes over the applied voltage. After the voltage has increased to V_{bat}, the current falls steeply. With IGBTs, the slope of the decreasing current di_c/dt can be adjusted by the gate resistor only with limited effect. The slope di_c/dt causes an inductive voltage peak at the parasitic inductance, and the forward recovery voltage peak V_{FRM} of the freewheeling diode adds to this. The voltage peak amounts to

$$\Delta V = L_{par} \cdot \frac{di_c}{dt} + V_{FRM} \tag{10.5}$$

The higher the nominal voltage range, the more significant the part of the V_{FRM} may be. See Sect. 5.6 especially Fig. 5.15.

As the main difference to the MOSFET and to the bipolar transistor, the IGBT features a tail current at turn-off. A measurement of the turn-off of the IGBT including the tail current is shown in Fig. 10.6. The current falls down to the value I_{tail}, and then it goes down slowly during the time interval t_{tail}. The measurement of the end point of the tail current is difficult since it decreases very slowly. The time of the tail current t_{tail} is determined by the recombination of the remaining charge carriers in the device. With the typical high charge-carrier lifetime in the NPT IGBT, t_{tail} can amount to several µs, while t_{rv} is in the order of some 100 ns and t_{fi} is in the range of 100 ns.

During the time of the tail current, the voltage is high and the losses created in this interval cannot be neglected. In practice, the determination of the turn-off energy per pulse is mostly done with an oscilloscope; the product of current waveform and voltage waveform is executed and integrated over the turn-off time interval. A simplified estimation can be made with

$$E_{off} = \frac{1}{2} \cdot V_{bat} \cdot I_C \cdot t_{rv} + \frac{1}{2} \cdot (V_{bat} + \Delta V) \cdot I_C \cdot t_{fi} + \frac{1}{2} I_{tail} \cdot V_{bat} \cdot t_{tail} \tag{10.6}$$

IGBTs are usually used in bridge configurations. The IGBT is turned off mostly with a negative gate voltage, as in Fig. 10.6. In the blocking mode, a voltage of −15 V is applied, sometimes also a smaller voltage of −8 V. This does not have,

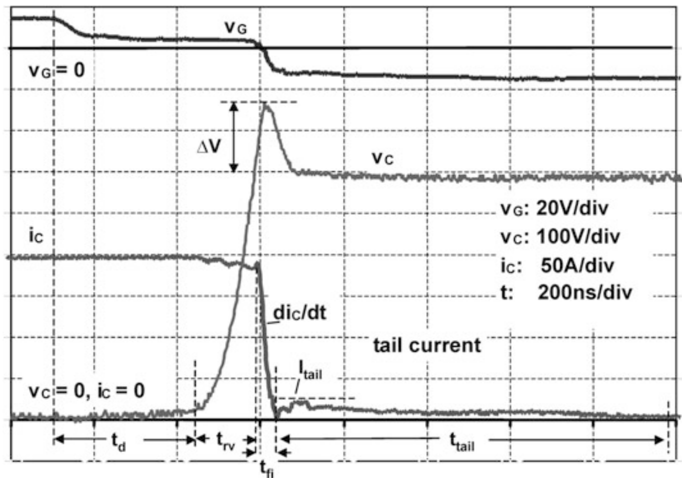


Fig. 10.6 Turn-off of an NPT-IGBT (200 A/1200 V module BSM200GB120DN2, manufacturer Infineon). T = 125 °C, R_{Goff} = 3.3 Ω

beside of the waveform of the gate voltage, much effect on the waveforms of current and voltage at turn-off.

10.4 The Basic Types PT-IGBT and NPT-IGBT

In the very first IGBT-structures the n^+-substrate of the MOSFET was replaced with a p^+-substrate. These structures were very sensitive to latch-up of the parasitic thyristor. The behavior could be improved by adding a medium-doped n-layer, a so-called n-buffer, between p^+-substrate and lowly doped n^--layer. This n-buffer must have a sufficient doping N_{buf} [Nak85]. The electric field can penetrate into the n-buffer, a trapezoidal shape of the electric field is given. From this characteristic, the notation Punch-Through IGBT or *PT-IGBT* was derived (In the proper meaning of the word, this notation is not correct, see Sect. 5.1). The structure is shown in Fig. 10.7.

As already mentioned, α_{pnp} must be adjusted and must not be too high. The doping of the buffer N_{buf} is of influence on it. In the section on the bipolar transistor, α was composed of two terms

$$\alpha = \gamma \cdot \alpha_T \tag{10.7}$$

For the estimation of the emitter efficiency γ, the Eq. (7.23) can be used; for the given case of a p-emitter it can be written as [Mil89]

$$\gamma = \frac{1}{1 + \frac{\mu_n}{\mu_p} \cdot \frac{N_{buf}}{N_{sub}} \cdot \frac{L_p}{L_n}} \tag{10.8}$$

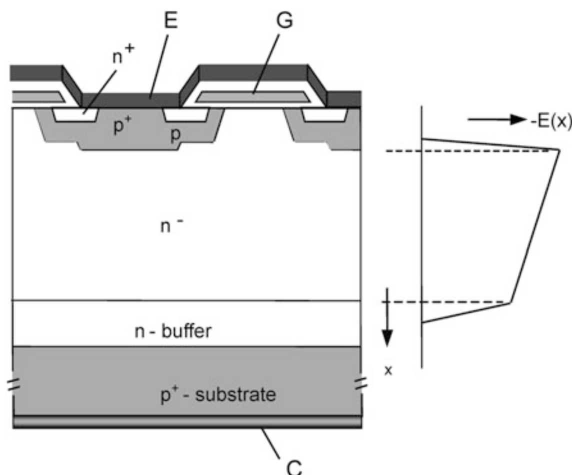


Fig. 10.7 PT-IGBT, structure and field shape

Since L_p, the diffusion length of holes in the n-buffer, and L_n, the diffusion length of electrons in the substrate, are in the same order of magnitude, and since μ_n/μ_p amounts to 2–3, it is difficult to make the denominator in Eq. (10.8) significantly larger than 1, as long as the doping of the buffer N_{buf} is not in the same range of magnitude as the doping of the p$^+$-substrate N_{sub}. Any control of γ will be difficult. For the PT-IGBT, we can therefore assume $\gamma \approx 1$.

Thus the adjustment of α_{pnp} in the PT-IGBT is done by the transport factor α_T. According to Eq. (7.29), the width of the base-layer w_B and the diffusion length L_p in the substrate are important factors for this[1]:

$$\alpha_T = 1 - \frac{w_B^2}{2 \cdot L_p^2} \tag{10.9}$$

and L_p was given in Eq. (2.118)

$$L_p = \sqrt{D_p \cdot \tau_p} \tag{10.10}$$

With a low carrier lifetime, L_p, α_T and finally α_{pnp} are reduced. For the reduction of the carrier lifetime, the technologies for the creation of recombination centers as described in Chap. 4 are used. Platinum diffusion, electron irradiation, irradiation with He^{++}-ions or protons, or a combination of two of these processes is used; the specific process is different for different suppliers and their special device generations. The devices feature low turn-off losses.

The fabrication of the basic material for PT-IGBTs is done with an epitaxy process. N-buffer and n–layer are deposited on a p$^+$-substrate. This technology is well applicable in the voltage range up to 600 V. For 1200 V devices, the epitaxial technology demands an increased effort because of the necessary thick epitaxial layer. PT-IGBTs dominated the applications at blocking voltages up to 600 V over several years.

As an alternative concept, the so-called *NPT-IGBT* (Non-Punch-Through IGBT) was introduced. It is based on a suggestion of Jenö Tihanyi [Tih88] and was first realized by Siemens (today Infineon) [Mil89]. The structure is shown in Fig. 10.8. The shape of the electric field is triangular. The device for the same blocking voltage, which is given by the area under the line $E(x)$, must therefore be designed with a much thicker base width w_B. For the very first types of NPT-IGBTs, additionally a relatively high distance between the end of the space charge at $x = w$ and the p-collector layer at $x = w_B$ was chosen. Thus the effective base of the pnp-transistor at high voltage is $w_B - w$. A widened effective base, as expressed with Eq. (10.9), can somewhat reduce α_T and therewith α_{pnp}.

[1]Equation (10.9) is usually derived for low injection. The considerations, however, will also be valid for high injection, qualitatively.

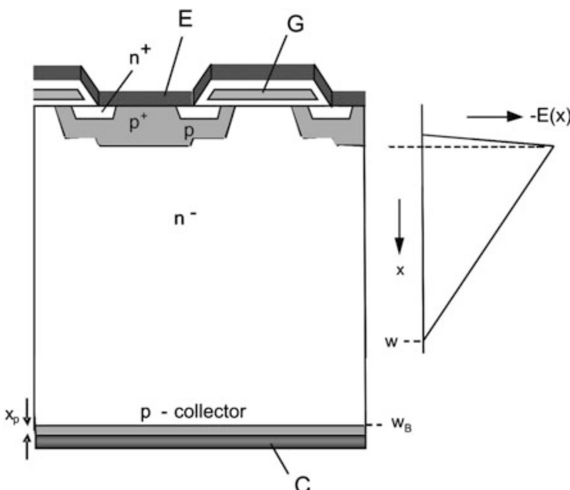


Fig. 10.8 NPT-IGBT. Structure and field shape

The main control of α_{pnp} is done via the emitter efficiency γ. The p-collector layer is lowly doped and its penetration depth is very shallow; with this the low emitter efficiency is adjusted. In Sect. 3.4, the Eq. (3.100) was derived for the emitter efficiency, for p-emitter with $p_L \approx n_L$ it is written as

$$\gamma = 1 - q \cdot h_p \frac{p_L^2}{j} \tag{10.11}$$

To achieve low γ, the emitter parameter h_p must be large. For the p-emitter, it can be expressed by Eq. (3.98)

$$h_p = \frac{D_n}{p^+ \cdot L_n} = \frac{D_n}{p^+ \cdot x_p} \tag{10.12}$$

For small x_p, the diffusion length L_n can be replaced by the very shallow penetration depth x_p of the p-emitter. At low p^+ and $x_p < 1$ µm for the given structure, the emitter efficiency γ will be low. High h_p means a large contribution of the emitter recombination at the total recombination. Special measures to reduce the carrier lifetime are no longer necessary.

The NPT-IGBT is therefore very robust against latch up, and it features high short-circuit robustness [Las92]. A further advantage results from this type of control of the plasma modulation. The temperature dependency of the forward voltage is very suitable for parallel connection. The voltage drop V_C at a constant collector current I_C and a constant gate voltage V_G is increasing with temperature, at the typical operation current range of I_C.

The voltage drop across the lowly doped middle region was estimated for a pin diode by means of Eq. (5.47) as

$$V_{drift} = \frac{w_B^2}{(\mu_n + \mu_p) \cdot \tau_{eff}} \tag{10.13}$$

with the effective carrier lifetime

$$\frac{1}{\tau_{eff}} = \frac{1}{\tau_{HL}} + \frac{h_p \cdot p_L^2}{w_B \cdot \overline{p}} + \frac{h_n \cdot p_R^2}{w_B \cdot \overline{p}} \tag{10.14}$$

In this Eq. (10.14) the two last terms on the right-hand side stand for the contribution of the emitter regions. Compared to pin diodes, a different profile of the plasma is given for the basic IGBT types (see Fig. 10.8). However, Eq. (10.13) may be used for estimation. For a PT-IGBT structure, the lifetime τ_{HL} dominates the effective lifetime τ_{eff}. τ_{eff} increases with increasing temperature. Depending on the used recombination centers, τ_{HL} increases by a factor of two to four for a temperature increase from 25 to 125 °C. The influence of τ_{eff} dominates in Eq. (10.13), and V_{drift} and therewith V_C decrease with increasing temperature.

For an NPT-IGBT the carrier lifetime τ_{HL} in the lowly doped middle region is adjusted as high, and both terms of emitter recombination dominate in the influence on τ_{eff}. With increasing temperature also τ_{HL} increases, but this is of low effect to τ_{eff}, since the emitter terms dominate, and their temperature dependency is weak. The temperature dependency of voltage across the drift region according to Eq. (10.13) is therefore dominated by the temperature dependency of the mobilities. The mobilities decrease strongly with increasing temperature, it holds $(\mu_n + \mu_p)_{(125\,^\circ\mathrm{C})} \approx 0.5 \cdot (\mu_n + \mu_p)_{(25\,^\circ\mathrm{C})}$. V_{drift} and therewith V_C increase with increasing temperature.

This V_C temperature dependency, also assigned as "positive temperature coefficient of V_C", on the one hand leads to increased conduction losses at high operation temperature. But on the other hand, this is of advantage if devices are connected in parallel: if one of the devices has a lower V_C because of production-induced scattering of V_C and therefore takes more current, its temperature will increase. Then its V_C increases and its current is reduced. With this a negative feedback is given, the system of parallel-connected devices is stabilized.

The manufacturing process of the NPT-IGBT can be controlled more exactly; the adjustment of the emitter efficiency at the collector side can be executed very accurately with ion implantation technology. The NPT-IGBT has become the dominating device in application, because of its robustness and because of the suitable behavior at parallel connection. Meanwhile, also NPT-IGBTs for 600 V have been developed and established in the market.

At turn-off, the NPT-IGBT has a long tail current, see Fig. 10.6. The PT-IGBT has a shorter, but higher tail current. This will be more understandable after the internal plasma distribution has been considered.

10.5 Plasma Distribution in the IGBT

Using the example of the NPT-IGBT, the internal distribution of the charge carriers, the plasma distribution, shall be investigated. Figure 10.9 shows a simulation of this shape of the plasma in the on-state mode for a 1200 V IGBT at different forward-current densities. The lowly doped n-base of the IGBT is flooded with free carriers. Because of neutrality, $n \approx p$ holds for the bipolar device. The shape of the hole distribution is therefore almost identical to the electron distribution shown in Fig. 10.9. The cell structures are on the left-hand side in Fig. 10.9; the vertical coordinate x is the same as in Fig. 10.8. The right-hand side p-collector layer has a penetration depth of less than 1 μm for the NPT-IGBT, it cannot be recognized.

Compared to the plasma distribution of a diode (see Fig. 5.6), the plasma distribution is strongly reduced at the side of the cell structures for the conventional IGBT. It corresponds to the plasma shape in a pnp-transistor; the collector of the IGBT in this case is the emitter of the pnp-transistor, compare Fig. 10.2. In a transistor, typically the plasma density decays from the emitter side to the collector side, see Fig. 7.6 for comparison.

Considering the rated voltage of 1200 V, the figure shows a base width w_B of 250 μm which is very wide. The first NPT IGBTs have been designed with w_B = 220 μm. According to Eq. (5.1) and the respective Fig. 5.5, a base width of 110 μm would be sufficient for the condition of a triangular field shape. However, a very wide n-base was typical for the first generations of NPT-IGBTs.

At turn-off with inductive load, the device must first take over the voltage, while the load current is still flowing. The development of the internal plasma at the turn-off process is shown in Fig. 10.10.

Up to the time t = 0.69 μs the voltage increases up to the value of the applied battery voltage. During the voltage increase, the charge in the left part of the base is extracted quickly. After the battery voltage is reached, the current decays steeply

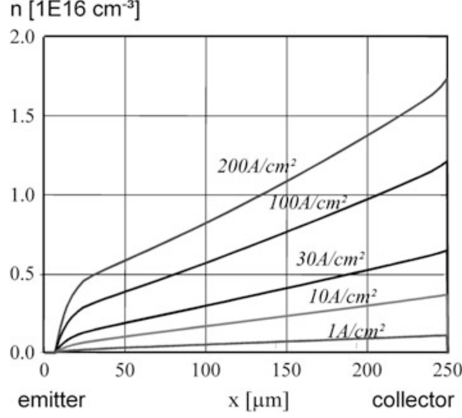


Fig. 10.9 Electron density at on-state, 1200 V NPT-IGBT. Figure from [Net99] © 2008 isle Steuerungstechnik und Leistungselektronik GmbH

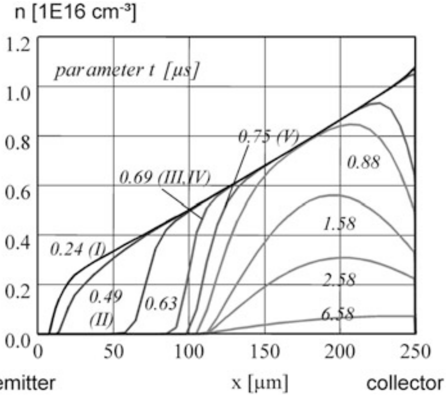


Fig. 10.10 Electron density in a 1200 V NPT-IGBT at turn-off of 80 A/cm^2. At the instant (III, IV), the device has taken the applied battery voltage, the current decreases. Figure from [Net99] © 2008 isle Steuerungstechnik und Leistungselektronik GmbH

down to the value of the tail current, similar as in Fig. 10.6. The extracted part of the base extends to approximately $x = 90$ µm. After $t = 0.75$ µs, the remaining charge on the right-hand side of the base is removed. In Fig. 10.10, this process is no more driven by the electric field. The applied voltage amounts to 600 V, the space charge has extended up to 100 µm and has taken up the voltage. There is only a very low increase in the extension of the space charge in the further process. For the removal of the carriers in the part close to the collector, recombination is the determining mechanism. Because of the high carrier lifetime, there is still a considerable charge in the device even up to $t = 6.58$ µs; during all this time t_{tail}, the tail current flows.

The voltage across the device has the value of the battery voltage during the time interval t_{tail}. Therefore, a main part of the turn-off losses is created by the tail current.

The PT-IGBT shows a shorter, faster decaying tail current because of its lower width of the base layer. However, PT-IGBTs have a more hanging-down plasma distribution in the conduction mode, caused by the reduced carrier lifetime. Compared to Fig. 10.9, the plasma density is lower at the emitter side, but higher at the collector side. As a result, the tail current in PT-IGBTs is shorter, but higher.

At the turn-off process described before, in which first the voltage increases before the current decays - switching with inductive load and also assigned as "hard switching" - the turn-off losses in the PT and NPT-IGBT are similar. However, the long tail current of the NPT-IGBT is a disadvantage in a turn-off process of the type "soft switching", at which at a circuit-induced zero crossing of the voltage or close-to-zero crossing the current is turned off, and the voltage increases slowly. Only a low share of the carriers is then removed in the first phase of the turn-off process. During the tail current time, the voltage increases and can lead to an additional increase in the current at the time t_{tail}. With this, additional losses are created.

10.6 Modern IGBTs with Increased Charge Carrier Density

The first IGBT-generations had a plasma distribution as shown in Fig. 10.9. This distribution is expected for a bipolar pnp-transistor, compare Fig. 7.6. Therefore, it was first supposed that the IGBT will have similar limits as a bipolar transistor: high voltage drop V_C in the on-state, and purely suitable for a blocking voltage above 1700 V. However, the IGBT could be improved.

Figure 10.10 shows that the plasma stored at the emitter side of the IGBT is quickly removed by the electric field while the voltage grows at the turn-off process. The plasma distribution, which is high at the collector side, low at the emitter side (see Fig. 10.9), leads to the effect that the main part of the charge carriers is removed in the tail phase. At the emitter side, the density of charge carriers could be significantly enlarged, without much increase in the turn-off losses. A higher density of carrier plasma will lead to a reduction of the voltage drop V_{drift} in the base layer, and consequently to a lower on-state voltage drop V_C. To achieve this, it was supposed during a long time that a new device will be necessary. Research and development activities on MOS-controlled thyristors (MCTs) and similar devices were started. However, it was discovered that the IGBT is capable to achieve the desired internal profile of the carrier plasma, and that the new devices might not be necessary.

10.6.1 Plasma Enhancement by High n-Emitter Efficiency

The effect of a possible increased plasma density at the emitter side was shown by Kitagawa et al. in 1993. They designated such a device as "Injection Enhanced Insulated Gate Bipolar Transistor (IEGT) [Kit93]. Kitagawa et al. discovered the effect at a Trench-IGBT, designed for 4.5 kV. The device had an increased charge carrier density at the emitter side and a surprisingly low voltage drop at forward conduction. The principal effect can also be explained with a planar IGBT, however [Lin06].

Figure 10.11 shows a partial section of an IGBT structure. In a first simplified investigation, the IGBT can be divided in two areas: an area of a bipolar pnp-transistor, and an area of a pin diode.

In the *transistor area*, the IGBT behaves as a pnp-transistor in the saturation region. The collector side of the IGBT corresponds to the emitter side of the pnp-transistor, see Fig. 10.2. The density of free carriers is high at the junction J_1 and decreases towards the junction J_2; at the junction J_2 it approximates to zero. For comparison, see Chap. 7 on the bipolar transistor (Fig. 7.6) and its description. From this, a shape of the carrier plasma as drawn in Fig. 10.9 results. In this figure, the junction J_2 is located at $x \approx 7$ µm. This shape of the plasma holds if it is determined by the bipolar transistor.

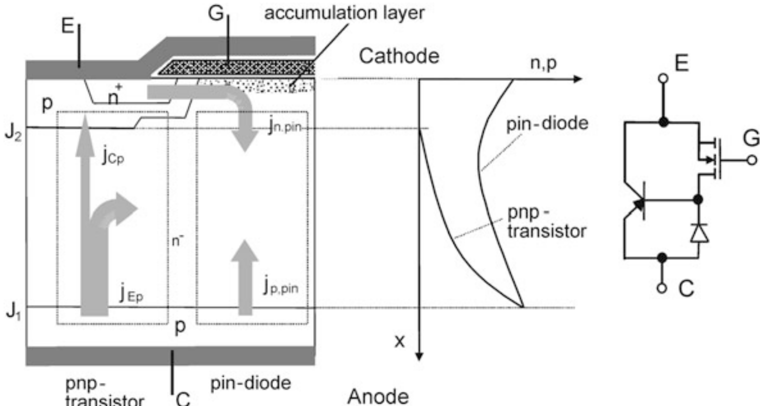


Fig. 10.11 Partition of an IGBT in a pnp-transistor area and a pin-diode area, carrier plasma distribution in both areas, equivalent circuit containing the pin-diode with MOS-switched n-emitter

In the diode area beneath the gate, similar conditions are given in the IGBT as in a pin diode. The electrons are supplied by the MOS channel. The MOS channel behaves as an ideal emitter; the current at this point is pure electron current. At the semiconductor surface between the p-wells, an accumulation layer of electrons will build up, created by the positive voltage at the gate above this area. Holes coming from the junction J_1 will not find a path here. The plasma distribution in this area will approximate the plasma distribution of a pin diode, as it was described in Chap. 5 with Fig. 5.6 and its discussion.

The voltage drop in the diode area, which is - in first approximation - the voltage drop V_C of the IGBT in forward conduction, is given by

$$V_C = V_{Ch} + V_{drift} + V_{J1} \tag{10.15}$$

whereby V_{Ch} is the voltage drop across the channel, V_{drift} is the voltage drop in the base layer and V_{J1} is the diffusion voltage of the pn-junction J_1. To make V_{drift} low, the relation of diode area to pnp-transistor area should be as high as possible. This can be achieved if the cells have a high distance. V_{drift} will decrease with increased cell distance (cell pitch). But then the cell density of the IGBT is decreased, and therefore also the voltage drop across the channel V_{Ch} will increase, as explained in the Chap. 9 on the MOSFET. For the planar IGBT in Fig. 10.11, one finds a cell distance at which the result for V_C has a minimum.

A more detailed investigation is given by [Omu97]. The whole top-side cell structure is considered to be an n-emitter. In the area between the p-wells, beneath the gate oxide, an accumulation layer of free electrons is built up, caused by the positive voltage at the gate. This accumulation layer is summarized with the p-wells to an effective n-emitter. An n-emitter of high efficiency shall be created. For this, the emitter must inject a high electron current. For the efficiency of an n-emitter, results are in analogy to (3.99)

$$\gamma = \frac{j_n}{j} \tag{10.16}$$

Here, j_n is the electron current delivered by the channel. Equation (10.16) can be written in a new form if $j = j_n + j_p$ is used

$$\gamma = \frac{j - j_p}{j} = 1 - \frac{j_p}{j} \tag{10.17}$$

To achieve a high γ, it is necessary

(a) to increase the distance between the cells. In [Omu97] it is shown that the share of j_n at the total current increases with increasing cell distance.
(b) to take care that j_p is only a small share of the total current density j. The hole current j_p flows across the p-well, see Fig. 10.11. If the area of the p-wells is reduced, also j_p is reduced. In the planar structure in Fig. 10.11, this happens by reduction of the p-well area.

So it is possible to increase the density of the plasma and therewith to decrease the voltage drop V_{drift} if the flow of the hole current j_p is reduced. The plasma distribution close to an emitter can be strongly influenced by its efficiency. On the example of the MOS-Controlled Diode (MCD), we have seen in Fig. 5.39 that the voltage drop in the drift layer can be controlled by the emitter injection. There, in the same way, we had an effective emitter efficiency of the p-well *and* the channel. In the MCD, the combination forms a p-emitter, and the current via the n-channel is the minority current. By injection of minority carriers from the channel, the emitter efficiency was reduced with the aim to reduce the plasma concentration at the side of the p-emitter. In the IGBT, the p-well, the channel and the electron accumulation layer form an n-emitter. The holes flowing from the p-well are the minority carriers. Reducing the minority carrier current j_p will increase the efficiency of the n-emitter; with this, the plasma density close to the emitter is increased. More charge carriers are available for current transport in the wide base layer, and the voltage drop V_C is decreasing.

This principle can be realized with planar structures and with trench structures. The trench structure offers special advantages for this. An example is the Infineon trench IGBT [Las00] . The trench cell is shown in Fig. 10.12. Emitter layers of the n^+-type and channels are arranged only in the middle of the cell between the two trenches. Outside the cell, a layer without contact to the emitter is arranged. Compare the IGBT trench cell with the MOSFET trench cell in Fig. 9.6. The mode of action of the trenches is different for the IGBT and for the MOSFET. At the MOSFET, a high part of the surface had to be equipped with channels, since the channel resistance R_{Ch} determines a main part of the total voltage drop; it is made low by arranging many channels in parallel. In the IGBT, the channel resistance is of minor effect to the voltage drop in the on-state. This voltage drop in the IGBT is determined by the plasma density in the middle layer, since the IGBT is a bipolar device. This plasma density must be high to reduce the voltage drop. Therefore, one has to reduce j_p to increase γ.

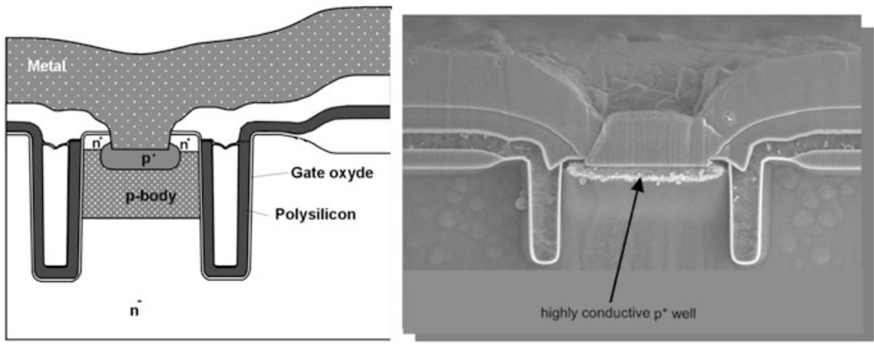


Fig. 10.12 IGBT trench cell. Structure (left), picture of a cross section of a cell made with a raster electron microscope (right). Figures from T. Laska, Infineon Technologies

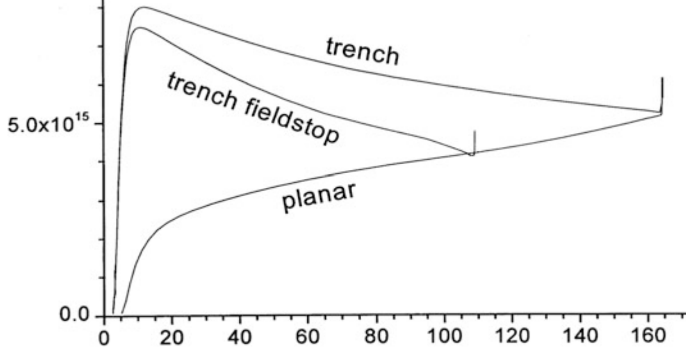


Fig. 10.13 Plasma distribution in a conventional IGBT (planar), in a trench-IGBT and in a trench fieldstop IGBT. Infineon figure [Las00b]

A comparison of the plasma density in the trench IGBT with the conventional NPT-IGBT according to Fig. 10.8 is given in Fig. 10.13 [Las00b]. The emitter is on the left-hand side of Fig. 10.13, the collector is on the right-hand side. The line for "planar" is the conventional IGBT, there the plasma distribution is as expected for a pnp-transistor with its emitter on the right-hand side. With the trench IGBT, we get an enhanced plasma concentration at the IGBT emitter side. It approximates the plasma distribution in a pin-diode.

In [Tak98], it was even shown that the trench IGBT voltage drop V_C decreases if some of the cells are not contacted. The plasma density increases in the same amount, as the part of contacted cells is reduced, as long as the number of not contacted cells is not too high. V_C is reduced, because the effect of plasma enhancement is much higher than an increased voltage drop across the channel.

10.6.2 The "Latch-up Free Cell Geometry"

Essential for modern IGBTs is to avoid latch-up of the parasitic thyristor at turn-off. This requirement must now be met under the condition of increased plasma density. The cell must have a structure which is optimized for this requirement. Figure 10.14 shows a detail of a trench cell. The hole current will flow mainly close to the electron current because of the condition of neutrality. Therefore, it flows close to the channel, and in its further way to the contact at the p^+-region it must flow underneath the n^+-source region. The n^+p-junction is biased in forward direction. If the voltage drop V_p that is generated by the hole current below the n^+-layer across the length L reaches the order of magnitude of the built-in voltage V_{bi} of this n^+p-junction, then the n^+-region will inject holes. The consequence will be latch-up of the parasitic thyristor, the turn-off capability is lost, the control of the device is lost and destruction of the IGBT will follow.

Using the depictive representation in Fig. 10.14, this voltage drop V_p can be simplified according to [Ogu04] with

$$V_p = \int_0^L R \cdot j_p \cdot y \cdot dy = \frac{1}{2} \cdot R \cdot j_p \cdot L^2 \qquad (10.18)$$

where R is the sheet resistivity of the p-layer below the n^+-source in $\Omega/\square$ and L is the length of the source layer. To keep V_p below the built-in voltage (≈ 0.7 V at 25 °C and decreasing with temperature) even at very high current density, especially L must be small, but also ρ must be kept low. As one can recognize in Fig. 10.12, L is designed very short in modern trench IGBTs. A highly doped p^+-layer spreads as far as possible in direction of the trench below the source layer. With these design measures, destructive latch-up can be avoided even at high current density.

This design measure, termed to be a "latch-up free cell geometry" [Las03] is realized by a highly conductive p^+ well adjusted on a submicron scale (Fig. 10.14) concerning the distance of this p^+ well to the trench sidewall. This layer forms a

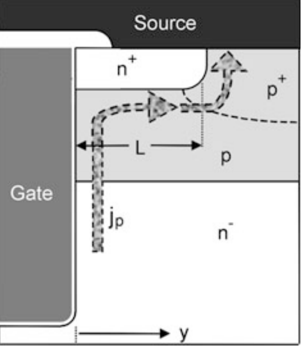


Fig. 10.14 Detail of an IGBT trench cell

resistance R_S (see equivalent circuit Fig. 10.2b). If R_S is low enough, the parasitic npn bipolar transistor is effectively suppressed, and latch-up of the thyristor, containing both the npn- and pnp bipolar transistors, will not occur. The discussed measure is not restricted to the trench cell. The same measure is possible for a planar structure.

We will come back on this later in Chap. 13. In devices for high blocking capability, a mode of operation with dynamic avalanche occurs at turn-off, an additional hole current is generated and the current density may even be locally increased. IGBTs with well-designed cell structures overcome even these high stress conditions.

10.6.3 The Effect of the "Hole Barrier"

The possibility to increase the plasma density below the emitter cells is also given by the implementation of an additional n-doped layer. This was first shown by [Tai96] on the example of a trench IGBT, the structure was denominated as "Carrier Stored Trench Gate Bipolar Transistor" (CSTBT). The effect is explained in [Tai96] in the following way: at the n^-n^+-junction a diffusion potential of approximately 0.17 V is built up, it hinders the outflow of holes. This additional layer was designated as a hole barrier.

This n-doped layer below the p-well acts in the same way in a planar IGBT [Mor07, Rah06], as shown in Fig. 10.15. In the n-doped layer, the hole current is the minority carrier current. It decreases strongly before it enters the p-well. In Eq. (10.17), this reduces j_p and increases the emitter efficiency γ of the total top-side cell structure, in which we again summarize p-well, n-channel and n accumulation layer to an effective n-emitter. The consequence is an increased plasma density, as shown in Fig. 10.15, right-hand side.

The hole barrier hinders the outflow of holes. To hold the condition of neutrality, additional electrons are delivered by the n-channel. The density of plasma increases, see Fig. 10.15, right-hand side.

A disadvantage of this measure is that the increased doping below the blocking junction J_2 decreases the blocking capability. This must be compensated with a slightly increased thickness of the n-base of the IGBT, and this increases the forward voltage drop. A part of the achieved advantage is lost again, but this effect can be kept minimal [Lin06].

A combination of the hole barrier with the trench structure is done in the "Carrier Stored Trench Gate Bipolar Transistor" (CSTBT) of the manufacturer Mitsubishi (see Fig. 10.16). The n-layer as a hole barrier is arranged below the p-layer within the trench structure. Below the hole barrier, holes accumulate; to keep neutrality, additional electrons are delivered effectively by the channel. The density of plasma is increased locally.

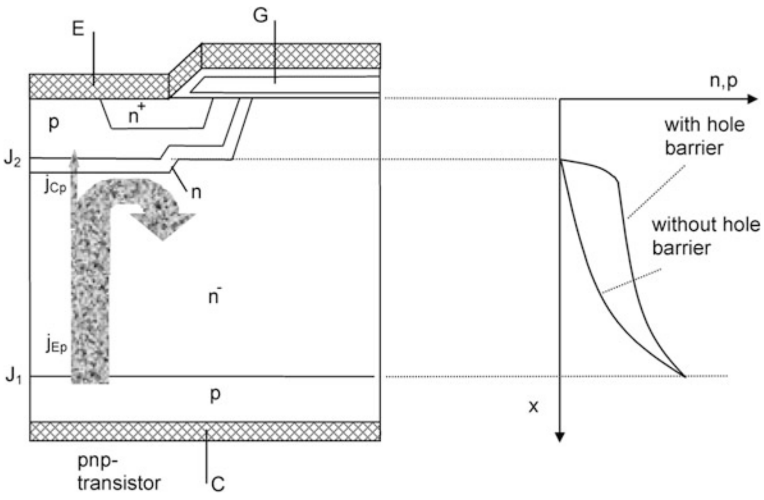


Fig. 10.15 Increase of the density of free charge carriers by a hole barrier

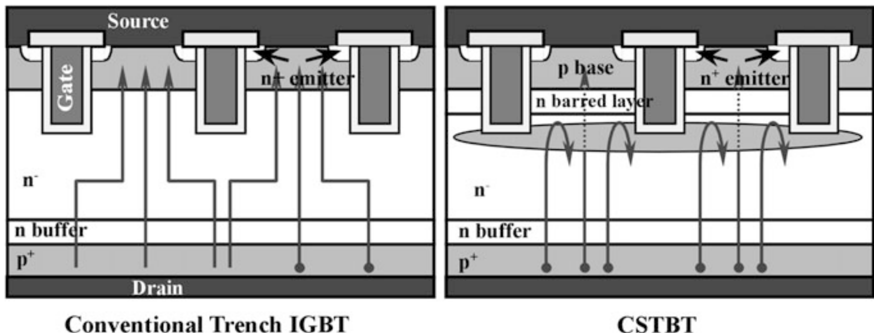


Fig. 10.16 Carrier Stored Trench Gate Bipolar Transistor (right-hand side) in comparison to a conventional trench IGBT (left-hand side). Pictures from Mitsubishi Electronics

The hole barrier can be combined with the before-described measures of increasing the distance between the cells and decreasing the lateral extension of the p-layers. In a trench IGBT, this can be done very easily by debarring a part of the cells. These cells are denominated as "plugged cells" [Yam02]. The polysilicon in these cells, which forms the gate area, is shorted to the emitter metallization, and the cell is prevented from building an n-channel. This measure additionally has the advantage that the current I_{Dsat} in the active area is reduced, and therewith the current at a short circuit event is reduced.

10.6.4 Collector Side Buffer Layers

Besides the described increasing of the plasma density, every modern IGBT additionally uses the effect of a reduced width of the n-base layer by designing a trapezoidal electric field instead of a triangular one. This is achieved with an n-layer with increased doping in front of the p-collector layer. The denomination differs for different manufacturers: "fieldstop", "soft punch through", "light punch through", etc., but it is more or less the same.

The plasma density for the "trench fieldstop" IGBT has already been shown in Fig. 10.13. In a trench-fieldstop IGBT, the width of the n-base is shortened compared to a trench IGBT. Figure 10.17 shows the new structure. In front of the collector layer, the doping density is increased (n-fieldstop). The space charge is trapezoidal if a voltage close to the specified blocking voltage is applied. This is sketched in Fig. 10.17, left-hand side. In reality, it is only a moderate trapeze, it is closer to a triangle and not almost rectangle, as one could conclude from Fig. 10.17.

From the viewpoint of the shape of the electric field, the fieldstop IGBT is similar to a PT-IGBT. The base width w_B is significantly reduced at the same blocking voltage in an IGBT with collector side buffer layer compared to an NPT-IGBT. The voltage V_{drift} that drops across the base layer is proportional to w_B^2 according to Eq. (10.13). Hence the voltage drop V_C can be reduced significantly with this design.

However, this is the only common feature with the PT-IGBT. The trench-fieldstop IGBT is not fabricated like the PT-IGBT on a p^+-substrate with an epitaxial n-layer, it is rather fabricated from a homogeneously n-doped wafer. The fabrication of a collector layer with exactly adjusted emitter efficiency is done similarly as with an NPT-IGBT [Las00b]. α_{pnp} is adjusted by the emitter efficiency; an emitter of low penetration depth and low doping is used. No charge carrier lifetime reduction is done. The mode of function of the trench-fieldstop IGBT is closer to the NPT-IGBT than to the PT-IGBT. As a result, the temperature dependency of V_C at is similar to that of the NPT-IGBT. The desired "positive temperature coefficient" remains.

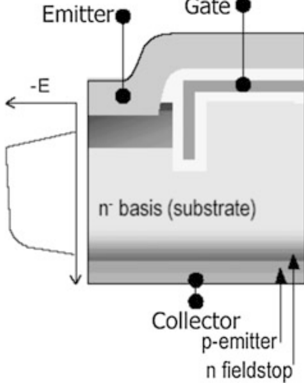


Fig. 10.17 Structure of the Infineon trench-fieldstop IGBT. Figure from Infineon

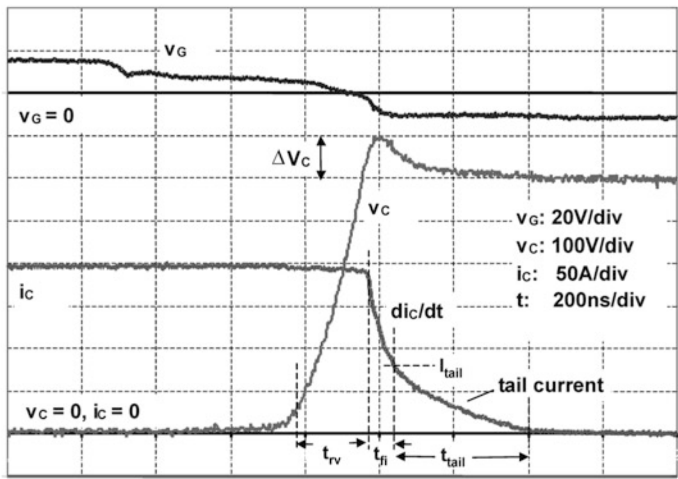


Fig. 10.18 Turn-off of the trench fieldstop IGBT (FF200R12KE3 from Infineon, 200 A 1200 V module). T = 125 °C, $R_{Goff} = 5\,\Omega$

Figure 10.18 shows the turn-off behavior of the trench-fieldstop IGBT. If one compares this figure with the turn-off behavior of the NPT-IGBT in Fig. 10.6, one can see a significantly shortened tail current. Additionally, the fall time of the current t_{fi} is increased and di_C/dt is decreased; this leads to a lower voltage peak ΔV_C.

The shortening of the tail current happens by the effect that a lower part of the stored plasma in the on-state remains close to the collector side at turn-off. The removal of the stored charge is done in a big amount during the voltage rise time t_{rv}. If the applied voltage V_{bat} is increased above the voltage of 600 V, as applied in Fig. 10.18, the tail current is shortened further, and finally it can disappear completely. Then the space charge spreads across the whole middle layer. If V_{bat} is increased further, a current snap-off similar to the behavior at reverse recovery of a snappy diode will occur at the end of the fall time t_{fi}. This current snap-off can lead to high overvoltage peaks.

10.7 IGBTs with Bidirectional Blocking Capability

For some applications of power electronics, e.g. the matrix converter, a both-side blocking device is necessary. The IGBT structure contains a p-layer forming a pn$^-$-junction at the collector side. The basic structure, as drawn in Fig. 10.1, has the capability to take an electric field at the bottom-side pn$^-$-junction as well as at the top-side n$^-$p-junction. It has similarities to the structure of a thyristor which has a blocking capability in both directions. However, the back-side pn$^-$-junction has no

defined junction termination. Such planar junction terminations for shallow pn-junctions need microstructures. The processes for the fabrication of junction terminations are only possible at the top side of the wafers, since semiconductor technology has been optimized for microstructures on one side of a wafer.

A possible solution is to lead the back-side pn-junction to the front side of the wafer. This can be done by the diffusion of a deep p-layer which reaches through the whole wafer; this technology is called "diffusion isolation" [Tai04]. A schematic drawing of an IGBT with this deep edge diffusion is shown in Fig. 10.19. With a deep diffusion in the area where the wafer is cut in the last production step, the collector-side pn^--junction is connected to the front side of the wafer. Now it is possible to apply a junction termination for both directions. This is done in Fig. 10.19 for the forward direction with potential rings, as known from Fig. 4.20. The junction termination for the forward direction ends at the channel stopper. As junction termination for the reverse direction, a field plate structure is applied (see Fig. 10.19).

The device blocks the voltage in both directions like a thyristor. The bidirectional blocking IGBT must be dimensioned as NPT-type. A buffer layer in front of the p-collector would reduce or even eliminate the blocking capability in reverse direction. Therefore, a minimal thickness w_B of the base layer for a triangular field shape is required. In [Nai04], this is given with 200 μm for a 1200 V reverse blocking IGBT; the electric field in the volume is very similar to that in Fig. 10.8.

Since a buffer layer cannot be applied, the reverse blocking IGBT cannot have a forward voltage drop V_C as low as other modern IGBTs. But V_C is surely lower than the one of a serially connected IGBT and diode.

In the intended matrix converter applications, the reverse blocking IGBT works as freewheeling diode at some switching events; it is passively turned-off like a diode. Because of the high carrier lifetime in IGBTs, a high reverse recovery peak I_{RRM} and a high reverse recovery charge occur. In one concept [Nai04], electron irradiation was used and a reduction of I_{RRM} of about 10% was achieved. The shorter carrier lifetime increases the on-state losses; a trade-off must be made.

Further work is done to improve the reverse blocking IGBT. The deep diffusion on the right-hand side of Fig. 10.19 has, however, also the consequence of lateral diffusion which will be approximately spread by 0.8× of the diffusion depth to the

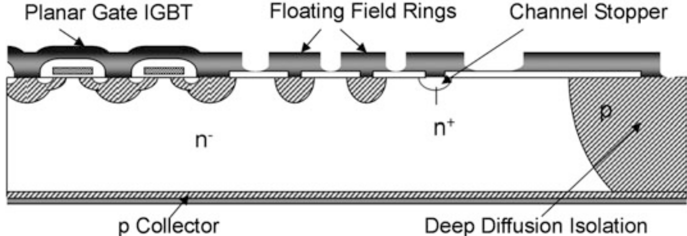


Fig. 10.19 Symmetrically blocking IGBT with edge diffusion. Figure from [Ara05] © 2005 EPE

side. With a wafer thickness in the range >100 µm, this deep p-layer will also be very wide. This leads to a loss of area for the region which takes the current, a loss of active area, and a high share of area for the junction termination which cannot be used for current transport. Research and development work is done for improved structures, for example to replace the deep diffusion zones by deep trenches [Ara05].

10.8 Reverse Conducting IGBTs

If in a part of the collector area an n^+-layer is implemented, see Fig. 10.20, a diode is formed by this n^+-layer and the p-body. There are several advantages of the monolithic integration of the diode. Primarily, it is the reduction of the chip size. Thermal calculations in of the diode and the IGBT in a three-phase inverter application [Rut07] have led to the result that a great saving of the area is possible, while the RC-IGBT stays almost at the same size like the IGBT and the diode area can be saved.

The concept of an RC-IGBT in a productive volume was first realized with an optimization for lamp ballast applications [Gri03]. This application needs 600 V blocking voltage, a current in the range of 1 A, and there is no hard commutation of the diode. These devices could be realized successfully without optimization of the diode.

To satisfy the requirements of hard switching applications, the reverse recovery behavior of the integrated diode has to be optimized. For this diode, the n-doped regions on the backside act as a cathode emitter (see Fig. 10.20), while the p-body of the IGBT and the highly p-doped anti-latch-up region near the front side act as an anode emitter of a freewheeling diode integrated into the chip in this way.

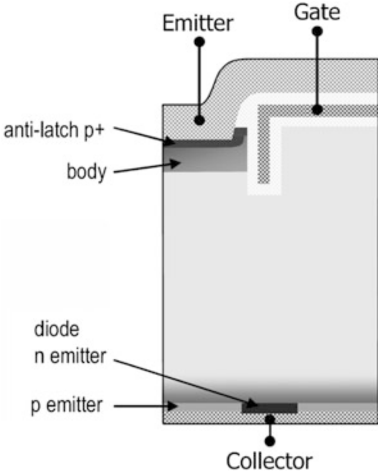


Fig. 10.20 Reverse conducting IGBT with trench gate. Figure from [Rut07] © 2007 IEEE

Unfortunately, such a diode with a highly doped p-emitter has a plasma distribution as shown in Fig. 10.21. It is very high at the anode side and lower at the cathode side. This will lead to a high reverse recovery peak and snappy reverse recovery behavior, as discussed in detail in Chap. 5. The internal plasma distribution as it must be for a soft recovery diode is shown for comparison in Fig. 10.21, too.

To reduce the plasma density at the cathode side, different methods have been tested. Irradiation of He^{2+}-ions is applied for the reverse conducting IGBT in [Tai04b]. The reverse recovery current peak I_{RRM} was reduced close to the value of a commercial freewheeling diode. But irradiation technologies can have a disadvantage. In [Rut07], it is reported that particle irradiation causes also undesirable trapped charges in the gate oxide and the silicon-gate oxide interface. These charges give rise to a drop of the threshold voltage and a broader parameter distribution. Therefore, the way used in [Rut07] is to reduce the efficiency of the anti-latch p^+-emitter by reducing the implantation dose. The advantage of this method is that there is no effect on the V_C. However, a reduction of this doping dose is limited by the overcurrent turn-off robustness (latch-up robustness). It was found that the potential for a reduction of Q_{RR} by decreasing the p^+-emitter dose at the same robustness as today's devices is between 15 and 25% [Rut07]. With platinum diffusion, Q_{RR} was reduced by 60%, and the increase in V_C of the IGBT was only 0.1 V [Rut07].

The measures to improve the IGBT and to improve the freewheeling diode contradict each other in some aspects. However it is expected that reverse conducting IGBTs will become available also for switching with inductive load, since a reduction of costs in the power electronic system is possible by saving half of the number of power dies. These solutions might be successful in applications in which the requirements to the diode reverse recovery are medium, e.g. for low current

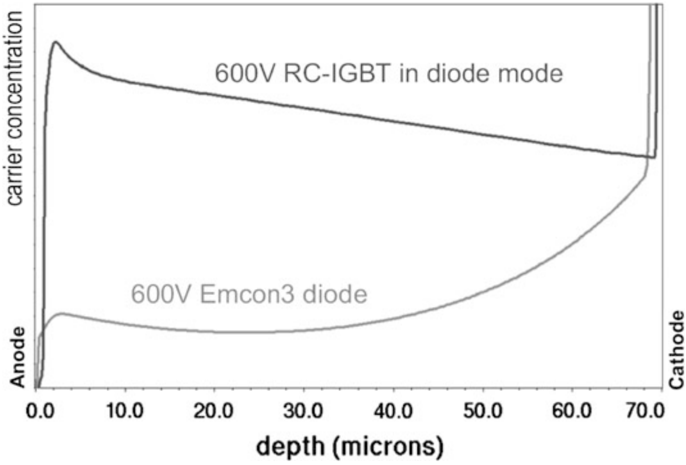


Fig. 10.21 Carrier concentration profile in a 600 V EMCON3 compared to a 600 V RC-IGBT in diode mode. Figure from [Rut07] © 2007 IEEE

applications up to 1 kW, since the energy stored in the parasitic inductance is small at low current. There is a wide field of such applications, such as inverters in air conditioners, refrigerators, and others.

A further interesting concept for a reverse conducting IGBT rated 3300 V was presented in [Rah08]. The requirements to the freewheeling diode are highest in the intended high power motor drive application. The new structure is shown in Fig. 10.22.

This structure uses the fact that there are measures which are not conflicting in their effect on the plasma distribution as well in the IGBT as in the diode. An n-layer below the p-well is used, as it was named as “hole barrier” in Fig. 10.15. It decreases the injected hole current, for the effective emitter of the IGBT it acts in an enhanced emitter efficiency. For the diode, it acts as reduced p-emitter efficiency.

On the collector side, the n^+-layer acts as a kind of anode short, known from GTO-thyristors, it will reduce the IGBT collector side emitter efficiency. Doping of the collector side p-layer, as well as width and doping of the n^+-layer must be carefully adjusted to fulfill the requirements for the diode as well as for the IGBT. For the diode, additional a local lifetime control region below the p-well is applied, as known from the CAL-diode in Chap. 5. The generation of recombination centers is executed with masked irradiation of light ions. The reverse recovery of the diode can further be improved by opening the channel. Injected electrons by the channel in the diode conducting mode will reduce the p-emitter efficiency, as it is described on hand of the MOS controlled diode (see Sect. 5.7).

Investigations in [Rah08] show that the IGBT as well as the diode could even profit in some aspects due to the integration. For the application it is a high advantage. The effective IGBT area is increased by 50% and the effective diode area even by 200%. A potential for rating 50% higher current for IGBT in modules with the same area is given, however only a part of it can be used in practice because of the thermal resistance of the housing sets a limit.

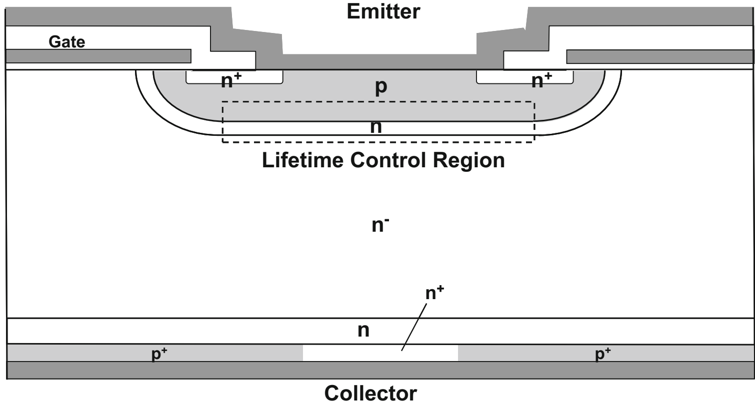


Fig. 10.22 3.3 kV RC-IGBT with optimized integrated diode. Figure following [Rah08] © 2007 IEEE

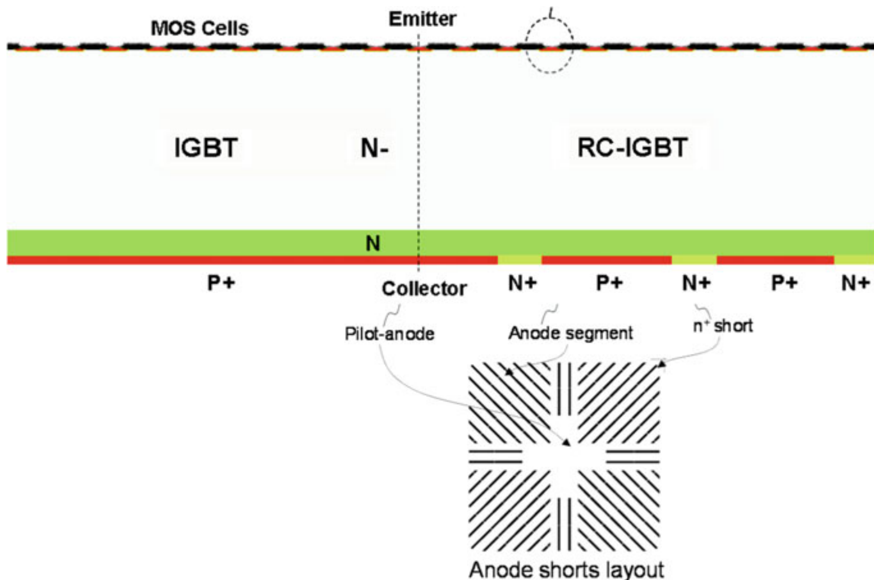


Fig. 10.23 Bimode-Insulated-Gate-Transistor BIGT. Fig. adapted from [Rah09] and [Sto15]

Realized RC IGBTs, however, showed a snap-back phenomenon in in the forward characteristics: First the n-region carry the current. As the p-layers starts to inject carriers, the forward voltage snaps back to a lower voltage. To avoid this unwanted behavior, the device is divided into an area with usual IGBT with no n-shorts (pilot IGBT) and an area with reverse conducting possibility. The device was denoted as Bimode-Insulated-Gate-Transistor BIGT [Rah09]. It is shown in Fig. 10.23. The optimal design of the n-shorts was found to be in the form of radial stripes [Sto11], an approximation to this is shown in Fig. 10.23 on the bottom side.

A 6.5 kV reverse conducting IGBT, where the freewheeling diode is integrated into the IGBT, was presented in [Wer14, Wer15]. The device is denoted as "Reverse Conducting IGBT with Diode Control" (RCDC) (Fig. 10.24).

In transistor operation the former area of the diodes is additionally used, during diode operation additionally the former transistor area. With this integration, a module with the same outline has now a rated current of 1000 A compared to 750 A of the former version. Especially interesting is the capability to strongly control the diode characteristics by the gate. In diode conduction mode, the gate is optimally loaded with −15 V. The flow of electrons is stopped, the p-emitter efficiency becomes high. The internal plasma is shown in Fig. 10.25. The voltage drop in conduction mode becomes very low, in the range of only a little bit above 2.5 V, see Fig. 10.26. Short before turn-off of the diode, a "desaturation pulse" of +15 V is applied during, for example, 15 µs. An n-channel injecting electrons is formed parallel to the p-doped diode emitter region forms. The p-emitter efficiency now becomes very low, the forward voltage is high (Fig. 10.26). The internal plasma shape of carriers becomes optimal for turn-off with low reverse recovery current

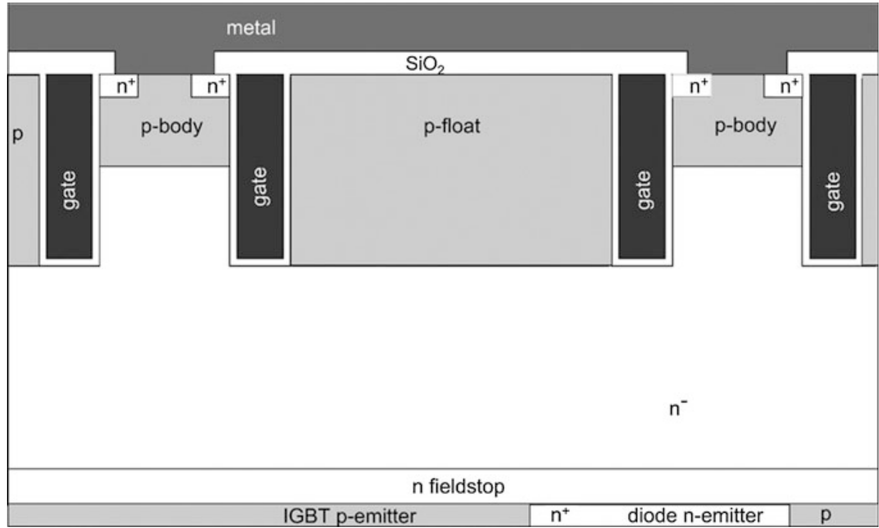


Fig. 10.24 Reverse conducting trench IGBT with integrated diode. Adapted from D. Werber, Infineon [Wer15]

maximum (Fig. 10.25). After a short dead time, the IGBT is turned-on and the diode is commutated in reverse direction. This leads to low turn-off losses of the diode (reduced by 40%) and low turn-on losses of the IGBT (reduced by 34%). The function of a field-controlled diode is shown in an impressive way.

A further advantageous aspect of the reverse conducting IGBT is expected in respect to reliability. In the typical motor drive application conditions, in one sinus half-wave the IGBT, in the other the diode is heated up. The absence of inactive periods can reduce the temperature ripple and will increase the lifetime of the module. For details, see Chap. 12. In 2017, Reverse conduction IGBTs are available for 4500 V [Dug17] and 1200 V [Tah16, Osa17]. The reverse conducting IGBT is still object of research and development.

10.9 The Potential of the IGBT

It has been shown that with modern IGBTs an improved distribution of the internal plasma was achieved. Therefore the IGBT applications could be extended to voltage ranges which were formerly only possible with thyristors. The development of devices with new structures, especially of a MOS-controlled thyristors and derivatives as IGBT-followers, has been placed back or suspended since the new structures were not necessary. The desired advantage has been possible with the IGBT to a great extent. IGBTs are commercially available up to 6.5 kV blocking

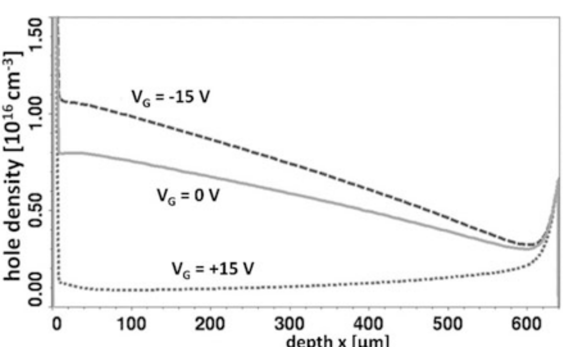


Fig. 10.25 Simulated carrier concentration of the RCDC in diode conduction mode along a vertical cut through a backside region with n-emitter for different gate voltages at 125 °C. Fig. from [Wer15], PCIM Europe 2015

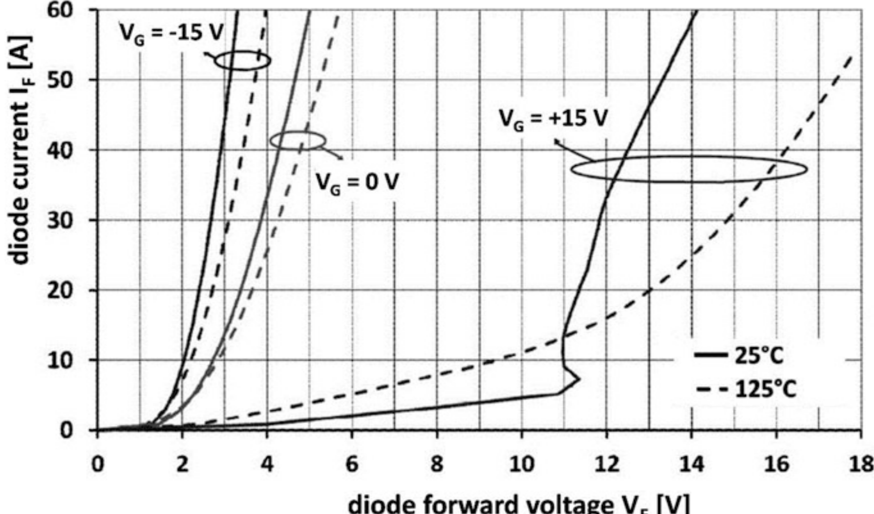


Fig. 10.26 Measured RCDC characteristics of diode conduction mode for gate voltages $V_G = -15$, 0, +15 V at room temperature and 125 °C. The nominal current per chip amounts to 28 A. Fig. adapted from [Wer14] and [Wer15]

voltage since more than ten years. Development works on IGBTs for the voltage range 8 or 10 kV were done, however these devices did not come to market.

However, further intensive work in research and development has to be done to decrease the conduction and switching losses of IGBTs. With IGBTs very low forward voltage drop and thereby very low conduction losses can be achieved [Sum12, Sum13]. A new structure as example is shown in Fig. 10.27. In the structure in Fig. 10.27a the current flowing from collector C to emitter E crosses the narrow Mesa between gate G and control gate CG. If a positive voltage is applied at CG, the hole current flow is suppressed. The structure approximates to an ideal n-emitter with j_p approximating 0 and $\gamma \rightarrow 1$. With this high emitter efficiency, the carrier density at the n-emitter side can be made very high, see Fig. 10.27b. The

on-state losses then will be low. With the signal at the control gate, the turn-off losses can be controlled. With a negative voltage at CG, the carrier density will be low and the turn-off losses are reduced. The achieved progress compared to the "IGBT-limit" calculated by Nakagawa [Nak06] is to be seen as a dot in Fig. 10.28.

The IGBT-limit was calculated in [Nak06] for an IGBT with ideal n-emitter of efficiency $\gamma_n = 1$. The lines for IGBTs in Fig. 10.28 are calculated as $R_{on} = V_C/I_C$ at the operation point of rated current, this means they are simplified for bipolar devices. In the publication [Sum13], for a 1.2 kV IGBT at 300 A/cm^2 a forward voltage drop of 1.6 V is achieved, this is even close to the forward voltage of a thyristor rated for this blocking voltage.

In HVDC applications, today IGBTs with rated voltage of 3.3 and 4.5 kV are used. 6.5 kV and more could be the future voltage rating to need less devices in series connection for the output voltage of the converter which is 500 kV or even more. A 6.5 kV IGBT of today has a V_C in the range 3.7 V. This is much higher than the value theoretically possible at the Nakagawa-limit.

An exemplifying drawing (T = 125 °C) for a state of the art 6.5 kV is shown in Fig. 10.29. A state of the art module (2016) is rated to 750 A and has a voltage drop of 3.7 V at rated current and 125 °C. It contains 24 IGBT-Chips with an approximated active area of 1.21 cm^2 each. This results in a $R_{on}{\cdot}A$ of 143 mΩcm^2 (see Fig. 10.28, "actual Si-IGBT"). At the "IGBT-Limit" according [Nak06], 7.5 mΩcm^2 would be possible. It is assumed that one can approximate to this limit up to a factor of two meaning 15 mΩcm^2 are achieved. This leads for same power loss density to a current conduction capability of 2300 A and a forward voltage drop of 1.45 V.

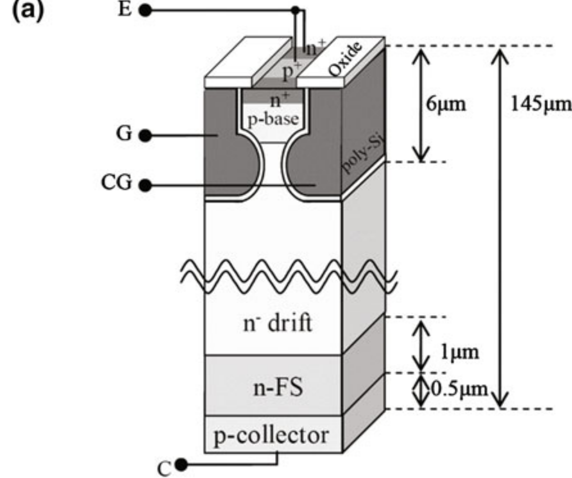
If this aim would be reached, the number of IGBT modules for a power converter with same output power could be reduced down to one third, and, compared to a solution with state of the art IGBTs, only 15% of conduction losses would be dissipated. However, an optimization must consider all requirements, especially overload capability and short-circuit withstand capability. Additionally, the switching losses have been neglected in this consideration. However, some high-power applications, for example the Modular Multilevel Converter for HVDC applications, use low switching frequencies. Therefore, such an aim is not unrealistic.

Additionally, IGBTs for the most frequently used 1200 V voltage range have a high capability of reduction of losses. It is achieved that IGBTs with a further reduction of the forward voltage down to less than 1.5 V will be possible. The losses in inverters for motor control can be reduced strongly.

Investigating Fig. 10.28 again, it is visible that for voltages above about 3 kV IGBTs can be superior to SiC unipolar devices regarding the forward voltage drop. SiC bipolar devices are possible, however due to the deep impurity levels of acceptors, incomplete ionization takes place for highly doped p-emitter regions. There is work in research for SiC IGBTs. However it will take time until they become mature devices.

IGBTs with additional integrated functions, as the reverse blocking IGBT and the reverse conducting IGBT, are expected to become commercially available. There is significant progress in reverse conducting IGBTs.

Fig. 10.27 IGBT structure with very low forward voltage drop. **a** Structure, **b** Flooding with free carriers. Figs from [Sak13] © 2013 The Institute of Electrical Engineers of Japan



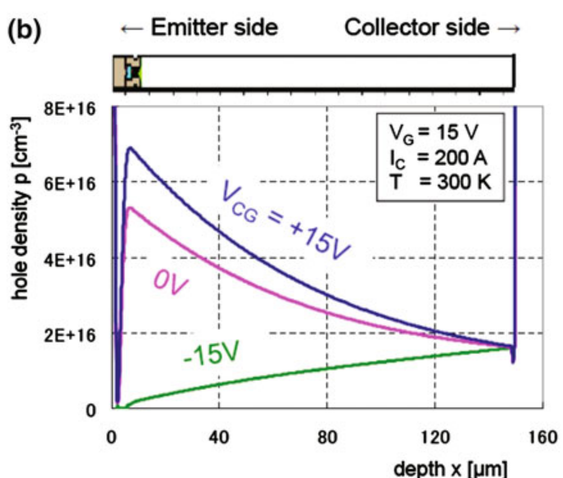
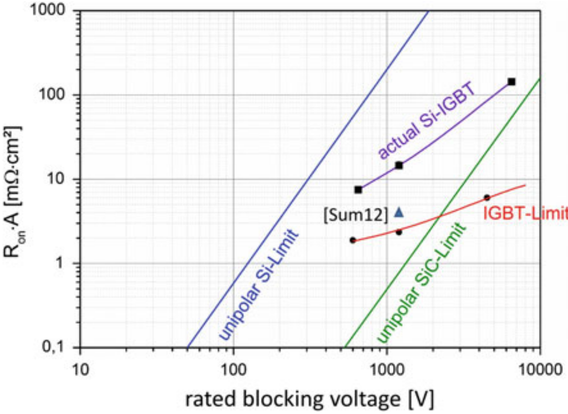


Fig. 10.28 Comparison of the on-resistance for the structure in Fig. 10.27 (marked as [Sum12]) with the on-resistance of unipolar devices (Si-Limit, SiC-Limit), with state-off-the-art IGBTs (actual Si-IGBT) and with theoretically possible IGBT-Limit, which is referred as Nakagawa-Limit. Figure inspired by [Nak06] and [Sum12]



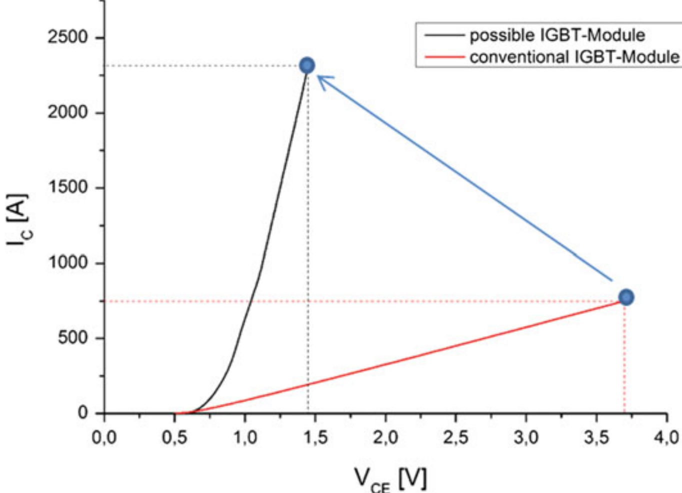


Fig. 10.29 Exemplifying drawing of the forward characteristics (T = 125 °C) for a state of the art 6.5 kV module and the possible progress at approximation to the Nakagawa-limit up to a factor of two

With the progress in IGBTs it is possible to control more and more power in a small device area. But with this, the dissipated losses per area are also increasing. These losses have to be extracted by the package and the challenges and requirements for the packaging technology are increasing. This is the subject of the next chapter.

References

[Ara05] Araki, T.: Integration of power devices – next tasks. In: Proceedings of the EPE, Dresden (2005)

[Bal82] Baliga, B.J., Adler, M.S., Grey, P.V., Love, R.P.: The insulated gate rectifier (IGR): a new power switching device. In Proceedings of the IEDM, pp. 264–267 (1982)

[Bal83] Baliga, B.J.: Fast-switching insulated gate transistors. IEEE Electron Device Lett. **4** (12), 452–454 (1983)

[Bec80] Becke, H.W., Wheatley, Jr C.F.: Power MOSFET with an anode region. United States Patent Nr. 4,364,073, 14 Dec 1982 (filed 25 Mar 1980)

[Dug17] Dugal, F., Baschnagel, A., Rahimo, M., Kopta, A.: The next generation 4500 V/3000A BIGT Stakpak modules. In: Proceedings PCIM Europe 2017, pp. 765–769 (2017)

[Gri03] Griebl, E., Hellmund, O., Herfurth, M., Hüsken, H., Pürschel, M.: LightMOS – IGBT with integrated diode for lamp ballast applications. In: PCIM 2003, p. 79ff (2003)

[Iwa17] Iwamuro, N., Laska, T.: IGBT history, state-of-the-art, and future prospects. IEEE Trans. El. Dev. **64**(3), 741–752 (2017)

[Kit93] Kitagawa, M., Omura, I., Hasegawa, S., Inoue, T., Nakagawa, A.: A 4500 V injection enhanced insulated gate bipolar transistor (IEGT) in a mode similar to a thyristor. In: IEEE IEDM Technical Digest, pp. 697–682 (1993)

[Las92] Laska, T., Miller, G., Niedermeyr, J.: A 2000 V non-punchtrough IGBT with high ruggedness. Solid State Electron. **35**(5), 681–685 (1992)

[Las00] Laska, T., Lorenz, L., Mauder, A.: The field stop IGBT concept with an optimized diode. In: Proceedings of the 41th PCIM, Nürnberg (2000)

[Las00b] Laska, T., Münzer, M., Pfirsch, F., Schaeffer, C., Schmidt, T.: The Field Stop IGBT (FS IGBT) – a new power device concept with a great improvement potential. In: Proceedings of the ISPSD, Toulouse (2000)

[Las03] Laska, T., et al.: Short circuit properties of trench/field stop IGBTs design aspects for a superior robustness. In: Proceeding 15th ISPSD, pp. 152–155, Cambridge (2003)

[Lin06] Linder, S.: Power semiconductors. EPFL Press, Lausanne, Switzerland (2006)

[Mil89] Miller, G., Sack, J.: A new concept for a non punch through IGBT with MOSFET like switching characteristics. In: Proceedings of the PESC' 89, vol. 1, pp. 21–25 (1989)

[Mor07] Mori, M., et al.: A planar-gate high-conductivity IGBT (HiGT) with hole-barrier layer. IEEE Trans. El. Dev. **54**(6), 1515 (2007)

[Nai04] Naito, T., Takei, M., Nemoto, M., Hayashi, T., Ueno, K.: 1200 V reverse blocking IGBT with low loss for matrix converter. In: Proceedings of the ISPSD '04, pp. 125–128 (2004)

[Nak84] Nakagawa, A., Ohashi, H., Kurata, M., Yamaguchi, H., Watanabe, K.: Non-latch-up 1200 V 75A bipolar-mode MOSFET with large ASO. In: Proceeding IEEE International Electron Devices Meeting, Dec 1984, pp. 860–861

[Nak85] Nakawaga, A., Ohashi, H.: 600–1200 V bipolar mode MOSFETS with high-current capability. IEEE-EDL **6**(7), 378–380 (1985)

[Nak06] Nakagawa, A.: Theoretical investigation of silicon limit characteristics of IGBT. In: Proceedings of the ISPSD, Neapel (2006)

[Net99] Netzel, M.: Analyse, Entwurf und Optimierung von diskreten vertikalen IGBT-Strukturen, Dissertation. Isle-Verlag, Ilmenau (1999)

[Nic00] Nicolai, U., Reimann, T., Petzoldt, J., Lutz, J.: Application Manual Power modules, ISLE Verlag (2000)

[Ogu04] Ogura, T., Ninomiya, H., Sugiyama, K., Inoue, T.: 4.5 kV injection enhanced gate transistors (IEGTs) with high turn-off ruggedness. IEEE Trans. Electron Devices **51**, 636–641 (2004)

[Omu97] Omura, I., Ogura, T., Sugiyama, K., Ohashi, H: Carrier injection enhancement effect of high voltage MOS-devices – device physics and design concept. In: Proceedings of the ISPSD, Weimar (1997)

[Osa17] Osawa, A., Higuchi, K., Kiamura, A., Inoue, D., Takamiya, Y., Yoshida, S., Gohara, H., Otsuki, M.: The highest power density IGBT module in the world for xEV power train. Proc. PCIM Europe **2017**, 1761–1766 (2017)

[Plu80] Plumer, J.D., Scharf, B.W.: Insulated-gate planar thyristors: I-Structure and basic operation. IEEE Trans. Electron Devices **27**(2), 380–387 (1980)

[Rah02] Rahimo, M., Kopta, A., Eicher, S., Kaminski, N., Bauer, F., Schlapbach, U., Linder, S.: Extending the boundary limits of high voltage IGBTs and diodes to above 8 kV. In: Proceeding ISPSD 2002, Santa Fe, USA, pp. 41–44

[Rah06] Rahimo, M., Kopta, A., Linder, S.: Novel enhanced–planar IGBT technology rated up to 6.5 kv for lower losses and higher SOA capability. In: Proceeeding ISPSD 2006, Naples, pp. 33–36 (2006)

[Rah08] Rahimo, M., Schlapbach, U., Kopta, A., Vobecky, J., Schneider, D., Baschnagel, A.: A high current 3300 v module employing reverse conducting IGBTs setting a new benchmark in output power capability. In: Proceeding ISPSD, Orlando, FL (2008)

[Rah09] Rahimo, M., Kopta, A., Schlapbach, U., Vobecky, J., Schnell, R., Klaka, S.: The Bi-mode insulated gate transistor (BiGT) A potential technology for higher power applications. In: Proceeding ISPSD09, p. 283 (2009)

[Rog88] Rogne, T., Ringheim, N.A., Odegard, B., Eskedal, J., Undeland, T.M.: Short-circuit capability of IGBT (COMFET) transistors. IEEE Ind. Appl. Soc. Annu. Meet. **1**, 615–619 (1988)

[Rus83] Russell, J.P., Goodman, A.M., Goodman, L.A., Neilson, J.M.: The COMFET – a new high conductance MOS-gated device. IEEE Electron Device Lett. **4**(3), 63–65 (1983)

[Rut07] Rüthing, H., Hille, F., Niedernostheide, F.J., Schulze, H.J., Brunner, B.: 600 V reverse conducting (RC-) IGBT for drives applications in ultra-thin wafer technology. In: 19th International Symposium on Power Semiconductor Devices and IC's, ISPSD '07, pp. 89–92 (2007)

[Sak13] Sakane, H., Sumitomo, M., Arakawa, K., Higuchi, Y., Asai, J.: Injection Control Technique for High Speed Switching with a double gate PNM-IGBT. In: The Papers of Joint Technical Meeting on Electron Devices and Semiconductor Power Converter, IEE Japan, Paper No. EDD-13-046 SPC-13-108 (2013)

[Scf78] Scharf, B.W., Plummer, J.D.: A MOS-controlled triac device. In: Proceeding IEEE International Solid-State Circuits Conference, pp. 222–223 (1978)

[She15] Shenai, K.: The invention and demonstration of the IGBT. IEEE Power Electron. Mag. June 2015

[Sto11] Storasta, L., et al.: The radial layout design concept for the bi-mode insulated gate transistor. In: ISPSD, San Diego, USA (2011)

[Sto15] Storasta, L., Rahimo, M., Häfner, J., Dugal, F., Tsyplakov, E., Callavik, M.: Optimized power semiconductors for the power electronics based HVDC breaker application. In: Proceedings PCIM 2015, Nuremberg (2015)

[Sum12] Sumitomo, M., et al.: Low loss IGBT with partially narrow mesa structure (PNM-IGBT). In: Proceedings ISPSD (2012)

[Sum13] Sumitomo, M., et al.: Injection control technique for high speed switching with a double gate PNM-IGBT. In: Proceedings ISPSD, Brügge (2013)

[Tah16] Takahashi, M., Yoshida, S., Tamenori, A., Kobayashi, Y., Ikawa, O.: Extended power rating of 1200 V IGBT module with 7G-RC-IGBT chip technologies. In: Proceedings PCIM Europe 2017, pp. 438–444 (2016)

[Tai96] Takahashi, H., Haruguchi, H., Hagino, H., Yamada, T.: Carrier stored trench-gate bipolar transistor (CSTBT) – a novel power device for high voltage application. In: ISPSD '96 Proceedings 8th International Symposium on Power Semiconductor Devices and ICs 20–23 May 1996, pp. 349–352, 1133 (1996)

[Tai04] Takahashi, H., Kaneda, M., Minato, T.: 1200 V class reverse blocking IGBT (RB-IGBT) for AC matrix converter. In: Proceedings of the 16th ISPSD, pp. 121–124 (2004)

[Tai04b] Takahashi, H., Yamamoto, A., Aono, S., Minato, T.: 1200 V reverse conducting IGBT. In: Proceedings of the 16th ISPSD, pp. 133–36 (2004)

[Tak98] Takeda, T., Kuwahara, M., Kamata, S., Tsunoda, T., Imamura, K., Nakao, S.: 1200 V trench gate NPT-IGBT (IEGT) with excellent low on-state voltage. In: Proceedings of the ISPSD, Kyoto (1998)

[Tih88] Tihanyi, J.: "MOS-Leistungsschalter", ETG-Fachtagung Bad Nauheim, 4.-5. Mai 1988, Fachbericht Nr. 23, VDE-Verlag, S. 71–78 (1988)

[Wer14] Werber, D., Pfirsch, F., Gutt, T., Komarnitskyy, V., Schaeffer, C., Hunger, T., Domes, D.: 6.5 kV RCDC for increased power density in IGBT-modules. In: Proceedings of the 26th ISPSD, Waikoloa, pp. 35–38 (2014)

[Wer15] Werber, D.: A 1000A 6.5 kV power module enabled by reverse-conducting trench-IGBT-technology. In: Proceedings PCIM 2015, Nuremberg (2015)

[Yam02] Yamada, J., Yu, Y., Donlon, J.F., Motto, E.R.: New MEGA POWER DUAL™ IGBT module with advanced 1200 V CSTBT chip. In: Record of the 37th IAS Annual Meeting Conference, vol. 3, pp. 2159–2164 (2002)

Chapter 11
Packaging of Power Devices

11.1 The Challenge of Packaging Technology

The operation of a power semiconductor device produces dissipation losses. The order of magnitude of these losses shall be estimated in the following example:

IGBT module BSM50GB120DLC (Infineon) mounted on an air-cooled heat sink
Operation conditions: I_C = 50 A, V_{bat} = 600 V, R_G = 15 Ω, T_j = 125 °C,
f = 5 kHz, duty cycle $d = t_{on}/(t_{on} + t_{off}) = 0.5$
The following parameters can be extracted from the data sheet:
Forward voltage drop: V_C = 2.4 V
Turn-on energy loss per pulse: E_{on} = 6.4 mW s
Turn-off energy loss per pulse: E_{off} = 6.2 mW s

For details on E_{on} see Figs. 5.20 and 5.21. A simplified calculation can be done with Eq. (10.4), a more exact determination is done with the oscilloscope, see Eq. (9.33). For details on E_{off} see Fig. 10.6, for simplified calculation Eq. (10.6) is useful.

Losses created by the leakage current can usually be neglected in modern IGBT and MOSFET applications. Therefore total power dissipated in the device is the sum of on-state and switching losses:

$$P_V = P_{cond} + P_{on} + P_{off} = d \cdot I_F \cdot V_C + f \cdot E_{on} + f \cdot E_{off} \tag{11.1}$$

This amounts to 123 W for the given example. These losses are marginal compared to the controlled power of approximately 30 kW. To calculate the efficiency an

© Springer International Publishing AG 2018
J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8_11

additional free-wheeling diode has to be taken into account; for most applications a half-bridge configuration of two switches in series must be considered. Nevertheless, the efficiency of the power control circuit is in the range of 98%.

However, 123 W of power losses have to be extracted from an IGBT switch with an area of about 1 cm^2 which requires a heat flux density of 123 W/cm^2 or 1.23 MW/m^2. The heat flux density can even amount to the 2–3 fold value for an assembly on a water-cooled heat sink and with maximum utilization of the module capability. Figure 11.1 relates this power loss to that of other heat sources.

The heat flux density of a power semiconductor chip exceeds that of a stovetop of a conventional kitchen stove by more than one order of magnitude and outranges a Pentium 4 microprocessor. Therefore, a power module has to provide a high thermal conductivity. Additionally, a power device package has to meet a number of requirements:

- High reliability, i.e. a long lifetime in application and therefore a high durability under alternating load conditions (power cycling stability)
- High electrical conductivity of the components to achieve low undesirable (parasitic) electrical properties (parasitic resistance, capacity and inductivity)
- For power modules additional electrical insulation between switches and between circuit and heat sink.

The solution to this problem is by no means trivial and it is today one of the most exciting challenges for engineers. Power modules are the prevalent types of packages in power electronic applications and they will be discussed in detail in the following chapters.

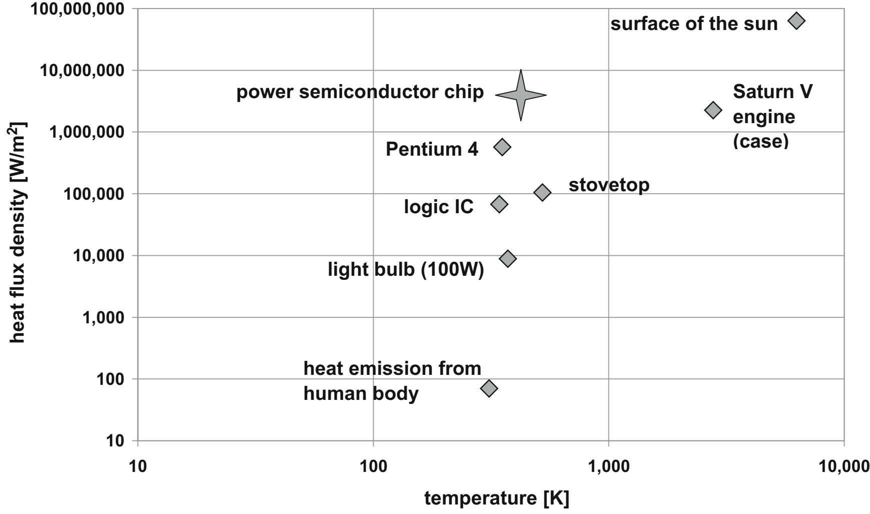


Fig. 11.1 Heat flux density of different heat sources, inspired by Dr. W. Tursky, Semikron

11.2 Package Types

A substantial criterion for the selection of an appropriate package type is the power range of the semiconductor device. A survey of power ranges is given in Fig. 11.2.

Discrete packages are prevailing in the range of small power. These packages are soldered to a laminated 'printed circuit board' (PCB) for application. Since the generated power losses are relatively small, the requirements for heat dissipation are unincisive. These packages are mostly designed without internal insulation with the consequence, that only a single switch can be integrated in one package. The most common package of this type is the 'transistor outline' (TO) package.

The discrete design has to fulfill the following functions:

– Conduction of load current and control signals
– Dissipation of heat
– Protection against environmental influences

Capsules also belong to the discrete packages. They are applied for the high power end of the range that is not yet reached by power modules. Capsules are not equipped with an internal insulation. They can be cooled from two sides. A power chip can have the size of a whole wafer in the peak performance range. Therefore, the circular footprint of the capsule is the ideal package for circular chips. Because of its shape, this package is also known as "hockey puk".

A thyristor from Mitsubishi for 1.5 kA with 12 kV blocking voltage is packaged in a capsule. Thyristors in capsules from Infineon are specified for 3 kA with

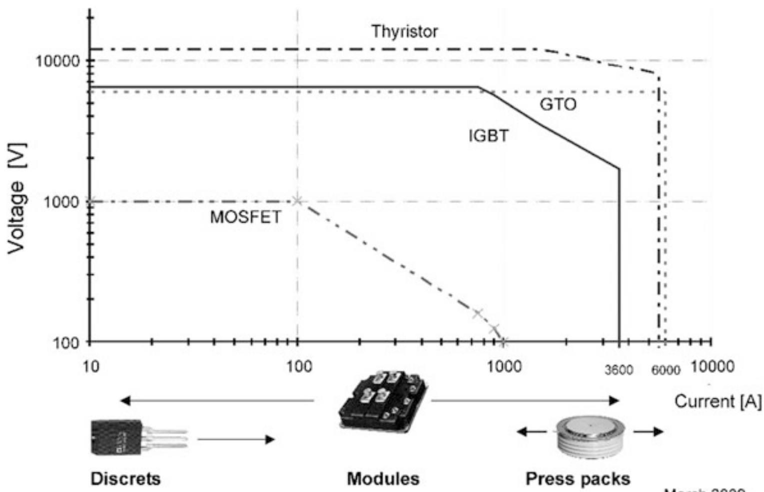


Fig. 11.2 Power range of modern semiconductor devices (2017) together with the predominant package type

8.2 kV, and recently developed thyristors for HVDC-applications are rated for 5.6 kA with 8 kV blocking voltage. The ‘chip’ in these packages consists of a complete 6 inch wafer with a diameter of approximately 150 mm.

Mitsubishi offers a gate turn-off thyristor (GTO) in a capsule with a chip fabricated from a single 6 inch wafer (150 mm) which is specified for 6 kA with 6 kV blocking voltage.

In contrast to discrete packages power semiconductor modules are characterized by:

- an insulated architecture in which the components of the electrical circuit are dielectrically insulated from the heat dissipating mounting surface,
- several single functions (phase leg circuit), often with paralleling of chips

Power semiconductor modules are dominating in the range of more than 10 A for blocking voltages of 1200 V and above. They are characterized by the integration of multiple functions (for example converter-inverter-brake topologies, CIB) in the lower power range. In the high power area, Infineon offers a module with 6.5 kV IGBTs and the associated freewheeling diodes with a continuous maximum current of 900 A. For 1200 V blocking voltage, Infineon produces a module specified for 3.6 kA continuous current which contains 24 IGBT chips in parallel and 12 freewheeling diodes in parallel. These examples show that modules have penetrated deep into the high power range which was formerly dominated by capsules. This trend will continue.

11.2.1 Capsules

Figure 11.3 displays the internal construction of a capsule in a simplified schematic view. The silicon device (e.g. a thyristor) is mounted between two metal discs in order to homogenize the pressure and to avoid pressure peaks. Molybdenum is the ideal material for this purpose because of its great hardness and its well adapted coefficient of thermal expansion. In the architecture shown in Fig. 11.3 the silicon device is rigidly coupled to one of the molybdenum discs on the anode side and pressed to the second molybdenum disc on the cathode side. Alignment features to center the chip inside the package are not shown in Fig. 11.3 for clarity, neither is the gate contact spring displayed, which is guided by a slot in the cathode compression piece to the center of the silicon device. The package is hermetically sealed by welding together the two metal latch rings.

A complete electrical and the thermal contact will only be established by the application of a defined pressure to the package, which is typically in the range of 10–20 N/mm^2.

The interconnection between the silicon device and the molybdenum discs can vary for different package sizes and manufacturers. For small chip diameters up to 5 cm, solder interfaces are feasible. But care must be taken to select a solder

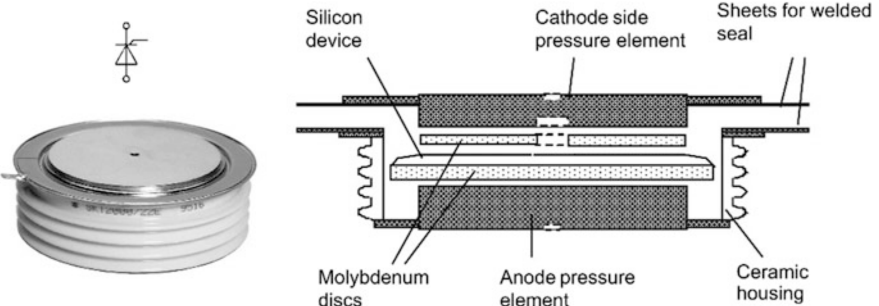


Fig. 11.3 Internal construction of a capsule (simplified)

material, which will show only little plastic creep under high pressure. For larger diameters of the device alloyed interfaces are generally preferred. Also designs without any rigid connection between Mo and Si are available, which allow a floating of the power device. A progressive technology for the interconnection of silicon and molybdenum is a diffusion sinter technique: The partners to be connected are equipped with a noble metal plating, a silver powder is applied to the connecting surfaces and a very reliable connection is established by sintering the interface layer at a high pressure and temperatures of approximately 250 °C [Kuh91].

Mostly conventional devices are packaged in capsules: Diodes, thyristors, GTOs and the GCTs derived from the GTO. The advantages of capsules are:

- Compact design with good relation between device surface area and package surface area
- Cooling of both device surfaces
- No wire bonds – wire bonds generally represent a reliability constrictive feature
- Few or no rigid interconnections between materials with different coefficients of thermal expansion

A high reliability can be expected from the last two factors. The disadvantages of capsules are:

- No dielectric insulation – the user has to provide for insulation in the application
- Higher effort in the mounting assembly – a defined uniaxial high pressure must be established and maintained

Due to its advantages the capsule package was also adapted as a package for IGBTs. But IGBTs are today produced only in a small chip size compared to thyristors. The reason is the high cell density of modern IGBT chips which would result in a yield problem caused by single cell defects with increasing chip size.

The largest commercially available IGBT has an area of 300 mm^2. Furthermore, the simplicity of paralleling fast switching IGBTs compared to the difficulty of paralleling slow switching thyristors, together with the thermal disadvantage of large area chips, does not produce a market pressure to develop larger area IGBTs. However, for the adaptation of the capsule package for IGBTs, the parallel arrangement of quadratic chips in a so called ‘presspack IGBT’ is a technological challenge.

An example for a presspack IGBT is illustrated in Fig. 11.4. The chips are assembled on a large molybdenum disc, each chip equipped with a collector side small Mo square. Alignment frames are positioning the chips relative to each other. Small Mo squares with cut-outs for the gate contact area are placed on the emitter contacts. The gate connection is implemented by springs, which are guided by another alignment structure. The upper pressure element has to transmit a uniform pressure to each of the chips below. To press each of the 21 paralleled IGBTs with an identical pressure requires maintaining very tight tolerances for every part of the package. For understanding of the system, electrical as well as thermal contact resistances have to be considered [Pol13a].

Integrated in the upper pressure element, a printed circuit board – carrying the gate resistors in ‘surface mounted device’ (SMD) technology – is installed. The complex construction of a presspack IGBT results in a considerable increased demand on the precise alignment of a multitude of parts and on the allowable part tolerances compared to a semiconductor module. Higher reliability in active power cycling compared to modules was expected, however not achieved in experimental tests [Tin15]. At high power load a deformation of the presspack resulting from internal temperature gradients occurred. It leads to inhomogeneous pressure distribution and even partially opening of contacts at outer positions [Pol13b]. The found failure modes were gate oxide damage and micro arcing [Tin15].

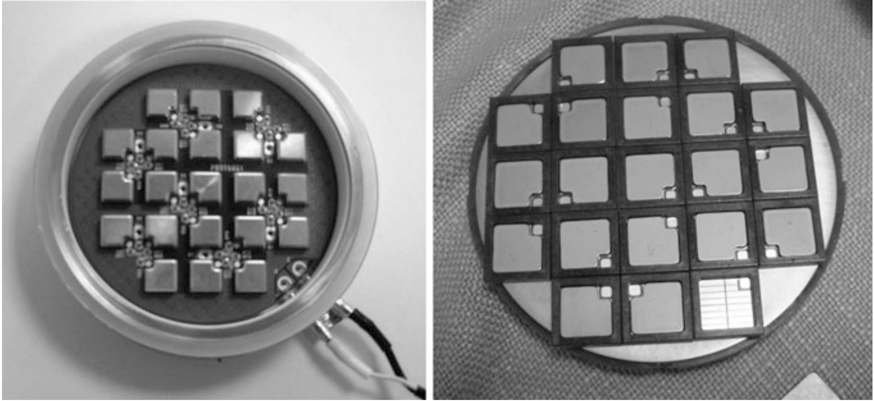


Fig. 11.4 Presspack IGBT: emitter pressure element (left), arrangement of the chips (right)

11.2.2 The TO-Family and Its Relatives

Discrete packages are also very common in the lower power range. Today, this field is dominated by the 'transistor outline' (TO) family. The principle design is shown in Fig. 11.5.

The TO package family comprises an extensive set of standardized package outlines, the most popular representatives are the TO-220 and the TO-247 package. In these standard packages, the power silicon chip is soldered directly to a solid copper base, which serves as mounting surface. Therefore, the package has no inherent electrical insulation. The contact leads or contacts legs are fixed by a 'transfer mold' housing. One of the leads is directly connected to the copper base, the others are connected to the load and control contact areas on the silicon chip by aluminum wire bonds (Fig. 11.5).

The difference in thermal expansion between the silicon chip and the copper base limits the reliability of this package. An improvement in this respect is the ISOPLUS package introduced by IXYS. As illustrated in Fig. 11.6, a ceramic substrate replaces the solid copper base, thus adapting a technology which is successfully applied in power modules. This design exhibits a number of advantages compared to the standard TO package:

- better adaptation of thermal expansion resulting in higher reliability
- internal insulation
- smaller parasitic capacity compared to a standard TO package mounted on a heat sink with an external insulation polyimide foil (for details refer to Sect. 11.5)

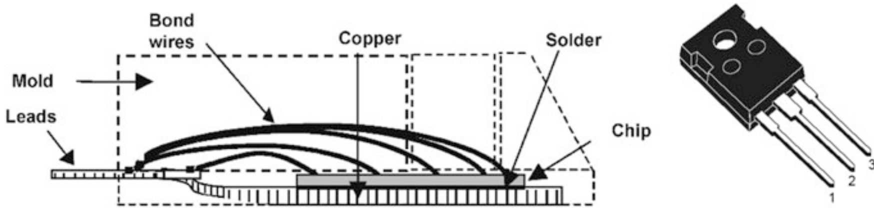


Fig. 11.5 TO package, principal design

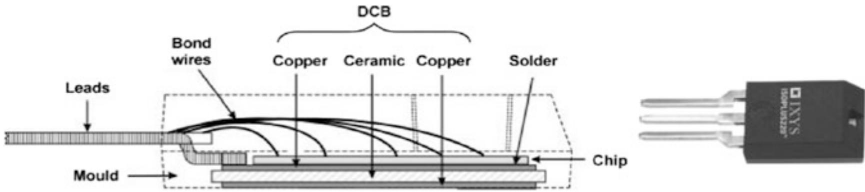


Fig. 11.6 ISOPLUS package with TO outline, but with insulated base

At the first glance, the smaller thermal conductivity of the ceramic layer with respect to copper appears as a serious drawback. However in a system, where several discrete packages – which typically show different potentials at their copper base – are mounted on a single heat sink, the ceramic insulation system generally is superior to the insulation of standard TOs by externally applied electrically insulating foils.

The prevalent power device packaged in a TO housing is the MOSFET. For this device, a drastic reduction of the electrical on-state resistance R_{on} has been accomplished in the last years. As a consequence, a fundamental weakness of this package design is emerging: The TO package has a parasitic electrical resistance in the same order of magnitude as the on-state resistance of a modern MOSFET device!

The contact leads are a major limiting factor. Their electrical resistance can be calculated by

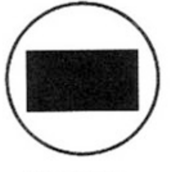
$$R_Z = \rho \cdot \frac{l}{A} \tag{11.2}$$

Considering a copper input lead and a copper output lead with a cross section of 0.5 mm^2 and a length of 5 mm each, the specific electrical resistance of copper ρ_{Cu} = 1.69 μΩ cm yields a total resistance of 0.34 mΩ. For a mean current of 50 A, the power loss

$$P_Z = R_Z \cdot I^2 \tag{11.3}$$

dissipated in these leads amounts to approximately 0.85 W. Since the contact leads are cooled only marginally, they are heated up by the ohmic losses to temperatures, which can get close to the melting temperature of the solder alloy applied for the PCB solder contact [Saw00]. This effect damages the solder contacts and reduces the reliability.

Since the through holes for PCB mounting are standardized and since insulation requirements demand to maintain minimal clearance distances between the leads, the lead cross section cannot by increased by simply implementing wider leads (Fig. 11.7). But it was possible to increase the cross section by improving the shape of the leads as shown in the right schematics in Fig. 11.7 and thus enhance the current capability of the TO package by 16%. This upgraded version of a TO-247 package is labeled as 'super-247' package by the manufacturer.



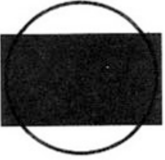


Fig. 11.7 Reduction of the electrical resistance of contact leads in a TO package [Saw00]

The evolution of PCB technology established the ‘surface mounted device’ (SMD) technology for component packages. This technology allows to assemble components on both sides of the PCB and facilitates multilayer PCBs with buried inner layers. Since the through hole mounting technique of classical TO packages is not compatible with this assembly process, a new generation of packages was developed as shown in Fig. 11.8. The standardized package outlines were equipped with contact leads for SMD mounting. For the ‘super’ version of this package, not only the contact leads are designed as short as possible – allowing more area for larger silicone devices – but also the wire bond connections were optimized. These improvements result in a reduction of the parasitic inductance of the package of 33% according to [Saw00].

A general weakness of TO packages and their SMD counterparts is the use of aluminum wire bonds, which contribute to the parasitic ohmic resistance of the package. Improvements are attempted by the implementation of thicker wires and/or by an increase of the number of wire bonds. Additionally, the parasitic inductance of the bond wires and the limited heat transport capability supported the search for better alternatives, e.g. by replacing the emitter bonds by copper metal sheets.

A revolutionary solution was introduced by the US American company International Rectifier, which completely eliminates the problematic contact leads, as well the bond wires. This ‘DirectFET’ package is displayed in Fig. 11.9. The same package is offered as “CanPack” from Infineon. The emitter and gate contacts of the silicon device are equipped with a solderable surface metallization. A so called ‘drain clip’ is attached to the drain contact of the device by a solder connection. This package is mounted on the PCB surface in a ‘flip chip’ fashion, were the SMD compatible solder connections for the gate, emitter und drain are established in a single reflow solder step.

Beside the low effort mounting procedure, the advantages of this package concept stem from the facts, that virtually no limitation of the current capability originates from contact leads and that parasitic inductance generated by bond wires is completely eliminated. Furthermore, a double-sided cooling of the package is possible, whereas the drain clip can dissipate significantly more heat than can be extracted through the PCB.

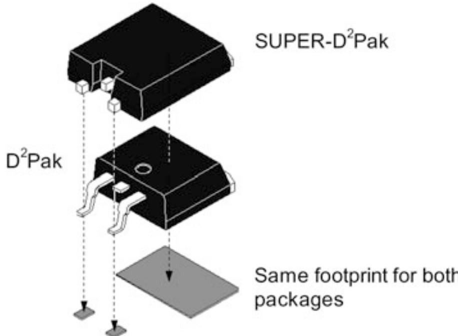


Fig. 11.8 For SMD technology optimized package design [Saw00]

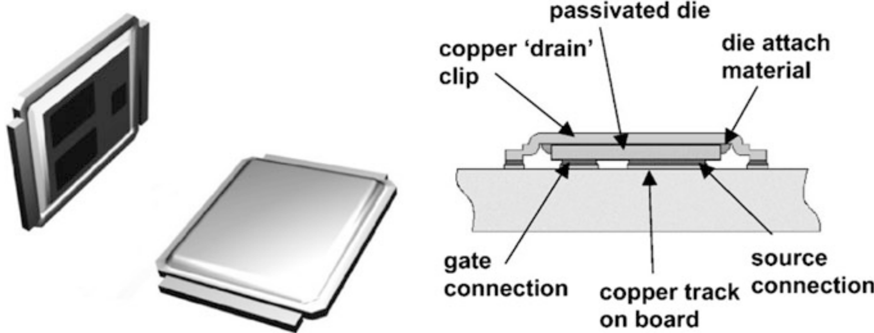


Fig. 11.9 DirectFET package [Saw01]

Nevertheless, no complete encapsulation is provided by this package, which leaves the sensitive silicon device unprotected against humidity and corrosive atmosphere influence. Nevertheless, a high humidity reverse bias test (see Chap. 12) showed positive results [Hof13]. Furthermore, the visual access to the solder interconnection beneath the package is nearly impossible, which impedes the quality control of the assembled PCB. The application and field experience with this new package design will show, if this concept will prevail.

However, the combination of the 'lead frame' construction with the 'transfer mold' technology, as was developed and optimized with the discrete TO packages, has lead to a powerful group of descendants: the transfer mold 'intelligent power module' (IPM) packages. In these packages, the advantages of both technologies were merged with the integration of various functions in a single package. Figure 11.10 shows an example of an IPM transfer mold package, containing a three phase inverter together with their driver ICs.

The internal structure of such a transfer mold IPM device as shown in Fig. 11.11 allows to comprehend the high potential of this packaging concept, which today is dominating the field of low power IPMs worldwide. Despite of all the limitations of

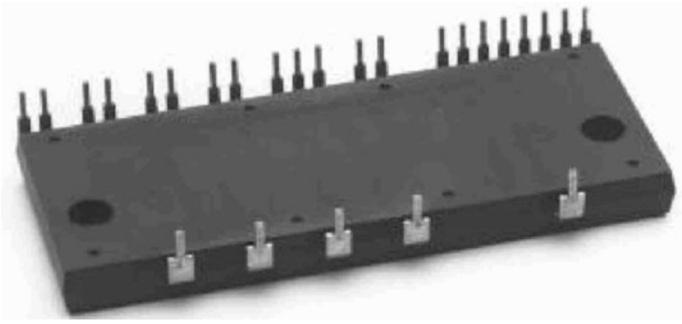


Fig. 11.10 Transfer mold DIP-IPM package from Mitsubishi

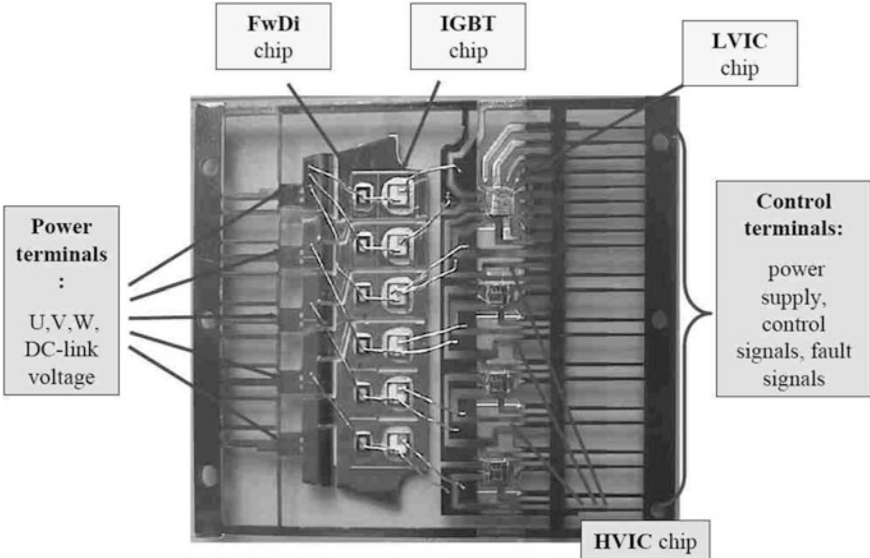


Fig. 11.11 Internal structure of the DIP-IPM package from Mitsubishi

this package design discussed before, the manufacturing process of these lead frame packages is highly optimized and very competitive. In production, the lead frames are connected to each other by the frame elements, forming a continuous band of such lead frames, ideal for automated assembly. After the chip soldering and wire bonding, the band is separated into single lead frames and subjected to a transfer mold process, which completely encapsulates the internal structure. Now the leads are fixed by the plastic encapsulation and the remaining supporting lead connections, which connect the ductile contact leads during the assembly, are stamped out.

Nowadays, more than 10 million of these transfer mold type IPM packages are produced every month, dominating the field of low power applications in the power semiconductor market.

11.2.3 Modules

As a result of the isolated construction, power modules provided substantial advantages in application. Soon after the first insulated power module was introduced by Semikron in 1975, this new architecture penetrated the market, even though the first design was rather complicated with a multitude of interfaces. Figure 11.12 shows the successor of this first power module, which today is still manufactured in large quantities. Shown here is the fifth generation design with an identical package outline as the first power module, but with an upgraded inner construction.

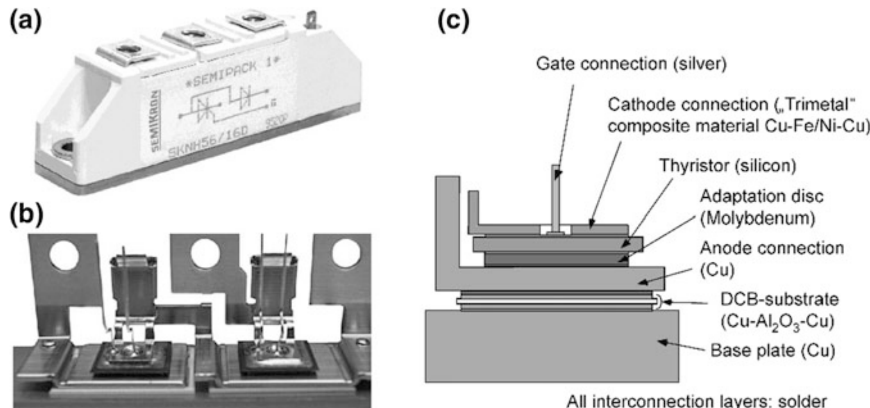
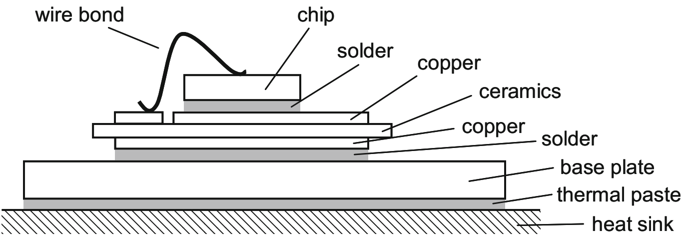


Fig. 11.12 Architecture of a classical thyristor power module: exterior view (**a**), inner construction (**b**) and cross section schematics showing the layer sequence (**c**)

The thyristor chip, which is equipped with solderable metallization on the anode, cathode and gate contact areas, is connected to the contact leads by solder interfaces. The cathode connector consists of a composite material with a coefficient of thermal expansion adapted to that of silicon. The anode contact of the silicon chip is joined to a molybdenum plate. This intermediate layer is required to accommodate the difference in thermal expansion between silicon and copper. The molybdenum plate is then soldered to a compact copper terminal, which conducts the current to the anode. The copper terminal is again soldered to the copper surface of a ceramic 'direct bonded copper' (DBC) substrate, than provides the electrical insulation. The substrate is attached to the base plate by another solder layer. In sum, the construction contains five solder layers. Despite of the complexity of this construction, this power module is manufactured these days in a high production quantity on an automated assembly line.

The cross section image in Fig. 11.12 visualizes the numerous interfaces that the heat flow has to overcome on his passage from the silicon device to the base plate and further into the heat sink, which is not shown in the schematics. Since every solder layer has a small risk of the formation of solder voids, the multitude of solder interfaces represent an increased risk factor for potential sources of error.

The introduction of advanced power devices like IGBTs or MOSFETs has induced the development of a package concept capable of housing multiple chips per electrical function in parallel. This architecture has emerged to the 'standard' or 'classical' module design in power electronics. The example displayed in Fig. 11.13 illustrates the general features of this concept: The top side contact of the silicon device is connected via aluminum wire bonds. The molybdenum adaptation plate and the bottom side copper terminal are completely eliminated. Trenches to form current tracks comparable to the familiar PCB in low power electronics structure the upper copper layer of the DBC substrate. Several power chips are



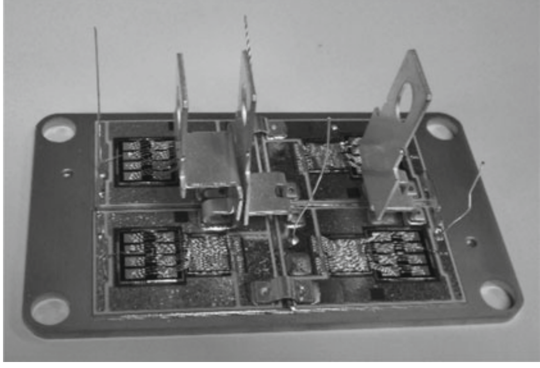


Fig. 11.13 Classical base plate module in schematic cross section (top) and as halfbridge IGBT module with two chips per switch (bottom)

directly soldered to these copper tracks and connected to other tracks by aluminum wire bonds. Powerful load current terminals are soldered to the load current tracks of the substrate.

Table 11.1 lists the layer thicknesses typically implemented in standard modules displayed in Fig. 11.13. This generic construction is found in 70–80% of all power modules produced by European manufacturers (Infineon, Semikron, IXYS, Danfoss, Dynex) and is also common in modules produce by Asian manufacturers.

The thickness of the ceramic layer is 0.63 mm in older generation modules, newer generation modules with base plate have a ceramic thickness of only 0.38 mm for improvement of thermal resistance. The thickness of ceramic plates is specified in the unit mil (1 mil = 1×10^{-3} inch = 25.4 μm), so that 0.635 mm is equivalent to 25 mil and 0.381 mm is equal to 15 mil. The substrate is attached to the base plate by a solder interface. Differences in solder thickness between 0.07 and 0.1 mm have a marginal impact on the thermal resistance.

For modules requiring a higher thermal conductivity or higher insulation strength, Al_2O_3 ceramics are replaced by AlN ceramics. The standard thickness of AlN is 0.63 mm, but for assemblies with extreme requirements with respect to insulation strength, ceramics with a thickness of 1 mm are applied. The fabrication of AlN substrates requires an additional process step compared to Al_2O_3 substrates, since no oxides are available at the surface to form an oxide-oxide interface as in the DBC production. Therefore, an oxide layer has to be generated first or other

Table 11.1 Layer thickness in module designs with base plate

	Standard module Al_2O_3 ceramic Cu base plate d (mm)		High power module AlN ceramic Cu base plate d (mm)		High power module AlN ceramic AlSiC base plate d (mm)	
Solder	0.05–0.1		0.05–0.1		0.05–0.1	
Copper	0.3		0.3		0.3	
Ceramics	Al_2O_3	0.381 0.635	AlN	0.635 1.0	AlN	1.0
Copper	0.3		0.3		0.3	
Solder	0.1 0.07		0.3–0.4 0.2		0.1	
Base plate	Cu	3	Cu	5	AlSiC	5
Thermal paste	0.05		0.04		0.04	

bonding techniques have to be applied. This increases the costs for AlN substrates. Furthermore, the coefficient of thermal expansion of AlN (and thus also the CTE of an AlN-DBC) is smaller than that of Al_2O_3. This amplifies the difference in thermal expansion between the substrate and a copper base plate and reduces the lifetime of this interface under thermal stress. A countermeasure is to increase the thickness of the solder interface to 200 µm or more to reduce the strain in the interface [Yam03]. Spacers implemented before the solder process assure a homogeneous solder thickness. Another option applied in some high performance power modules is to replace the copper base plate by AlSiC, a metal matrix composite material. An AlSiC plate is manufactured by first forming a matrix of SiC with a controlled porosity and secondly filling the pores with liquid aluminum. The material parameters are determined by the ratio of both components and can therefore be tailored to the application.

While the implementation of AlSiC as material for the base plate has the advantage of an adaptable thermal expansion, it has the disadvantage of a reduced thermal conductivity compared to copper. This was the major reason that a module design concept without a base plate emerged. Even for copper as base plate material the base plate adds to the total thermal resistance in the vertical direction between the chip and the heat sink, so that systems without base plate should be advantageous. Figure 11.14 shows a schematic cross section of this architecture.

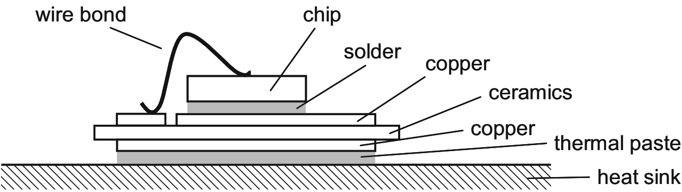


Fig. 11.14 Schematic cross section of a module without base plate

The applied substrates and solder materials are essentially the same as in modules with base plate. A value of 0.38 mm is the standard ceramic thickness for Al_2O_3 substrates, but thicknesses of 0.5 mm or 0.63 mm are also encountered. Especially for substrates with increased copper thickness for high current applications, thicker ceramics are preferred to enhance the mechanical robustness of the substrate (e.g. 0.4 mm Cu on both sides of a 0.5 mm Al_2O_3 ceramics). For AlN substrates 0.63 mm is the standard ceramics thickness. Other ceramic materials can easily be substituted in a module without base plate.

Modules without base plate are provided for example in the SKiiP-, MiniSKiiP- and Semitop module families from Semikron and in the EasyPIM series from Infineon. The packaging concept was available in several modules provided by IXYS for a long time. It is also applied in the insulated version of the TO-247 package. Since the solder interface between a base plate and the substrate is eliminated, only one single solder interface between the chip and the heat sink remains in this package type.

In contrast to base plate modules, no limitation of the substrate size must be obeyed in non base plate constructions. Therefore, complex circuits can be realized on a single substrate. The example of a substrate from a highly integrated module without base plate is shown in Fig. 11.15. It contains a single phase input rectifier and a three phase output inverter for a medium power frequency converter. Shunt resistors and a temperature sensor are also integrated.

The load, control and sensor terminals in this package concept are accomplished by identical springs. With this spring contact technology, a multitude of load and control contacts can be positioned at almost any location on the substrate, thus allowing a very flexible contact technology to realize a multitude of different circuits on the same package platform. Each spring can continuously conduct 20 A, higher currents can be attained by paralleling of springs.

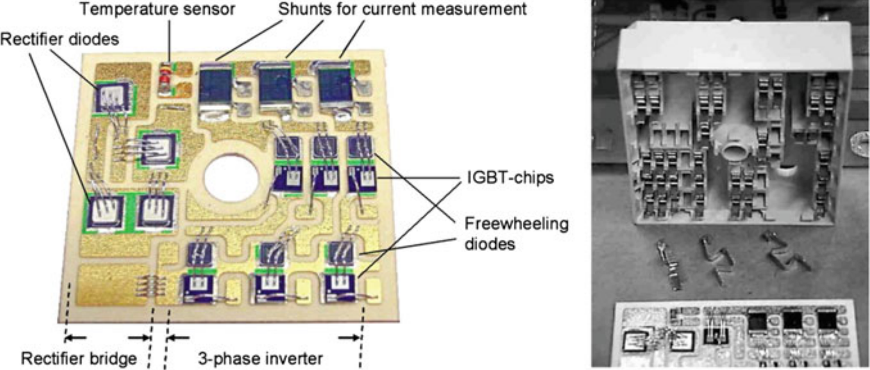


Fig. 11.15 Internal structure of a modern power module with input rectifier, output inverter and sensors (left) and housing with spring contacts (right) (MiniSKiiP by Semikron)

Modules without base plate are not limited in footprint size. They require minimum effort in production and also minimize the number interconnections in complex circuits, thus reducing the number of potential sources of failure. On the other hand, sophisticated pressure systems have to be integrated into modules with large footprints to ensure an optimum thermal contact between the substrate and the heat sink. The impact of the capability of this pressure system on the thermal interface between the module case and the heat sink results in different thicknesses of the thermal interface layer in Table 11.2. Finally, modules without base plate not only have advantages like the improved thermal gradients inside the module, which reduce the thermal stress and therefore increase the reliability under active thermal cycling [Scn99]. The absence of a base plate also has some drawbacks. First, the thermal spreading of the base plate no longer helps to reduce the temperature distribution across the chip. Therefore, smaller chip sizes are preferred in non base plate designs. The second drawback is the missing heat capacity of the base plate, which increases the thermal impedance of the module in a range between 50 and 500 ms for isolated overload events in this time range.

A common problem for all module designs is the interface between the module case and the heat sink surface. Due to the geometric tolerances of the contact surfaces and the modification of bow by the bi-metal effect caused by thermal expansion, no perfect metal-to-metal contact can be achieved. The gaps have to be filled with a 'thermal interface material' (TIM), which typically has a specific thermal conductivity in the range of 1 Wm^{-1} K^{-1}. Even though this conductivity is a factor 30 better than air, the conductivity is more than a factor of 100 worse than that of most metal layers. Therefore, the thickness of the thermal grease has to be kept as small as possible without the risk of air gaps.

The application of the optimum thermal grease thickness during the mounting process on a heat sink is a serious quality issue for many users of power modules. Semikron therefore delivers the SKiiP module family already mounted on the customer heat sink with a controlled thickness of thermal grease. For modules with base plate, the thermal resistance of the interface between module case and heat sink is specified by a typical value in the data sheet R_{thch}, which amounts to roughly 50% of the internal thermal resistance from chip to case. While this interface is

Table 11.2 Layer thickness in module designs without base plate

	Al_20_3 substrate d (mm)		AlN substrate d (mm)	
Solder	0.05–0.1		0.05–0.1	
Copper	0.3 0.4		0.3	
Ceramics	Al_2O_3	0.381 0.5 0.635	AlN	0.635
Copper	0.3 0.4		0.3	
Thermal paste	0.02–0.08		0.02–0.04	

difficult to establish in a controlled process, it is of greatest importance for the thermal characteristic of the power module in application. Recently, several manufacturers of power modules offer pre-applied TIM layers on the interface to the heat sink with optimized lateral distribution to simplify the mounting process for customers.

11.3 Physical Properties of Materials

The properties of the materials used in a package design are fundamental for the characteristics of the module. The most important parameters are the thermal conductivity and the coefficient of thermal expansion (CTE) of a material, but the electrical conductivity and the heat capacity are also of great interest. It is therefore inevitable to know and consider the properties of the materials prior to their implementation into a power module package.

A survey of the thermal conductivity of the most important materials in power electronic packaging is shown in Fig. 11.16. The best of the ceramic materials used for insulation feature thermal conductivities in the range of metals. Beryllium oxide, which exhibits the highest thermal conductivity, had been used in power module designs in the early days of module history. Nowadays, this material is implemented no more due to the toxicity of BeO dust and the resulting threats and limitations in handling and disposal of this material. Second in line of the ceramic insulators in this survey is AlN. But substrates with AlN are more expensive than standard Al_2O_3 substrates, so that this material is implemented only when high power density requirements or a demand for high basic insulation makes it inevitable. Organic insulators like epoxy or polyimide (Kapton®) only provide a comparable low specific thermal conductivity.

Inherent to the performance of a power module in application are varying load conditions, which generate temperature swings. Differences in the thermal

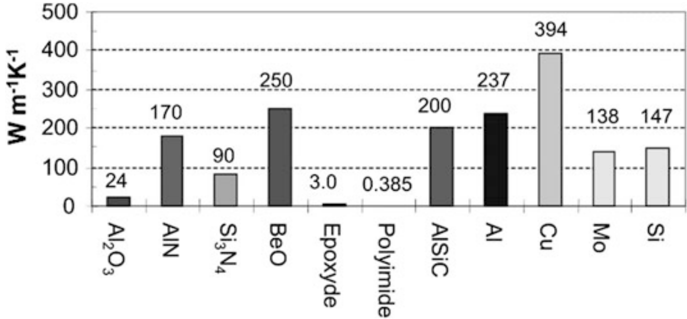


Fig. 11.16 Thermal conductivity at room temperature of different materials frequently used in packaging technologies

expansion of different materials stress the package. To minimize the stress induced by the thermal expansion between different adjacent layers, their CTE should be comparable (or more precisely in the presence of thermal gradients inside a stack of layers: the difference of the product of layer temperature and the CTE in adjacent layers should be as small as possible).

Figure 11.17 illustrates the fact, that the CTE of Si and Cu are quite different. It is therefore very unfavorable to connect both materials directly, as is the case in standard TO packages (Fig. 11.5). Implementing a ceramic substrate between the Cu base plate and the Si chip – which is a general concept of power modules – considerably reduces the thermal mismatch between adjacent layers and thus increases the lifetime. The adoption of Al_3O_3 DBCs with 0.3 mm Cu layers leads to similar stress in the two interconnection layers between chip and substrate and between substrate and base plate [Scn99]. Replacing Al_2O_3 DBCs with AlN DBCs reduces the stress from thermal expansion between the chip and the substrate but increases the stress between the substrate and a copper base plate. Implementing AlSiC base plates reduces the stress in the interface to the substrate in high performance power modules. The ratio of the two components of this metal matrix compound allows to adjust the CTE of the material to an optimal value for AlN substrates. On the other hand, a considerably reduced thermal conductivity is the consequence as shown in Fig. 11.16. Al_2O_3 as the prevailing ceramic material for power (DBC) substrates is from the thermal expansion point of view the best compromise to attach to silicon on one side and to copper on the other.

The organic insulation materials epoxy and polyimide (Kapton®) have a wide elastic deformation range, so that the coefficient of thermal expansion is not of interest and is therefore omitted in Fig. 11.17. Otherwise, these organic insulators are characterized by a much higher breakdown voltage (refer to appendix C, D) and thus can be implemented in very thin layers. Table 11.3 gives a compendium of standard material parameters and standard thicknesses, which are established in the packaging technology. The comparison shows, that a polyimide layer has a more than 10 times smaller thickness compared to ceramic insulators for an equivalent breakdown voltage.

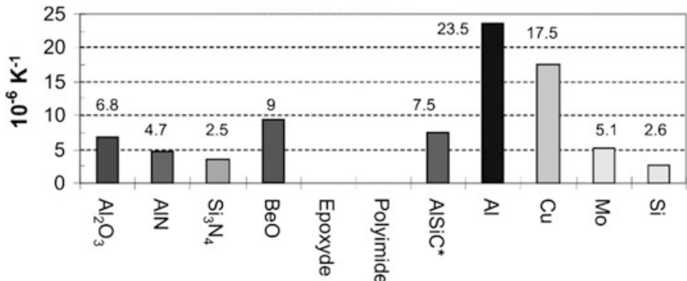


Fig. 11.17 Coefficient of thermal expansion (CTE) at 25 °C of materials frequently used in package technologies. (*) Depending on the composition of compound

Table 11.3 Standard layer thickness of insulators and emanating properties

Material	Standard thickness (μm)	Heat transfer coefficient ($WK^{-1}cm^{-2}$)	Capacity per unit area ($pFcm^{-2}$)	Breakdown voltage (kV)
Al_2O_3	381	6.3	22.8	5.7
AlN	635	28.3	12.5	12.7
Si_3N_4	318	28.3	25.6	4.5
BeO	635	39.4	11.8	6.4
Epoxy	120	2.5	52.4	7.2
Polyimide	25	1.5	138.1	7.3

Despite of the small layer thickness, substrates based on organic insulators exhibit a smaller thermal conductivity than ceramic substrates [Jor09]. Additionally, the small layer thickness provokes a high electrical capacity, which as parasitic capacity interacts detrimental with the power circuit.

The comparison of all properties of the insulation materials delivers that AlN is technically the best choice as insulating material for power semiconductor packages, if BeO is abandoned due to its toxic characteristics. AlN possesses the highest thermal conductivity and it is indispensable by virtue of its high breakdown voltage for modules with a blocking capability >3 kV. However, AlN exhibits due to its brittle structure an increased risk of fracture and thus inflicts a greater challenge for the industrial production of modules. The high bending strength of Si_3N_4 allows reducing the layer thickness to achieve a heat transfer coefficient comparable to AlN in the voltage range up to 1200 V. Si_3N_4 substrates are currently evaluated by several groups for automotive applications.

11.4 Thermal Simulation and Thermal Equivalent Circuits

11.4.1 Analogy Between Thermal and Electrical Parameters

The differential equations describing the physical process of one-dimensional heat conduction have the same form as the set of equations characterizing the one-dimensional electrical conduction. By exchanging the corresponding parameters, a thermal problem can therefore be transformed into an electrical problem and vice versa. Due to the equivalence of the differential equations, all operations performed for electrical networks can be transferred to thermal networks, especially the approximation of a continuous conduction line by a set of discrete elements in a lumped network. Since a variety of tools is available today for the simulation of electrical networks, thermal problems can be calculated by solving the equivalent electrical circuit.

The standard procedure is to first transform the thermal parameters into the corresponding or analogue electrical parameters. Then the corresponding equivalent network can be solved by applying advanced electrical network simulation tools. Finally, the results are transformed back into the thermal parameters. Table 11.4 gives a list of the fundamental corresponding parameters [Lap91].

From these fundamental parameters, other corresponding parameters can be derived. The electrical time constant as the product of resistance and capacity for example has its correspondence in the thermal time constant, defined as the product of thermal resistance and thermal capacity.

While, at the first glance, the correspondence between electrical and thermal parameters seems to be perfectly symmetrical, there is a difference which destroys that perfect symmetry. This difference is the explicit appearance of the temperature in the thermal equations. To examine this difference closer, let us consider the definition of the thermal resistance R_{th} between the geometrical locations a and b:

$$R_{th(a-b)} = \frac{T_a - T_b}{P_V} = \frac{\Delta T}{P_V} \tag{11.4}$$

In the electrical theory, Ohm's law postulates that the ohmic resistance is constant and therefore independent of the voltage, if the boundary condition of a constant temperature is fulfilled. This boundary condition reflects the fact, that material properties are generally dependent on temperature. But since the temperature is an explicit parameter in the definition of the thermal resistance, a correspondence to Ohm's law in the thermal theory with the boundary condition T = const. is not reasonable. This means, that the thermal resistance is always temperature dependent [Scn06].

The temperature dependence of the specific thermal conductivity of silicon, aluminum and copper illustrates Fig. 11.18 according to [EFU99]. Following [Poe04], the temperature characteristic of silicon can be approximated between −75 and +325 °C by the expression

$$\lambda_{Si} = 24 + 1.87 \times 10^6 \cdot \mathrm{T}^{-1.69}\,\mathrm{Wm}^{-1}\mathrm{K}^{-1} \quad \text{with T in Kelvin} \tag{11.5}$$

The thermal resistance (11.4) is constant only if λ is temperature independent. This applies in good approximation between −50 and +150 °C for Al and Cu and for most other materials. In power electronic systems, the thermal resistance of silicon amounts to only 2–5% of the total resistance, so that a negligence of its

Table 11.4 Equivalent electrical and thermal parameters

Electrical parameter	Thermal parameter
Voltage V (V)	Temperature difference ΔT (K)
Current I (A)	Heat flux P (W)
Charge Q (C)	Thermal energy Q_{th} (J)
Resistance R (Ω)	Thermal resistance R_{th} (K/W)
Capacity C (F)	Thermal capacity C_{th} (J/K)

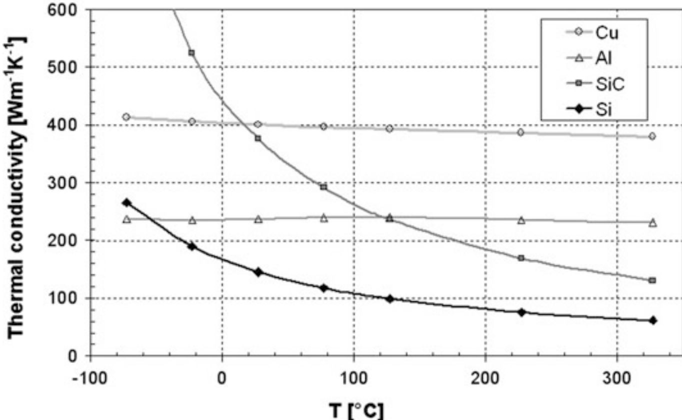


Fig. 11.18 Temperature dependence of the thermal conductivity of Si, SiC, Al and Cu. Data according to [EFU99] and [Fel09], solid line for Si calculated by Eq. (11.5)

temperature dependence results in most cases only in a small error. To eliminate this fundamental problem, a temperature dependent resistance can be simulated by using a voltage dependent resistor in the equivalent network, which is possible in most electrical network simulation tools.

Another general problem in thermal simulation is the interpretation of temperatures. Conventional reference points are the ambient temperature T_a, the heat sink temperature T_s, sometimes the case temperature T_c and the so called 'virtual' junction temperature T_{vj}. Three-dimensional systems that are not in a state of thermal equilibrium exhibit pronounced gradients of temperature in every layer of the system. So a single temperature values T_c or T_s must be clearly defined in a real system, in which the base plate (as the module case) and the heat sink surface are characterized by temperature distributions.

Especially, this holds true for the junction temperature T_j. In the power device where the power is dissipated, the greatest gradients of temperature are present. Thus, it is expedient to postulate a virtual junction temperature T_{vj} as a characteristic temperature of the silicon device. This parameter is defined by the measured voltage drop over a pn-junction for a small sense current, as was already discussed in Sect. 3.2.

The forward voltage drop of a pn-junction at very small current depends strongly on temperature. It is always decreasing with temperature. To use this effect to determine the temperature, the sensing current must be small enough, that a temperature influence of the sense current can be neglected. Typically a current density of 100 mA/cm^2 or lower is selected. Figure 11.19 shows a measurement of the forward voltage drop at the pn-junction of a 50 A diode measured at 50 mA as function of temperature. After determination of this calibration function, the junction temperature of the device can be measured by applying a sense current of 50 mA and measuring the voltage drop at an instant where the device is supposed

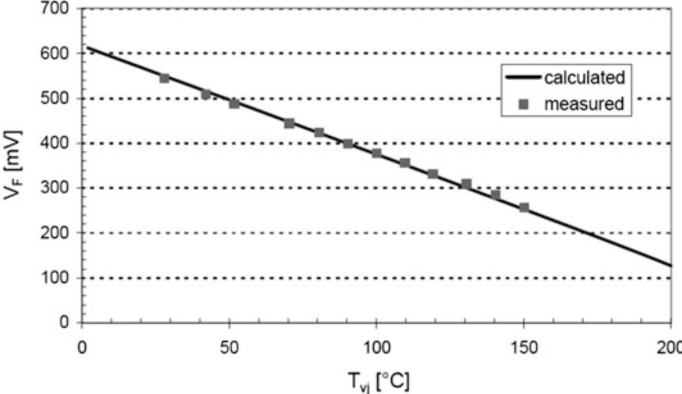


Fig. 11.19 Calibration of the pn-junction of a 50 A 1200 V Si-diode for use of the pn-junction as temperature sensor. Forward voltage drop at the pn-junction of the 50 A diode measured at 50 mA. Calculation according Eqs. (3.53) and (3.55)

neither to a forward current nor to a blocking voltage. Figure 11.19 compares the measured calibration function with the calculation according Eqs. (3.52), (3.53) and (3.55), resolved for V. As ideality factor, $n = 1.05$ was used in Eq. (3.55) for this fast recovery diode.

Together with the calibration of this voltage drop at different ambient temperatures, this method delivers a convenient technique to determine the virtual junction temperature of a device without intrusion into the package. This technique works well with diodes and IGBTs. The pn-junction between the gate and the cathode can be used for this method for thyristors. For the MOSFET, the inverse diode can be utilized for the measurement of the virtual junction temperature; in this case the sense current is applied in reverse direction.

As pointed out before, there is no constant temperature on the surface of a real power device in non-equilibrium condition. The edges of the silicon chip have a lower temperature than the center of the chip, because the heat flux can propagate not only vertically towards the heat sink, but it can also spread out away from the chip center, which can be envisaged by the cross-sectional illustrations in Figs. 11.13 and 11.14. This phenomenon is called heat spreading. The exemplary simulation in Fig. 11.20 illustrates the impact of a power dissipation of 200 W, homogeneously generated in the volume of a 12.5×12.5 mm^2 IGBT chip. The calculated temperature distribution reveals a center temperature, which is approximately 20 °C higher than the cooler edges of the chip.

An experimental validation of the simulated temperature distribution was presented in [Ham98]. The temperature was measured using a potential separated sensor, consisting of a phosphorescent powder at the end of a silica glass rod, which was excited by a laser. The temperature dependence of the phosphorescent radiation was used to measure the temperature at the tip of the silica glass rod. With this potential separated sensor, the surface temperature of an IGBT chip could be

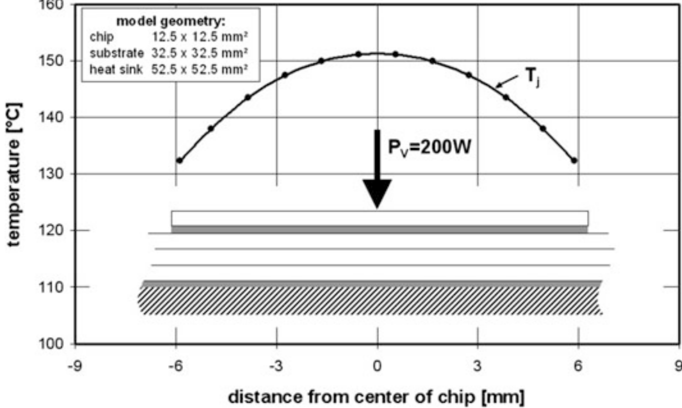


Fig. 11.20 Simulated temperature distribution in a silicon chip with layers according to Fig. 11.14. Illustration from [Scn06]

measured at different locations. The measured temperatures at the center and at the edge of the chip were related to the virtual junction temperature determined by the voltage drop for a sense current of 100 mA (Fig. 11.21). For a high load current, the temperature difference between center and edge was also found to be in the range of 20 °C. The results also show that the temperature T_{vj} is an average value for the real temperature distribution, which is shifted towards the hotter chip center temperature.

The reason for this shift towards the hot chip center temperature is found in the temperature characteristic of the voltage drop across the backside pn-junction of the IGBT (refer to Fig. 10.10, junction J_1) for small currents. Although the temperature

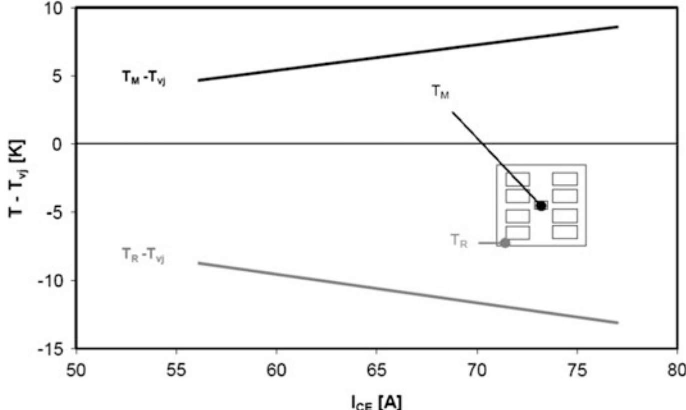


Fig. 11.21 Relation between the surface temperature of an IGBT and the virtual junction temperature T_{vj}, according to [Ham98]

coefficient of the forward voltage drop of an IGBT is positive in the range of nominal current, the voltage drop for small currents is only determined by the physical properties of the pn-junction and thus exhibits a negative temperature coefficient. This results in a smaller resistance of the hot chip areas and therefore implies a greater weight of the hot areas in the averaging process by the sense current flow. This is a desirable feature of the temperature measurement because the emphasis on the areas of higher temperature reduces the difference to the maximum temperature in the real chip. But it should be kept in mind, that the maximum temperature can still be considerably higher than the virtual junction temperature for large chips and high load currents.

Determining the junction temperature from the temperature dependence of electrical parameters of the chip itself has the great advantage that no sensors are necessary and thus no intrusion into a module package or modifications by the manufacturer are required. Any temperature sensitive electrical parameter (TSEP) of a device can in principle be applied for a sensorless measurement of the chip temperature [But14]. However, it should be emphasized that different temperature values will be delivered by different measurement methods. Especially for comparison between different measurements or to FEM simulation results, the geometrical interpretation of the measured temperature value is essential.

The most common TSEP method is the $V_j(T)$ method, also named $V_{CE}(T)$ method, as discussed above which delivers the virtual junction temperature T_{vj}. For fast switching devices like modern IGBTs the temperature can be measured 100 μs after turn-off of the load current. An extended investigation of the averaging effect of the sense current in [Scn09] shows that the measured temperature value corresponds to the area related average value of the temperature.

However, there are two major limitations of sensorless measurement methods. The first limitation is the fact, that most of these methods cannot directly be applied to determine the chip temperature in real switching applications like a PWM-operation in a frequency inverter. There is no simple possibility of applying defined operating conditions for TSEP evaluation in a real inverter operation.

The second limitation is attributed to the averaging process of these methods. For a single chip only one characteristic temperature value is delivered with no information on the temperature difference between the hot center and the cooler corners of the chip. This problem becomes even more pronounced in case of parallel chips. If one of the parallel chips has a higher thermal resistance – for example due to a deficient solder quality – its higher temperature will only have a small impact on the average value.

The measurement technique for determining the other reference point temperatures T_c and T_s is also not trivial. For the measurement of the case temperature T_c, which is a common reference point for classical modules with base plate, a drilling has to be incorporated into the heat sink exactly in the center of the silicon device generating the power losses as displayed in Fig. 11.22. This measurement therefore requires the knowledge of the exact position of the chips inside the module. This drilled hole interferes with the heat flux into the heat sink. However, due to the thermal spreading

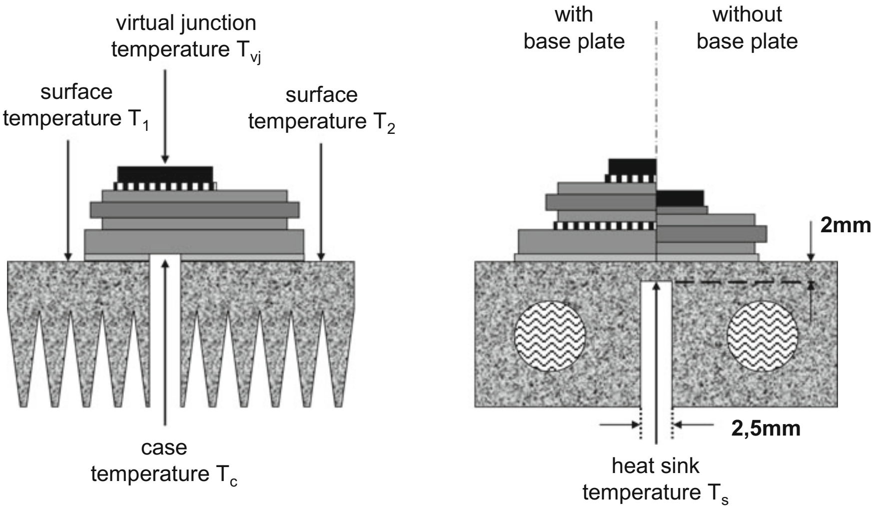


Fig. 11.22 Definition of the case temperature T_c and the heat sink temperature T_s

in the base plate of classical modules, this disturbance results only in a deviation of $\leq 5\%$ form the undisturbed value, as was verified by thermal simulation.

The impact of such a drilled hole for temperature sensing on a module without base plate is much more severe because of the non-existing spreading of a base plate. It was proposed by [Hec01] to replace the through hole in the heat sink by a blind hole, which only reaches up to 2 mm underneath the heat sink surface. This measurement configuration has the advantage, that the thermal interface between module and heat sink is integrated in the heat path. This method can be applied for any type of module. The reference temperature defined by this geometry is called heat sink temperature T_s (in older publications often indicated by T_h).

In contrast to the measurement of the virtual junction temperature, the measurement of the case or heat sink temperature is mostly restricted to equilibrium state conditions. The transient response of thermocouples is in the range of 100 ms and more, so they cannot be utilized for the measurement of fast temperature evolutions in power modules, which have a typical time constant for the internal thermal resistance of approximately 1 s.

These considerations shall illustrate that the thermal characterization of a power electronic system is never easy – neither by simulation nor by measurement. A lot of experience and a critical mindset are necessary to select the right model and to interpret simulation results. Temperature measurements have to be carefully reviewed as well. It is mandatory in scientific publications to discuss the applied methodology for temperature measurements for the interpretation of results. The thermal characterization is one of the most difficult tasks in power electronic systems and only succeeds by combining experimental skill with correctly applied thermal simulation.

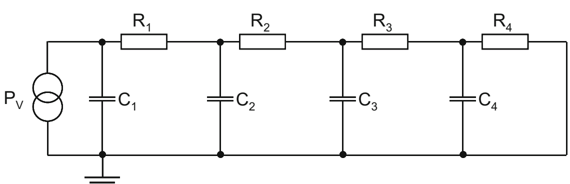
11.4.2 One-Dimensional Equivalent Networks

In a one-dimensional equivalent network, the power dissipated in a thermal heat source is represented by a current source. A network of resistors and capacitors represent the thermal resistances $R_{th,i}$ and the thermal capacities $C_{th,i}$ of the analogue thermal system. The ground potential is equivalent to the ambient temperature. In the physically correct Cauer-model, the thermal capacities are connected from each node of the model to the ground potential. If power losses are generated in the system, the temperature will rise in the nodes and thermal energy is stored in the capacitors. The stored energy is proportional to the temperature difference to the situation before the power losses were applied; therefore, this network correctly describes the physical reality (Fig. 11.23).

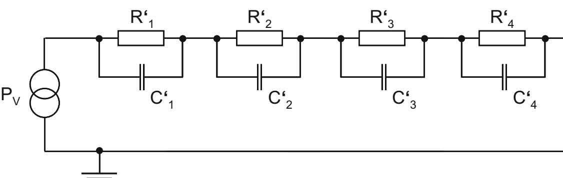
In contrast to the Cauer-model, the capacitors are connected in parallel to the resistors in the Foster-model. It should be noted that the values of the resistors and capacitors are different in both networks! The equivalent Foster-model has in total the same transient behavior as the Cauer-model with respect to the temperature of the first node next to the power source. While the internal nodes in a Cauer-model can be interpreted as geometrical locations in the system, this is not possible for internal nodes of the Foster-model. The feature, that pairs of Rs and Cs in a Foster-model can be exchanged without altering the transient response of the whole system might help to remember this important fact. The exchange of pairs of Rs and Cs in the Cauer-model will on the other hand alter the transient response of the system, as the exchange of layers in a real system would do. This missing link to the system geometry also implies that the values of the resistors and capacitors in the Foster-model cannot be calculated form material constants, as is the case for Cauer-models. Finally, a Foster-model cannot be divided, neither can two Foster-models be connected together, while both operations are possible for Cauer-models.

Fig. 11.23 One-dimensional thermal equivalent networks

Cauer-model, physically correct



Foster-model, parameters given in data sheets



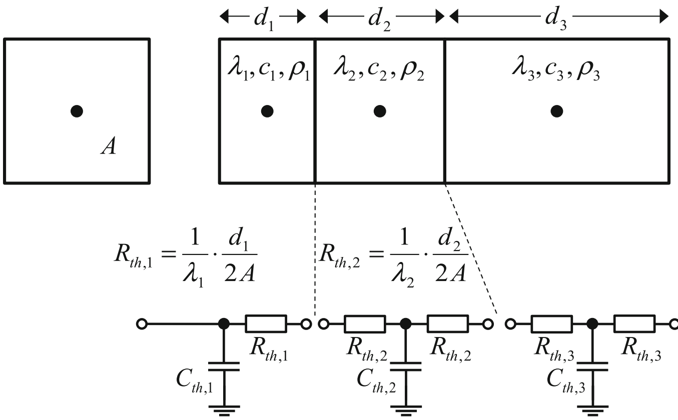


Fig. 11.24 Extraction of R_{th} and C_{th} values of a Cauer network from geometry and material constants of a layer system

Those severe restrictions in the application of Foster-models lead to the question, why we use this type of model at all. The answer is, that the time dependent thermal resistance – often referred to as the 'thermal impedance' Z_{th} of a system – can be expressed by a simple analytical expression. The model parameters R_i' and C_i' in this expression can be determined from the system response to a step function in power losses. By applying a least square fit algorithm, the parameters in the analytical expression can be optimized until the time response matches the transient system response, for example measured by a heating or cooling curve.

$$Z_{th} = R_{th}(t) = \sum_{i=1}^{n} R_i' \cdot \left[1 - \exp\left(-\frac{t}{\tau_i}\right)\right] \text{ with } \tau_i = R_i' \cdot C_i' \tag{11.6}$$

The values of the R_i' and τ_i' are often explicitly listed in data sheets of power packages. They allow a fast calculation of the transient response of a package to complex power distributions for application engineers.

The resistances and capacitors in the Cauer-model can be calculated straight forward from the material parameters and geometry:

$$R_{th} = \frac{1}{\lambda} \cdot \frac{d}{A} \tag{11.7}$$

$$C_{th} = c \cdot \rho \cdot d \cdot A \tag{11.8}$$

with layer thickness d, cross section area A, specific thermal conductivity lambda, specific heat capacity c, and specific density ρ. When the nodes of the equivalent network are chosen to be located in the center of gravity of each layer of homogeneous material, the thermal capacity of each layer is defined by the material

parameter and the layer volume. The resistance between two nodes is then composed of two contributions, defined by the material constants and half the thickness of each of the layers in contact (Fig. 11.24). Power losses generated homogeneously in a layer are represented by current injected in the node of the layer and the voltages at the nodes represent the average layer temperature.

With these extraction rules, a Cauer-model of a layer system can be derived, which allows a complete geometrical interpretation and all geometrical operations such as combining or dividing systems.

11.4.3 The Three-Dimensional Thermal Network

The one-dimensional Cauer-model is only a rough approximation for the complex geometry of layers in a real power module. The layers typically exhibit different cross sections and the resulting lateral heat spreading cannot be described by a one-dimensional model (refer Sect. 11.2, especially to the Figs. 11.13 and 11.14). The thermally high conductive copper layers extend the effective heat conduction area above the layers with a high thermal resistance (ceramic, thermal paste). These three-dimensional features can be accounted for by extending the Cauer-model to three dimensions as shown in Fig. 11.25. The nodes are arranged in a three-dimensional lattice; each node is connected to its neighbors via resistors.

The nodes in Fig. 11.25 are located in the center of each cuboid element. The resistors between the nodes are determined by the material parameters along the path, so that across the interface of two adjacent layers the resistors are determined by the material parameters of both layer materials.

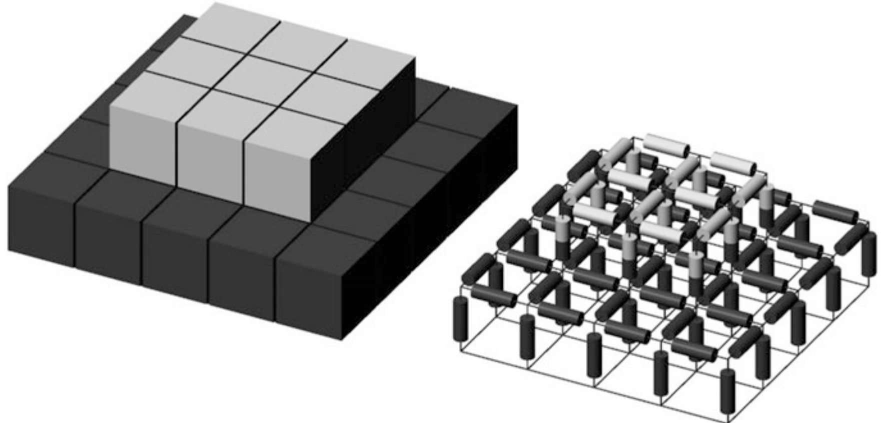


Fig. 11.25 Three-dimensional model for a simple two-layer system showing the lattice of nodes with the interconnecting resistors, diagram from [Scn06]

Figure 11.25 also exemplifies that the number of elements is increasing fast in the three-dimensional model. Even though this simple system contains only 2 layers with 9 nodes for layer 1 and 25 nodes for layer 2, a total number of 86 resistors are necessary to connect the nodes. Additionally 34 capacitors have to be connected from every node to the ground potential if the transient response of the system is of interest. For a realistic model of a power semiconductor package, several hundred of nodes with more than a thousand elements are adequate. Such complex networks require fast network simulation tools and a suitable pre-processor to generate the input data from a given geometry. However, the simulation results reveal the temperature evolution in layers, which are not accessible to measurement without considerable interference with the system. Furthermore, the three-dimensional simulation is the only way to calculate the impact of disturbances by a measurement setup and allows quantifying the offset with respect to an undisturbed system. Finally, fast transient temperature evolutions in all layers other than the silicon device are virtually not accessible by measurement [Hec01] and only simulation models deliver the accurate data necessary for the thermal characterization of a power module package.

11.4.4 The Transient Thermal Resistance

A simulation of the transient thermal resistance or thermal impedance Z_{th} based on a three-dimensional model is shown in Fig. 11.26. Three different power module packages using AlN-substrates are calculated for comparison: two modules with base plate according to Fig. 11.13 with layer dimensions as listed in Table 11.1 and one module design without base plate according to Fig. 11.14 with layer thicknesses as given in Table 11.2.

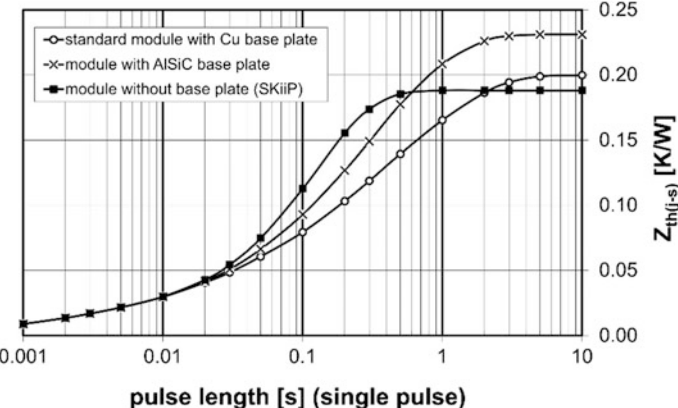


Fig. 11.26 Transient thermal resistance simulated for different module designs based on AlN substrates. Figure from [Scn99]

The thermal impedance is small for a short single pulse below 50 ms and it is independent of the module design, because the heat is almost completely stored in silicon chip and the substrate. For large pulse length, the thermal impedance approaches the equilibrium value of the thermal impedance, which is the thermal resistance R_{th}. The comparison of the different designs reveals that the thermal resistance of the module without base plate is moderately smaller than the thermal resistance of the module with a Cu base plate. This is a consequence of the thermal resistance of the base plate in vertical direction. The thermal resistance of the module with AlSiC base plate is considerably higher due to the inferior thermal conductivity of AlSiC. This inferior thermal conductivity also increases the thermal impedance in the intermediate range between 50 and 500 ms compared to the copper base plate modules. However, the Z_{th} of the module without base plate features the highest values in this interval. This is a consequence of the missing thermal capacity of the base plate. Therefore, the module design without base plate possesses less buffer capacity for single pulse overload conditions between 50 and 500 ms.

However, the single pulse response of a system is relevant for a limited number of applications, only. A single pulse event draws maximum advantage out of additional thermal capacity in the system, because there is an infinite time to dissipate the heat after the end of the pulse. The situation is different in case of repetitive pulses. Then there is only limited advantage of the additional thermal capacity, in our case the base plate.

The temperature evolution in a system [described by Eq. (11.6)] for an infinite series of constant power pulses can be calculated analytically. For such a series of constant pulses of the constant power P_{on} during the time t_{on} followed by no power for the time t_{off}, we can calculate the stationary maximum temperature swing:

$$\Delta T_{\max,\,\text{stationary}} = P_{on} \sum_{i=1}^{n} R_i' \frac{1 - \exp\left(-\frac{t_{on}}{\tau_i}\right)}{1 - \exp\left(-\frac{(t_{on} + t_{off})}{\tau_i}\right)} \tag{11.9}$$

We can also calculate the minimum temperature of the stationary temperature ripple generated by a sequence of constant pulses:

$$\Delta T_{\min,\,\text{stationary}} = P_{on} \sum_{i=1}^{n} R_i' \frac{\exp\left(-\frac{t_{off}}{\tau_i}\right) - \exp\left(-\frac{(t_{on} + t_{off})}{\tau_i}\right)}{1 - \exp\left(-\frac{(t_{on} + t_{off})}{\tau_i}\right)} \tag{11.10}$$

Both equations together deliver the stationary temperature ripple:

$$\Delta T_{ripple,stationary} = \Delta T_{\max,\,stationary} - \Delta T_{\min,\,stationary} \tag{11.11}$$

If we define the duty cycle as the ratio $d = t_{on}/(t_{on} + t_{off})$, we can discuss the impact of the thermal capacity on the system response in more detail. For very low duty

cycles, the single pulse characteristic as shown in Fig. 11.26 can be used as good approximation. In real applications, this applies to welding applications or some induction heating applications with long pauses between high current pulses. For motor drive applications, which are still the majority of all power electronic applications, the IGBTs as well as the diodes are under load during one half wave of the output current and under no load during the next half wave. It shall be simplified in a first approximation as a load duty cycle of 50%. Then the thermal impedance of the three systems of Fig. 11.26 compare as shown in Fig. 11.27. Here we see that the system without base plate is still inferior to the system with a copper base plate, but it is superior to the AlSiC base plate design in the whole frequency range. Due to the much smaller thermal conductivity of AlSiC, the heat stored in the base plate cannot be dissipated fast enough into the heat sink to give an advantage for the 50% duty cycle.

Equations (11.9) and (11.10) are also very useful to determine the ripple amplitude in equilibrium state. The difference between the maximum temperature and the minimum temperature defined by these equations delivers the amplitude of the temperature ripple in steady state condition. Figure 11.28 illustrates these dependencies.

An application example shall be given. We consider a dissipated power of P_{av} = 300 W in an application with an output power frequency of 5 Hz and a duty cycle of 50%. This is for example a slowly rotating electric motor fed by a variable speed drive with IGBTs. For the module with Cu base plate we get from Eq. (11.4) a temperature increase ΔT_{jav} = 60 K using the thermal resistance of 0.2 K/W as follows from Figs. 11.26 and 11.27 for the stationary condition. During the pulse length of 100 ms we have to calculate

$$\Delta T_{jmax} = Z^{\cdot}_{thjc} P_{on} \tag{11.12}$$

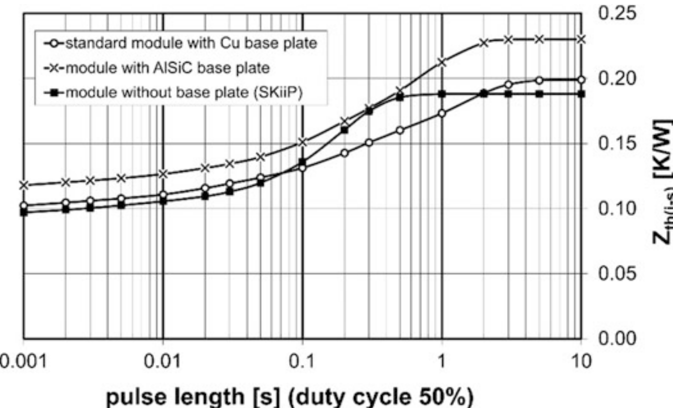


Fig. 11.27 Transient thermal resistance simulated for different module designs based on AlN substrates for a duty cycle of 50%

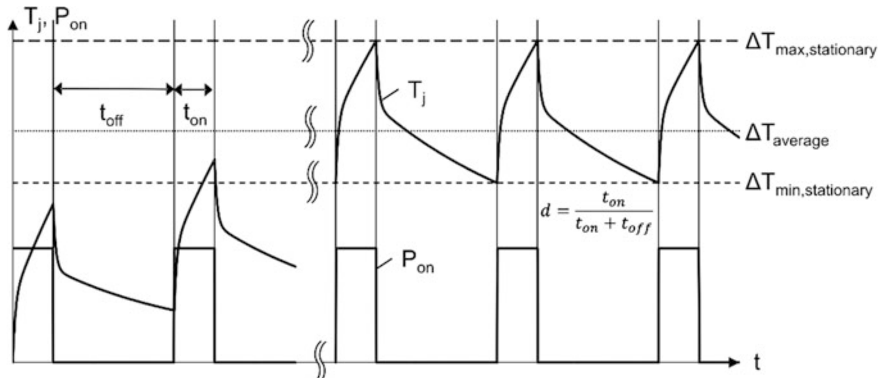


Fig. 11.28 Temperature ripple for sequence of constant power pulses P_{on}

with Z_{thjc} = 0.13 K/W from Fig. 11.27 and P_{on} = 600 W we get ΔT_{jmax} = 78 K. The difference of ΔT_{jmax} and ΔT_{av} is ½ of the temperature ripple in the special case of 50% duty cycle, so that the temperature ripple amounts to 36 K. For the module without base plate we get, using 0.19 K/W, as result ΔT_{jav} = 57 K; and with Z_{thjc} = 0.135 K/W from Fig. 11.27 we get ΔT_{jmax} = 81 K, the temperature ripple amounts to 48 K. For the AlSiC base plate results ΔT_{jav} = 69 K and ΔT_{jmax} = 90 K, the temperature ripple amounts to 42 K.

The Cu base plate module has a high thermal mismatch between Cu and AlN, therefore, in the viewpoint of high reliability only the other systems shall be considered. In the module without base plate we have a higher temperature ripple, in the AlSiC base plate module we have a lower temperature ripple, however at a significant higher temperature. Maximal temperature affects reliability as well as temperature ripple, for details on lifetime estimation see Chap. 12.

The used simplification of constant losses during t_{on} is not realistic in motor drive applications, in fact the on-state losses have a term proportional to $\sin^2(\omega t)$, the switching losses are proportional to $\sin(\omega t)$. The result of the comparison, however, will be similar.

11.5 Parasitic Electrical Elements in Power Modules

Every power module contains parasitic resistances and parasitic inductances causes by internal conduction tracks, as well as parasitic capacities provoked by parallel conductors separated by dielectric layers. Their influence is not negligible, especially during fast switching operation.

11.5.1 Parasitic Resistances

The significant contribution of external and internal leads to the total voltage drop in discrete packages was already addressed in Sect. 11.2. Figure 11.29 illustrates the evolution of package designs produced by International Rectifier (IR). Table 11.5 lists the characteristic parameters of these package types.

The substitution of wire bonds by a copper strap in the transition from the SO-8 package to the Copperstrap design reduces the parasitic resistance and the parasitic inductance. Since the progress in chip technology succeeded in reducing the on-state resistance to approximately 1 mΩ for a 40 V MOSFET, this package improvement was mandatory. The progress towards the Power-Pak housing is marked by a substantial improvement of the thermal resistance by implementing a solid copper base, which supplies an effective thermal path and at the same time represents the electrical drain contact. The ultimate progress was the development of the DirectFet package, which reduces the parasitic effects and the thermal resistance to a minimum.

The parasitic resistance in power modules is also considerable. For advanced power modules, the manufacturers often explicitly specify the parasitic resistance induced by the package in data sheets. Infineon indicates a value of 0.12 mΩ for the

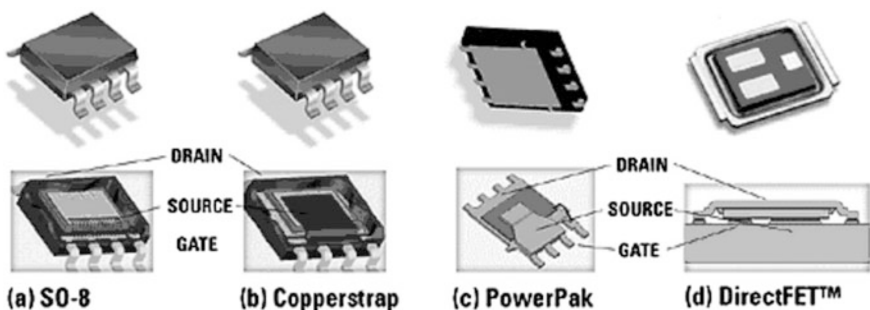


Fig. 11.29 Optimization of discrete package architecture concerning the reduction of parasitic resistance and induction, as well as the improvement of the thermal resistance, according to [Zhg04]

Table 11.5 Characteristic parameters of the package designs shown in Fig. 11.29 according to [Zhg04]

Package type	Electrical parasitic resistance (mΩ)	Parasitic inductance (nH)	R_{th} junction to PCB (K/W)	R_{th} junction to case (top surface) (K/W)
SO-8	1.6	1.5	11	18
Copperstrap	1	0.8	10	15
PowerPak	0.8	0.8	3	10
DirectFET	0.15	<0.1	1	1.4

high performance power module FZ3600R12KE3, which is a 1200 V IGBT module with a rated current of 3600 A. Thus, the parasitic resistance implies an additional voltage drop of 0.43 V at the nominal current of 3600 A. The on-state voltage drop of the IGBT has a typical value of 1.7 V. Therefore, the package provokes roughly 20% of the total voltage drop. Other high current modules feature comparable values.

If the 36 IGBT chips rated 100 A each in the Infineon power module would be replaced by 75 V 100 A MOSFET chips with an on-state resistance $R_{DS,on}$ of 4.9 mΩ, the voltage drop of the package would be in the same range as the voltage drop of the MOSFETs.

Even though the total parasitic resistance of a power module is an obstacle for power electronic applications because it generates additional losses and thus reduces the system efficiency, the consequences of internal parasitic resistances are even more severe, because they affect the static current distribution in high power modules with multiple parallel chips. This effect will be considered in more detail in the following simple example.

This simple model comprises 5 parallel diode chips with a rated current of 70 A each. The chips are soldered to a DBC substrate which is contacted to a heat sink by a thin layer of thermal interface material as shown in Fig. 11.14. The position of the cathode contact is on the left side, the anode contact is located on the right side of the substrate. The parasitic resistance of the Cu-layers on the DBC and of the wire bonds constitute an electrical network together with the (temperature dependent) forward voltage drop of the 5 diodes. With the material parameters and realistic assumptions on the geometry, the current distribution in this network can be solved with the boundary condition that the same voltage drop must occur for each current path.

In a first step, identical temperatures are assumed for all 5 diodes and the network is solved for 5 mm distance between the chips (Fig. 11.30a). Then the losses resulting from the individual currents through the diodes are inserted in an FEM model according to Fig. 11.30b and the calculated area-related average chip temperatures are determined in a second step. Re-inserting these temperatures into the network starts an iteration process which will converge against a thermal-electrical consistent solution for the given problem. The results are depicted in Fig. 11.31 (the lines are only drawn as a guide to the eye).

The solution for a distance between the chips of 5 mm, indicated by triangles, shows a variation of currents between 40 and 86 A while the temperatures vary between 115 and 150 °C with the maximum at the rightmost chip in Fig. 11.30a for the rated total current of 350 A. The main reason for this imbalance is the much smaller width of the anode current track on the DBC which leads to the maximum current and temperature for the chip at position 5.

Since the parasitic resistances between the chips are the root cause for the imbalance, a reduction of the resistances should improve the current distribution. This can be achieved by placing the chips closer together as illustrated by the second version in Fig. 11.30 with a distance of only 1 mm between the chips.

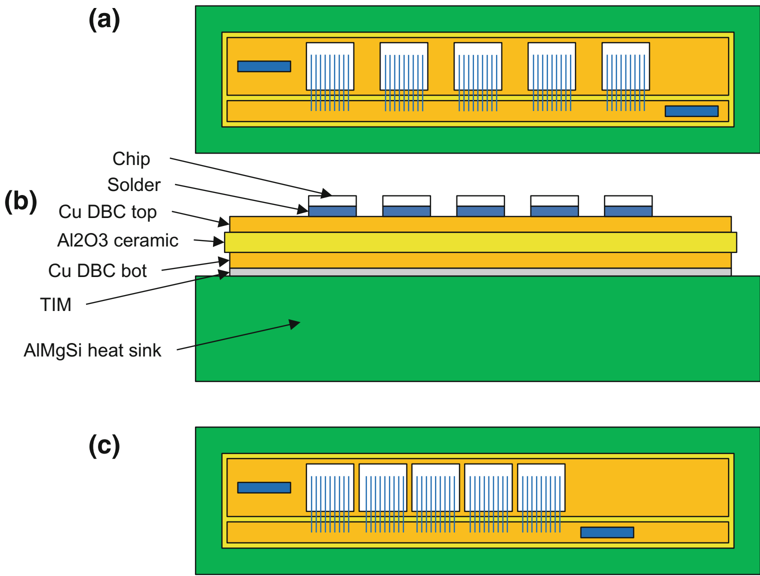


Fig. 11.30 Simple model of five parallel diode chips with 5 mm distance between chips in top view (**a**) and cross section (**b**) and version with 1 mm distance (**c**) [Scn15]

However, this layout change will also increase the thermal interaction between the chips. Repeating the same iteration procedure as for the layout with 5 mm distance between the chips leads to the result also displayed in Fig. 11.31 (indicated by the circles). It shows, that the maximum current can actually be reduced to ~77 A. However, the maximum chip temperature rises to 160 °C at chip position 4 for this layout with reduced parasitic resistance.

For this simple model the paralleling of diodes was considered. These diodes exhibit a negative temperature coefficient of the forward voltage drop, which amplifies temperature imbalances, since the hotter chip will attract even more current. For paralleling IGBTs with a pronounced positive temperature coefficient at nominal current the impact of thermal coupling would be less pronounced.

Although the absolute values shown in Fig. 11.31 are related to the specific model assumptions, the general trend applies to every high power module design with parallel chips: Each design will always be a compromise between the conflicting requirement of minimized parasitic resistance and minimized thermal coupling between the parallel chips. Further reduction of the on-state voltage for future power chip generations will increase the impact of unbalance in parasitic electrical resistance.

This example shows that it is an important goal to reduce the parasitic resistance in power modules, especially for high current and low voltage modules. However, the internal parasitic inductance has an even greater impact on the performance of the package.

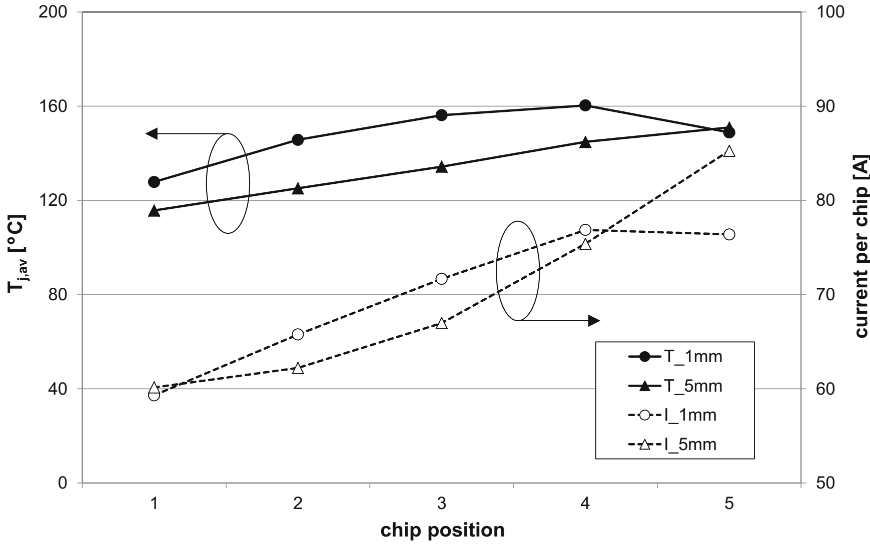


Fig. 11.31 Thermal-electrical consistent solution of the current distribution and the temperature distribution of the models defined in Fig. 11.30 [Scn15]

11.5.2 Parasitic Inductances

Every current lead is associated with a parasitic inductance. For the estimation of the magnitude of inductance, a rule of thumb applies for the inductance of current tracks:

$$L_{par} \approx 10\,\mathrm{nH/cm} \tag{11.13}$$

The inductance can be reduced by parallel arrangement of the plus and minus tracks; this technique is applied frequently in modern power modules. The module inductance is typically in the range of 50 nH for classical module designs. The more advanced packages, this value is reduced to 10–20 nH. These parasitic inductances affect the commutation circuit as shown in Fig. 11.32.

L1 and L6 represent the inductances of the DC-link capacitors and of the current tracks to the DC-link.
L2 is the inductance formed by the plus terminal and the current track on the substrate to the collector of the IGBT soldered to the substrate.
L3 is composed by the bond wires on the emitter of the IGBT and by the current tracks on the substrate to the AC-terminal.
L4 is synthesized by the current tracks from the AC-terminal to the cathode contact of the freewheeling diode, soldered to the substrate.

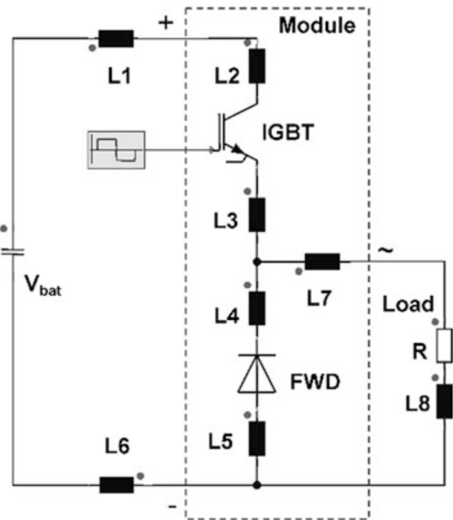


Fig. 11.32 Parasitic inductances in the commutation circuit

L5 consists of the bond wires on the anode contact of the freewheeling diode and the track to the minus terminal, including the terminal itself.
L8 represents the inductance of the load, which acts as the major current source during commutation. This inductance, as well as the series inductance of the AC-terminal L7, is not affecting in the commutation circuit. The effective internal parasitic inductances are all connected in series, so that they can be merged into a single module inductance L_{pm}.

$$L_{pm} = L2 + L3 + L4 + L5 \tag{11.14}$$

The total parasitic inductance L_{par} is then given by the series connection of the module inductance with the DC-link inductances.

$$L_{par} = L_{pm} + L1 + L6 \tag{11.15}$$

The impact of this parasitic inductance on the dynamic properties of a power module will be illustrated by two examples. In the first example, we will consider a frequency converter for a three-phase motor drive. This converter is formed by three IGBT half-bridge modules of the voltage class 1200 V with a nominal current of 800 A each.

The maximum DC-link voltage V_{DC} is 800 V and the parasitic inductance of each phase leg L_{par} is assumed 20 nH. The rate of current rise di_r/dt shall be

5000 A/μs. The voltage characteristic during commutation under these assumptions can be calculated by Eq. (5.72):

$$V(t) = -V_{DC} - L_{par} \cdot \frac{di_r}{dt} + V_{tr}(t)$$

Evaluation of this equation delivers an over-voltage peak of 100 V. Therefore, the maximum occurring voltage would amount to 900 V, which lies safely within the specification limits of the power modules. Furthermore, it is a typical feature of IGBTs, that the voltage $V_{tr}(t)$ is not exhibiting an abrupt cut-off, but rather decreases slowly after turn-on. Since $V_{tr}(t)$ has the opposite polarity compared to the voltage spike generated by parasitic inductance, no voltage spike above 800 V can be detected in an actual measurement. An example is shown in Fig. 5.21.

The second example considers half-bridge modules in an integrated starter-generator application for a 42 V vehicle power system of an automobile. The MOSFET power switches have rated current of 700 A each and blocking voltage of 75 V. As before, the parasitic inductance is assumed 20 nH and the di_r/dt shall be 5000 A/μs. Again, the over-voltage peak generated by the parasitic inductance would be 100 V, resulting in a total maximum voltage of 142 V. In contrast to an IGBT, the voltage decay in a MOSFET is rather abrupt after turn-on, so that it does not assist to reduce the total over-voltage spike. The resulting 142 V spike is clearly exceeding the maximum blocking voltage of the MOSFETs!

These examples illustrate a general feature in power electronic applications: Systems with high currents at a low voltage are most sensitive to parasitic inductance. Additionally, the problem of symmetric current paths is more severe in these applications.

To investigate this problem further, a parallel configuration with 5 IGBTs and the associated freewheeling diode with 1200 V blocking capability on a single DBC substrate is considered (Fig. 11.33a). The positions of the load terminals are indicated. A schematic circuit diagram for this design is depicted in Fig. 11.33b. The current tracks on the substrate are represented by the inductances L1 to L9, whereas L10 to L15 are symbolizing the wire bond connections.

While the current path relevant for commutation from the terminals via IGBT3 contains only four parasitic series inductances, the relevant current path via IGBT1 contains eight parasitic series inductances. Since the values of the parasitic inductances are in the same order of magnitude for the given geometry, a factor of 2 can be assumed between the parasitic inductance for the two IGBTs. This will result in a pronounced dynamical unbalance in the current distribution of this circuit during commutation. Moreover, the parasitic inductances can lead to oscillations between the chips, which will be investigated in Chap. 14.

It is difficult to find a symmetrical arrangement for a multitude of parallel chips in a high current power module; the example in Fig. 11.33 is all in all one of the better solutions. In this context, designs combining a chip connected via a geometrically short current track with chips connected in parallel via geometrically

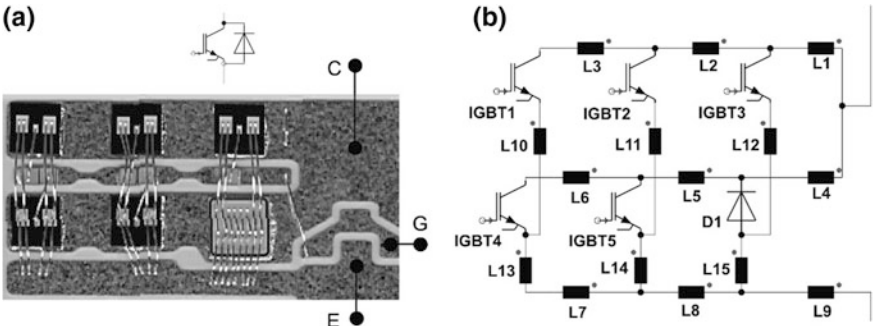


Fig. 11.33 (**a**) Realistic power circuit consisting of 5 parallel IGBT chips and one anti-parallel freewheeling diode chip (**b**) schematic circuit diagram showing the power devices plus the parasitic inductances formed by the current tracks

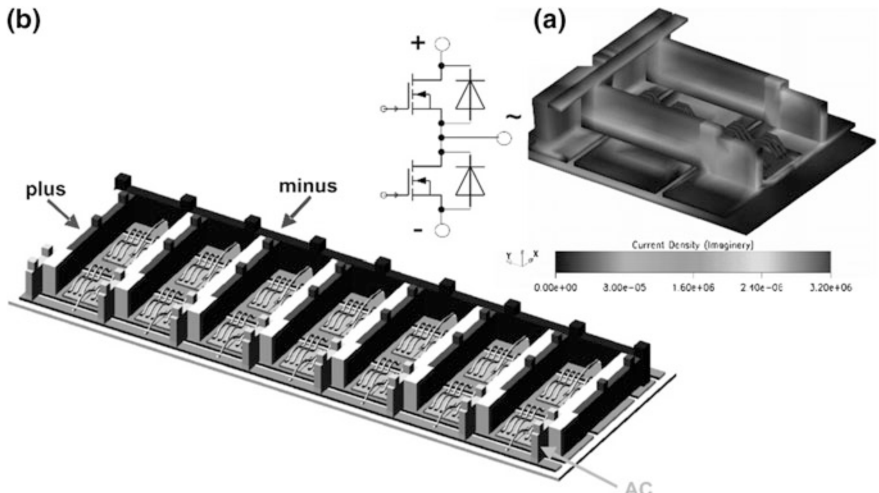


Fig. 11.34 Advanced MOSFET half-bridge configuration (**a**) Elementary cell with a simulated parasitic induction of 1.9nH (**b**) Highly symmetrical circuit design with 7 elementary cells in parallel per half-bridge [Mou02]

long tracks are especially problematic. In induction measurements, the low parasitic inductance along the short path will dominate the result, while internally extensive differences generate a severe dynamical imbalance.

Solutions to this problem have been proposed [Mou02], which denote a substantial progress especially for applications with high currents at low voltages. Figure 11.34 shows an example. The elementary cell is a half-bridge configuration of two MOSFET chips. The function of the freewheeling diodes is adopted by the

internal diode of the MOSFET switches. The design of the elementary cell was optimized by numerical simulation using a Fast-Henry-algorithm [Kam93], which allows to calculate the dynamical current distribution during high frequency commutation in the 3-dimensional model, with Skin-effect and eddy currents taken into account. The optimized cell design in Fig. 11.34a exhibits a parasitic inductance of 1.9 nH for a single elementary cell. By symmetrically paralleling 7 of these elementary cells and by connecting the DC-link bus bar as laminated metal sheets directly on top of the plus and minus terminals, a parasitic inductance in the sub-nH range was achieved. This module architecture is suitable for the application in an integrated starter-generator system as described in example 2 above.

11.5.3 Parasitic Capacities

Insulated substrates in a power module create a capacitor, which will also influence the dynamical characteristics of the power circuit, as illustrated in Fig. 11.35 for a simple construction.

The copper tracks on the substrate generate two capacities C_{PA} and C_{PK} connected to the ground contact, which is represented by the module case. The series connection of C_{PA} and C_{PK} is connected in parallel to the internal junction capacity of the diode. In more complex circuits, these capacities will also form capacitive coupling links to other parts of the circuit. The dimension of these capacities depends on the insulator material and the thickness of the insulating layer. Characteristic parameters are collected in Table 11.3, Sect. 11.3.

Since insulation layers of epoxy and polyimide have only a marginal thermal conductivity, but at the same time exhibit a very high breakdown voltage, layers of these materials are rendered very thin. This results in a high capacity per unit area and limits the applications for components with these insulation materials (i.e. IMS substrates) for fast switching devices.

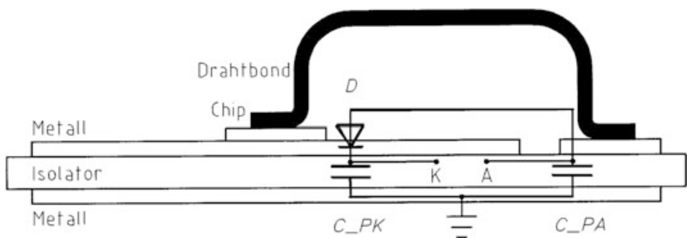


Fig. 11.35 Parasitic capacities in a package with a diode chip mounted on an insulated ceramic substrate [Lin02]

In an insulated TO-220 package, the cathode current contact features an area of 8 mm × 12.5 mm. Assuming a ceramic insulation layer of 0.63 mm thickness with a relative dielectric constant $\varepsilon_r = 9.8$ delivers a parasitic capacity on the cathode side C_{PK} according to

$$C_{PK} = \varepsilon_0 \cdot \varepsilon_r \cdot \frac{A}{d} \tag{11.16}$$

The evaluation of this equation yields a parasitic capacity of 14 pF for the given geometry, for a thinner ceramic layer of 0.38 mm the value rises to 23 pF. If the diode used in this example is a GaAs Schottky diode DGS10-18A, then the (voltage dependent) junction capacity is specified at 100 V with $C_j(100\ \mathrm{V}) = 22$ pF [Lin02]. This value lies in the same range as the cathode parasitic capacity of the package. Therefore, the dynamic characteristic will not be determined by the junction capacitance alone, it will rather be modified by the external parasitic capacity.

However, as stated before, the parasitic capacity parallel to the junction capacitance is determined by the series connection of C_{PK} and C_{PA}. This diminishes the problem, since the area of the anode side current track is generally much smaller than the area of the cathode track. The total parasitic capacity C_{PG} parallel to the junction capacity is given by

$$C_{PG} = \frac{C_{PK} \cdot C_{PA}}{C_{PK} + C_{PA}} \tag{11.17}$$

If C_{PA} is only 1/5 of C_{PK}, the total parasitic capacity parallel to the junction capacitance C_{PG} is only 1/6 of C_{PK}. This favorable condition is generally fulfilled in packages of the TO-family.

Now the example circuit is extended. In order to increase the blocking capability, two diodes in TO-packages are connected in series, while the geometry for each single diode is still in accordance with Fig. 11.35. The equivalent schematic circuit diagram for the extended example is depicted in Fig. 11.36.

The parasitic capacity parallel to the junction capacity C_{J1} of diode D1 is formed by C_{PK1} in series with the parallel connection of C_{PA1} and C_{PK2}. With two identical packages and the assumptions of the relation of areas discussed above, the parasitic capacity parallel to C_{J1} amounts to 6/11 C_{PK1} or 0.54 C_{PK1}.

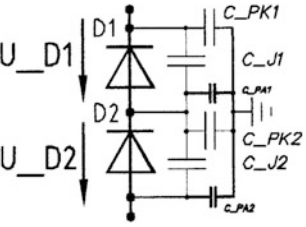


Fig. 11.36 Parasitic capacities in a series connection of two TO-220 diodes [Lin01]

The parasitic capacity parallel to C_{J2} of diode D2 is composed by the small capacity C_{PA2} in series with the parallel connection of C_{PK2} and C_{PA1}. This capacity calculates to 0.17 C_{PK1}.

Therefore, the parasitic capacities form an asymmetrical dynamical voltage divider, which generates different voltages drops over the two diodes during high frequency switching processes. This example illustrates, that parasitic capacities can lead to unfavorable effects, which do not become obvious at the first glance.

If TO-packages without internal insulations are mounted on a common heat sink, external insulation layers like polyimide foils have to be applied. These external foils also establish parasitic capacities, which exhibit even higher values, according to Table 11.3.

Especially challenging are power modules for wide bandgap devices, e.g. SiC-MOSFETs. The device package also contains, besides the known chip internal capacities C_{gs}, C_{gd} and C_{ds}, the capacities formed by the substrate insulator layers, e.g. Al_2O_3 or AlN, which are displayed in Fig. 11.37b as $C_{\sigma+}$, $C_{\sigma out}$ and $C_{\sigma-}$. Another point to consider is that $C_{\sigma out}$ is recharged with every switching event, and an undesired current is supplied into the heat sink. If the two L_{σ} and the $C_{\sigma+}$, and $C_{\sigma-}$ are unbalanced, they also generate a current into the heat sink [Fei15].

While the presented simple examples can be evaluated by analytical inspection, real multi-chip packages with a variety of chips and current tracks exhibit a much higher complexity, which makes it almost impossible to investigate by an analytical approach.

The situation is actually even more complex, since parasitic resistances, inductances and capacitances, as well as the junction capacities of the power devices have to all be considered at the same time. Their interaction during dynamic switching processes can form resonant circuits, which can cause oscillations [Gut01]. This will be discussed in Chap. 14. Today, software tools like the Fast-Henry algorithm allow simulating the electrical characteristics of complex 3-dimensional systems in detail. The analysis and optimization of power modules with respect to parasitic effects is possible and necessary to increase the reliability of power electronic systems.

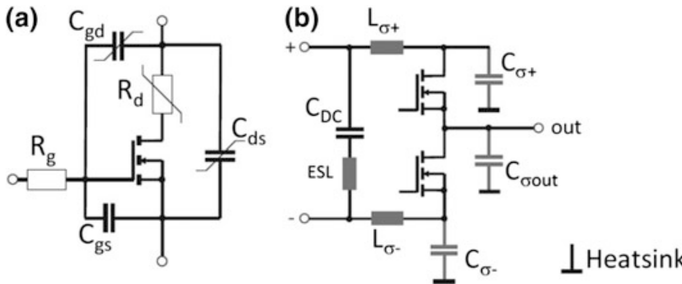


Fig. 11.37 Parasitic elements (**a**) In the chip (**b**) Switching cell parasitics. Figures from [Fei15]

11.6 Advanced Packaging Technologies

Power semiconductor packaging has become a key technology for the progress of power electronic devices. There are four basic challenges to be met:

1. The current density in power devices is continuously increasing. Even today, the package related voltage drop in a power module accumulates to a considerable percentage of the total voltage drop at nominal current. Thus, improved architectures with a reduced electrical resistance of the load current paths are required.
2. The increasing power density requirement of advanced power electronic applications enhances the power density per unit area. This development demands progressive technologies to extract the heat generated in modern power module designs.
3. The physics of silicon semiconductor devices allows maximum junction temperatures up to 200 °C for selected applications. It can be expected, that MOSFETs, IGBTs and freewheeling diodes with a voltage rating of 600 V can be operated up to a maximum junction temperature T_j = 200 °C after the necessary improvements of leakage current levels and of the reliability of the passivation. Wide bandgap devices on the basis of SiC and GaN are capable of even higher operation temperatures. In consequence, the reliability under extended temperature swings and extended maximum temperature must be ensured, especially with respect to active power cycles. The established standard module architectures are not capable to meet these requirements today; new materials and interconnection technologies must be developed.
4. The parasitic inductances and capacities must be minimized or else controlled, so that they are transformed from undesirable obstacles into functional elements of power electronic circuits.

Finding a solution to these challenges is a task, which is intensively addressed by research and development groups all over the world. The 'Center for Power Electronic Systems' (CPES), a consortium of 5 universities and several industry partners in the USA, has proposed to replace the aluminum wire bonds by a copper foil, which provides a larger effective cross section at the top side chip contact [Wen01]. The interconnection between this foil and the contact metallization of the chip is achieved by a 'dimple array technique', where only localized indentations in the foil are soldered to the chip. However, this technology has so far not been implemented in a series production and the expected increased lifetime during active power cycling has not yet been demonstrated.

The integration of the cooling system into the base plate has been proposed in order to improve the heat transfer of power modules. This concept eliminates the need for thermal interface materials between the base plate and the heat sink, which accounts for a considerable contribution to the total thermal resistance. The technique of integrating the cooling system into the DBC substrate goes even further [Scz00], because it eliminates the interface between the substrate and the base plate

as well. The substrate serves as an assembly layer for the chips and as a heat sink while providing an electrical insulation, thus combining three functions in a single element. A drawback of this proposal is the comparatively small cross section of the cooling liquid channels, which is responsible for a high pressure drop in the cooling system. This makes the system vulnerable to pollution particles in the cooling system. Moreover, the reliability of such a high tuned cooling system is of crucial importance for the device operation: A transgression of the maximum heat extraction capability would lead to the formation of a vapor layer between the cooled surface and the liquid flow, which would result in an instantaneous dramatic increase in the thermal resistance of the system. The short time constant of such a highly efficient system will result in an abrupt junction temperature increase, which will damage or possibly destroy the semiconductor device. This reliability issue is common to all highly effective cooling systems, which are currently investigated on the basis of heat pipes or based on impingement cooling methods.

11.6.1 Silver Sintering Technology

A task of fundamental importance is the accomplishment of a sufficient power cycling lifetime at high maximum junction temperatures T_j. A very promising approach to reach this goal is the 'low temperature joining technology' (LTJ). In this process, which is a diffusion sintering technique, a powder of silver particles is placed between the two surfaces to be joined. These surfaces require noble metal surface platings. An organic protection layer inhibits the silver particles to avoid diffusion of particles prior to the joining process. A heating process to approximately 250 °C during the application of a high pressure to the sinter interface dissolves this protective coating and activates the diffusion of the silver particles. This results in a densification of the powder layer to a porous rigid interconnection layer of high reliability [Mer02].

The properties of this interconnection layer are superior to solder interfaces in all parameters. The specific thermal conductivity of the sinter layer can be as high as 220 Wm^{-1} K^{-1} and is therefore almost a factor of 4 times higher than the thermal conductivity of a conventional SnAg3.5 solder layer. Together with a characteristic layer thickness of <20 µm, the sinter technology exhibits a reduced thermal resistance between the chip and the substrate compared to conventional solder layers of typically >50 µm thickness. The electrical conductivity is also improved due to the low specific electrical resistance of silver.

However, the major advantage of the silver diffusion sinter interface is the high melting temperature of the interconnection. This advantage can be illustrated by the concept of homologous temperature. Mechanical engineers use this concept to evaluate the reliability of an interconnection under mechanical stress. The homologous temperature is the ratio of the operation temperature divided by the melting temperature of the material in absolute temperature. Figure 11.38 displays the

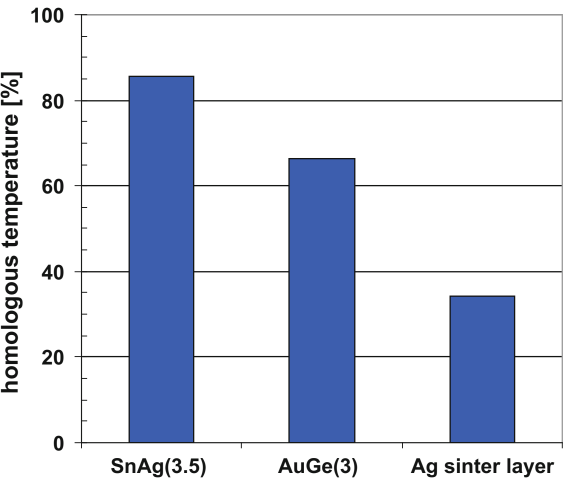


Fig. 11.38 Homologous temperature for a conventional solder interface SnAg(3.5) $T_{liquidus}$ = 221 °C, a high melting solder interface AuGe(3) $T_{liquidus}$ = 363 °C and the silver diffusion interface Ag $T_{liquidus}$ = 961 °C for an operation temperature of 150 °C

homologous temperature for a conventional SnAg(3.5) solder interface, a high temperature AuGe(3) solder interface and the silver diffusion interface, assuming an operation temperature of 150 °C.

Mechanical engineers consider interconnections operated below 40% of the homologous temperature as mechanically stable, between 40 and 60% as operated in the creep range, sensitive to mechanical strain, and above 60% as unable to bear engineering loads. Figure 11.38 illustrates clearly, that even solder interfaces with a liquidus temperature of 363 °C have a limited reliability for an operation temperature of 150 °C, while the silver diffusion technology can be expected to be reliable under mechanical stress.

Another advantageous feature of the silver diffusion technology, which might be overlooked at the first glance, is the absence of a liquid phase during the connection process. In a solder process, the solder interface passes through a liquid phase, while the temperature exceeds the liquidus temperature. During this phase of the solder process, the chip swims on a liquid film with the consequence of a series of fundamental problems:

- The chip might shift or turn out of its desired position. Solder jigs or solder stop layers are necessary to minimize this effect. Both countermeasures require a considerable margin, so that the position accuracy in a solder process is limited.
- The solder layer can exhibit a wedge-shape thickness distribution due to a variation of the surface wettability with considerable impact on the thermal resistance and thus on the reliability of the solder interface.
- Solder voids cannot be eliminated completely in an industrial series production.

Since the silver diffusion technology does not comprise a transition through a liquid phase, these problems well known from solder technologies are eliminated. In a well-controlled silver diffusion process, the chips are perfectly aligned and the interface has a homogeneous thickness without any large scale voids.

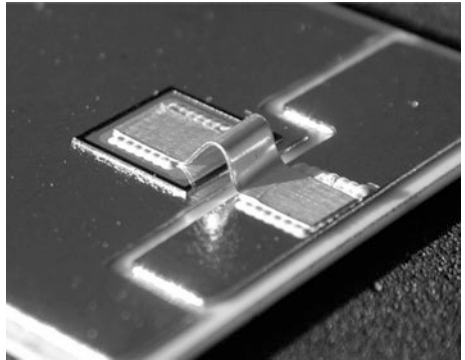


Fig. 11.39 Silver diffusion sinter technology applied to the bottom and top side chip contact, eliminating the traditional solder interface and replacing the wire bonds with a silver foil. *Source* TU Braunschweig

The diffusion sinter technology was adapted to the assembly of modern power devices like IGBTs, MOSFETs and freewheeling diodes in the middle of the 90s [Kla96]. Further process improvements verified that the simultaneous connection of multiple (different) power chips can be achieved in a single process step, which makes this technology compatible with modern series production [Scn97]. In 2008 the first commercially available series power module was introduced, which contains not a single solder interface [Scn08]. This module design combines the silver diffusion technology with the pressure system technology and spring contacts.

Experimental results confirm the expected high reliability under extreme power cycles. A power cycling test with ΔT_j = 130 K survived 30,000 cycles, which exceeds the estimated lifetime for classical base plate modules derived from an extrapolation of the LESIT curve (Eq. (12.2)) by more than a factor of 20 [Amr05]. This technology seems to be very promising even for maximum operation temperatures up to 200 °C – as was investigated in active power cycling test with ΔT_j = 160 K [Amr06] – which allows to extend the application of power modules to challenging environments, e.g. in the motor compartment of hybrid automobiles.

The silver diffusion sinter technology has the potential of replacing even the wire bonds by connecting a silver foil to the top side chip contact as shown in Fig. 11.39. This eliminates another weakness in the classical module architecture: the aluminum wire bond. This improvement reduces the parasitic resistance and inductance of the top side chip contact and further enhances the power cycling reliability [Amr05].

11.6.2 *Diffusion Soldering*

Another method to improve the reliability of the chip-to-substrate interconnection is the diffusion soldering technique. The principle is based on a process called 'Transient Liquid Phase Bonding', which has been applied in industry since many years mostly for Ti alloys [Mac92]. By applying this principle to SnCu alloys the homologous temperature of the chip interconnection can be significantly increased.

In a first step a thin eutectic alloy of Sn and Cu is liquefied at a temperature of 227 °C. The surfaces of the chip and the substrate are equipped with Cu layers so that copper is diffusing into the liquid solder and will form intermetallic phases, see Fig. 11.40. Close to the copper surfaces a copper rich phase Cu_3Sn with a melting temperature of 676 °C will form followed by a layer of a Cu_6Sn_5 phase with less copper content and a melting point of 415 °C. In the following diffusion step a high temperature will be maintained and the intermetallic phases will grow towards the center of the solder layer and progressively consume the Sn content in the eutectic alloy. Finally, the fronts of the Cu_6Sn_5 phases growing from both sides will meet in the center and the original solder alloy will be almost completely transformed into intermetallic phases with a melting point of at least 415 °C as shown in Fig. 11.41.

To maintain a reasonable diffusion time of several minutes, the bondline thickness must be in the range of 10 µm. This restricts the acceptable tolerances for the chip and the substrate surfaces. To mitigate this limitation a modification of this process was proposed: By dispersing Cu particles in the eutectic solder alloy, these particles could serve as additional sources for the diffusion of copper so that much thicker interfaces could be realized without increasing the diffusion time. The problem of this process however is the homogeneous dispersion of the Cu particles. If these Cu particles agglomerate in the paste of eutectic alloy, the intermetallic

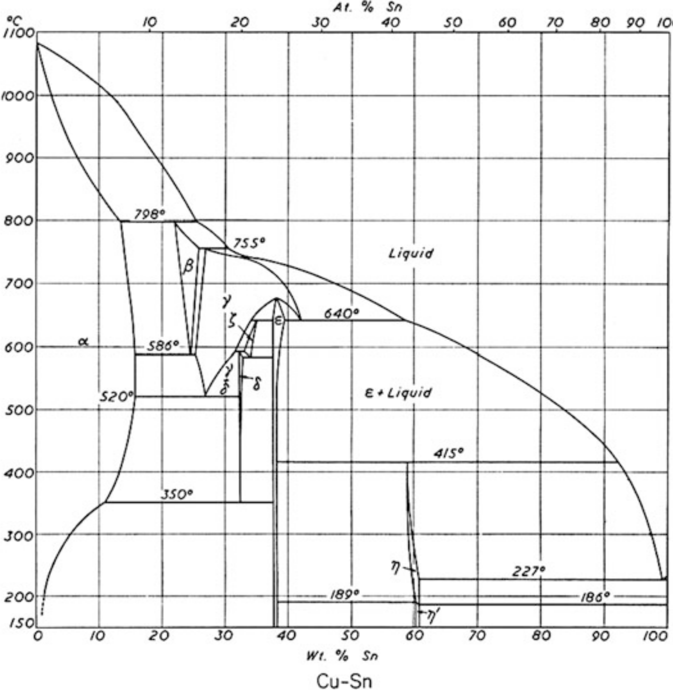


Fig. 11.40 Equilibrium phase diagram for Sn–Cu alloys [Smi76]

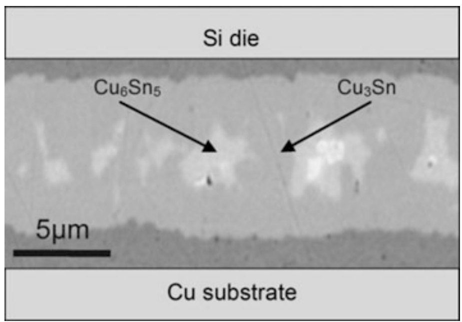


Fig. 11.41 REM image of a cross section of the interconnection obtained by diffusion soldering [Gut10]

phase grow will not be evenly distributed. In the worst case, agglomerated Cu particles can even form vertical pillars between the chip and the substrate and thus produce voids in the interconnection layer.

However, if chips with a Cu backside metallization are available and the specific process requirements a met, diffusion soldering can establish chip-to-substrate interconnections with a melting point of 415 °C and above, while maintaining the process temperature in the range of 250 °C. The increase of lifetime under thermo-mechanical stress is expected lower than the increase for Ag diffusion sintering, but diffusion soldering avoids the high pressure in the range of 40 MPa required for silver sintering.

Diffusion soldering in industrial scale was presented by [Gut10, Gut12]. Also a new solder process substrate to base plate is reported in [Gut10]. The new solder layer contains vertical intermetallic phases with high melting temperature (Fig. 11.42).

At temperature swings and the given CTE mismatch between the neighboring layers, it will be difficult for a crack to cross laterally through the harder vertical intermetallic phases. Therefore, crack propagation is strongly slowed down. This was confirmed by a power cycling test with high temperature swing of the base plate of ΔT of 100 K with aim to stress the interconnection base plate to substrate. The expected increase in power cycling capability by a factor of 2.5 to 3 was confirmed [Hen10].

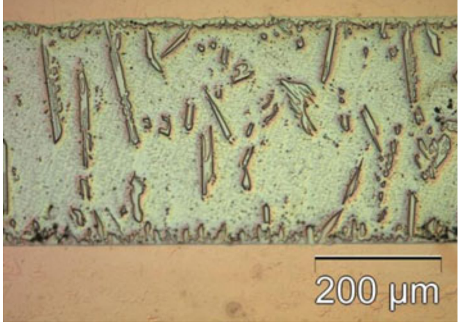


Fig. 11.42 Substrate solder layer containing vertical intermetallic phases. Figure from [Gut10]

11.6.3 Advanced Technologies for the Chip Topside Contact

Aluminum wire bonds with diameters between 300 and 500 μm are established for more than 30 years as topside chip contact in standard MOSFET or IGBT power modules (see Fig. 11.13). A wedge-wedge ultrasonic bonding machine can establish a single wire connection from the top side chip metallization – conventionally an approx. 4 μm Al layer – to the substrate current trace within fractions of a second. A single Al wire bond of 10 mm length can conduct currents of 18 and 50 A for a diameter of 300 and 500 μm, respectively, due to the high heat extraction capability of the standard power module mounted on an efficient heat sink.

The wire bond process is highly flexible and can be adapted to different layouts and design changes by a simple software update. This great advantage is accompanied by several aspects of concern: The wire bond contributes to parasitic resistance and inductance of the module, it is limited in current capability and lifetime – as will be discussed in more detail in Sect. 12.7.2 – and the wires will act as fuses at extreme over-currents, which will result in arcing and in consequence in an explosion with often considerable damage in the power electronic system.

In the first decade of this century, two driving factors have increased the demand for an improvement of the chip topside contact. Firstly, the increasing current density of the devices requires an increasing number of Al wire bonds. For modern low-voltage MOSFETs it is already very difficult to place the required number of Al wire bonds on the source contact. Secondly, the extension of the maximum junction temperature from 150 to 175 °C and in the future to even 200 °C requires a lifetime increase of the modules. The increased operating temperature range can only be exploited, if the lifetime is increased accordingly. As a rule of thumb, each 25 °C increase in operation temperature requires a lifetime increase by a factor 5, so that the lifetime of the standard module technology must be enhanced at least by a factor 25.

Cu Bond Wires

Cu bond wires were introduced by Infineon [Gut10]. Cu exhibits high yield strength of 140 MPa compared to Al with 29 MPa. Further, the CTE of Cu is with 16.5 ppm/K much better matched to Si (2 ppm/K) than Al with 23.5 ppm/K. Cu has higher electrical and thermal conductivity than Al. The power losses P_{bond} in a bond wire with length l are given by

$$P_{bond} = R_{bond} \cdot I^2 = \rho \cdot \frac{l}{A} I^2 \tag{11.18}$$

For the same wire length and diameter, the current for same P_{bond} increases by $\sqrt{\rho_{Al}/\rho_{Cu}}$. Special soft Cu bond wires were developed by soft annealing, leading to low deformation resistance [HER15], the specific resistance increases slightly compared to bulk Cu and is given as ρ_{Cu} = 1.8 μΩ cm. Cu has compared to Al a higher melting temperature. Finally, the fusing current per wire is increased, in

[Her12] an increase by a factor of 1.25–1.27 is given when comparing Cu to Al. For same wire diameter and length, a 50 K lower temperature in the loop of the bond wire is calculated for Cu compared to Al [Sie10].

A cross section of a Cu bond wire and a substrate with Cu bonds is shown in Fig. 11.43.

Cu wire bonding requires, compared to Al wire bonding, higher bonding forces, significant higher ultrasonic power and modified cutting tool concepts. A comparison of bond parameters is made in Fig. 11.44. Cu bonding is not compatible with Al-metallized device surface, since the high power and applied force could cause cracks in the semiconductor body. Additionally, IGBTs and MOSFETs have a cell structure containing a thin SiO_2 isolation layer between gate-emitter resp. gate-source, which is sensitive against too high local mechanical load. Cu wire bonding requires a Cu metallization of sufficient thickness, 5–50 μm are mentioned in [Her12]. Cu metallization is already known from the fabrication of microelectronic devices in which, for the submicron structures, high current densities occur. The metallization needs diffusion barriers to silicon to avoid contamination. The metallization technology was successfully transferred from microelectronics to power electronics.

Modules with Cu wire bonds can reach very high power cycling performance, the bond wire is no longer limiting the lifetime.

Cu exhibits compared to Al a higher thermal capacity. Therefore, it is of advantage to make the Cu layer in a significantly higher thickness than 4 μm as usual for Al metallization. However, due to the thermal mismatch to silicon this can

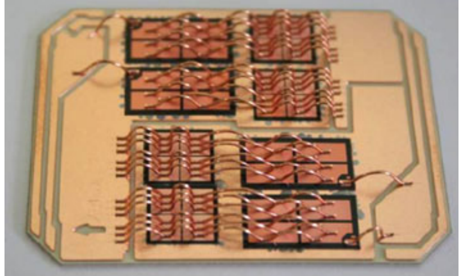


Fig. 11.43 Cross section of a Cu bond foot (top), view of a substrate equipped with dies with copper metallization and Cu bond wires. Figure from [Gut10]

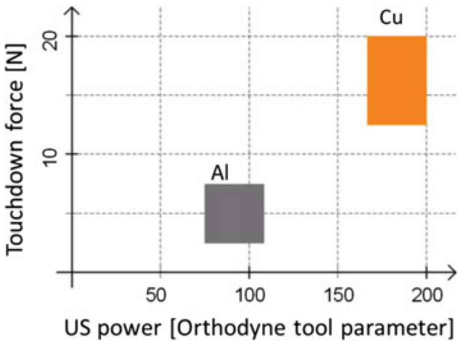


Fig. 11.44 Comparison of bond parameters for Al and Cu. Figure adapted from [Bec15]

lead to wafer bow, which is an obstacle for fine pattern of metallization. The challenge is more exposed for thin wafer technology. And in same time, 300 mm wafer technology for IGBTs is coming up. Therefore, several trade-offs have to be taken into account.

Bond Buffer Technology

Since only some dies from special manufacturers are available with Cu metallization, the bond-buffer technology was introduced by Danfoss. Here, a thin Cu plate ("bond buffer") is attached on the topside of the die with silver sintering. Figure 11.45 shows the structure of the so-called "DBB-technology". Beside silver sintering on the topside, this interconnection is also applied at the bottom side and replaces the chip solder layer.

The process just requires a noble-metal surface of the chip metallization. If this is applied, different dies from several manufacturers can be used. The Cu bond process can be executed with desired ultrasonic power since the cell structure is effectively protected. High power cycling capability is to be expected.

The structure gains advantage in thermal resistance and thermal impedance. The bottom side thin silver sinter layer improves the thermal resistance compared to solder layers. The bond buffer and Cu bond wires introduce additional thermal capacitance close to the chip which reduces Z_{th} for short time loads, especially at pulse duration of 10 ms and below [Rud12].

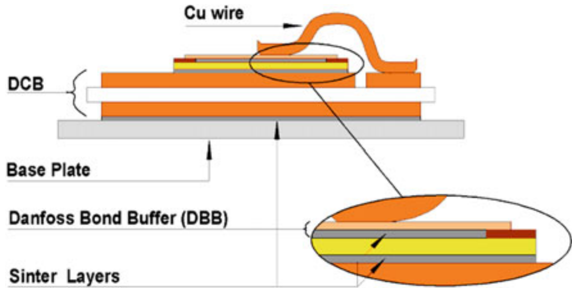


Fig. 11.45 Bond buffer technology. Figure from [Rud12]

The thickness of the Cu buffer, however, cannot be increased too high, otherwise cracking caused by high thermal mismatch in the topside metallization will occur during temperature cycling or power cycling. In contrast to Cu-metallization, the bond buffer foil will not cover the complete active area of the device. At very short overload events, such as short circuit with duration of 10 μs or lower, these cells have no energy storage buffer on top and will fail first.

Al-clad Cu Wire Bonds

With Al-clad Cu wires it is intended to combine the superior electrical and thermal characteristics of Cu with the advantage of mass production of the Al wire bond process [Dal06]. Figure 11.46 shows an example of a 300 mm wire. The Cu core diameter amounts to 230–250 μm, the Al coating layer to 25–35 μm.

During the bonding process, an Al to Al ultrasonic welding interconnection is established, therefore similar failure mechanisms under repetitive thermal load as for pure Al wires could be expected. However, the effective CTE of the wire bond is reduced and together with the lower electrical resistivity and the enhanced thermal conductivity, the crack initiation and propagation should be reduced. In fact, a significant increase of the power cycling capability was shown in [Sct12]. Figure 11.47 is showing the results.

Two types of cladded wires are compared in Fig. 11.47 with homogeneous Al bond wires. The type “hard” with hard Cu core shows the best power cycling results. It can be used for diodes with Al topside metallization. The type “soft” with soft-annealed Cu core lowers the risk of damage in sensitive devices with cell structures on topside and can be used for IGBTs. The power cycling tests in Fig. 11.47 have been executed with a “solder-free” package where Ag sinter layers are applied which are not limit the power cycling capability.

Even if the number of cycles to failure with the soft-annealed Cu is lower compared to “hard” Cu, it is a significant progress.

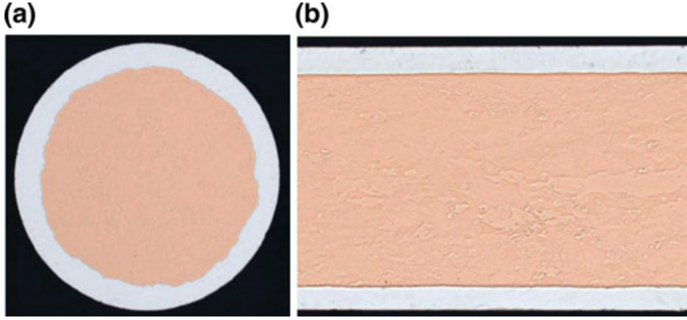


Fig. 11.46 Cross section of an 300 μm Al-clad Cu wire bond in (**a**) radial and (**b**) longitudinal direction. Figure from [Sct12]

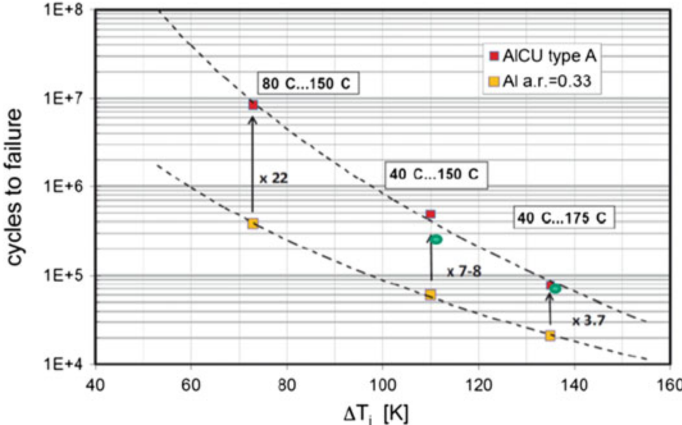


Fig. 11.47 Comparison of power cycling results of bond wires: *yellow square* Al-wires (lower line), *red square* Al-clad, hard Cu, *green circle* Al-clad, soft Cu

11.6.4 Improved Substrates

Aluminum oxide and Aluminum nitride substrates are established. With improvement of the die interconnection layer and the topside contact technology, the substrate will become the next limiting factor. Standard and new ceramic substrates are compared in Fig. 11.48. The parameters are given in Table 11.6.

The so-called HPS-substrate (high power substrate) consists of Al_2O_3 with 9% ZrO_2. Its standard thickness amounts to 0.32 mm [ELE09]. In Fig. 11.48 the temperature cycling capability of the different substrates is compared. AlN shows the weakest cycling capability, since in AlN conchoidal fractures in the ceramics occur [Dup06]. Al_2O_3 reveals a superior cycling lifetime, next HPS exhibits double number of cycles compared to Al_2O_3. Si_3N_4 is possible in two production processes. One is coating to allow application of the usual direct copper bonding process, the other is active metal brazing (AMB) which is a type of high temperature active soldering using AgCu alloys containing 1.5% Ti. Already in the coated version, Si_3N_4 shows a very high stability under temperature cycling. In the AMB version, it offers an excellent temperature cycling capability. For cycles from −30 to 180 °C, 780 cycles were executed without failure [Dup06]. In the results shown in Fig. 11.48, 5000 cycles were applied and no weakness occurred. Therefore, with this substrate even an increased thickness of Cu layers is possible, e.g. 400 or 500 μm. This leads to a high current capability of the substrate, and also to a very good head spreading effect [KYO05].

It is pointed out in [Miy16] that fracture toughness is the decisive parameter for the high stability of Si_3N_4 AMB, and the temperature cycling capability does not depend on flexural strength.

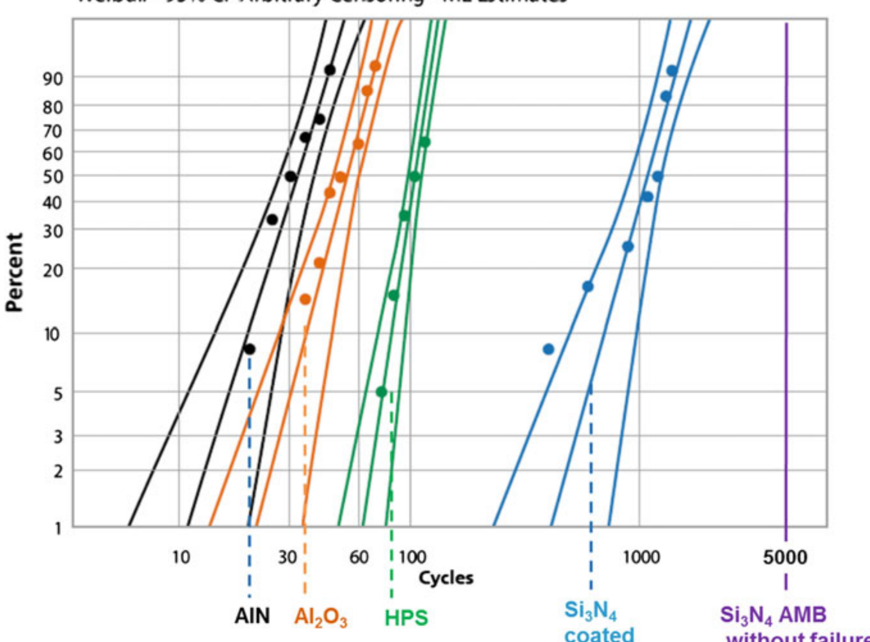


Fig. 11.48 Comparison of temperature cycling capability of ceramic substrates with both-side Cu layers. Temperature swing −55/+150 °C. Figure from [Goe12]

Table 11.6 Characteristic parameters of the substrates in Fig. 11.48

	Thermal conductivity (W/mm/K)	Thermal expansion (10^{-6}/K)	Tensile strength (MPa)	Thickness (mm)
Al_2O_3	0.024	6.8	400	0.38
HPS	0.026	7.1	600	0.32
Si_3N_4	0.06	3.4	800	0.32
AlN	0.17	4.7	270	0.63

Beside Cu-metallized substrates, shown in Fig. 11.48, also Al-metallized substrates are possible. The Al layers are covered with a Ni film or a Ni–Au film to allow for soldering. Al has a much lower yield stress (30–35 MPa) compared to Cu (85–100 MPa), and also a much flatter plastic characteristics [Dup06]. A high temperature cycling capability for cycles between −55 and +125 °C was shown. The disadvantage is a higher thermal resistance due to the lower thermal conductivity of Al. Further, it must be taken into account that Al layers show the effect of reconstruction, which can be an obstacle for power cycling capability. Reconstruction of metal layers will be discussed in Sect. 12.7.2

11.6.5 Advanced Packaging Concepts

The advanced packaging concepts implement the new technologies introduced in previous sections into products. Several solutions focus on different aspects of package improvement.

Mold Modules With Strong Heat Spreading

The Mitsubishi mold module has a top side contact soldered to a Cu lead frame [Mot12]. The architecture shown in Fig. 11.49 is named as DLB technology (Direct Lead Bonding). The module is encapsulated with epoxy-based mold compound as known from the TO-family.

As potential separation a so-called TCIL-layer (Thermally Conductive Insulation Layer) is used, which contains a thin laminated Cu-foil on the bottom side. Since the thermal conductivity of organic insulators is significantly lower than that of ceramics, a thick Cu layer is located above the insulation layer for lateral heat spreading, increasing the area for vertical heat transport. It is claimed that the thermal resistance is comparable to modules using AlN ceramic substrates [Mot12].

The Fuji Green Line technology is based on a Si_3N_4 substrate which is connected on both sides with a 1 mm Cu layer, see Fig. 11.50. The thick Cu layers act as effective heat spreaders, and the thermal performance is supported by the high thermal conductivity of Si_3N_4. The thermal impedance can be strongly reduced at medium time constants [Hor14].

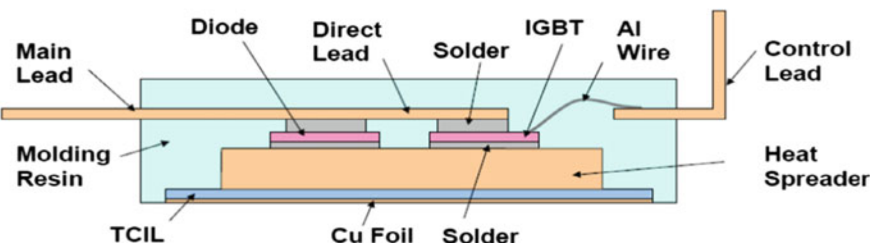
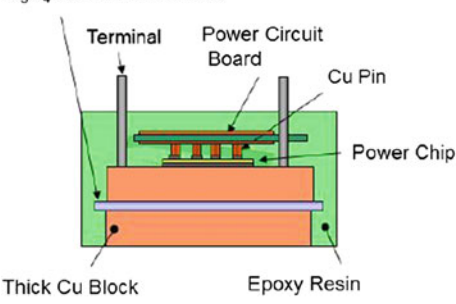


Fig. 11.49 Mitsubishi mold module, DLB technology. Figure from [Mot12]

Fig. 11.50 Fuji Green Line module. Figure from [Hor14]



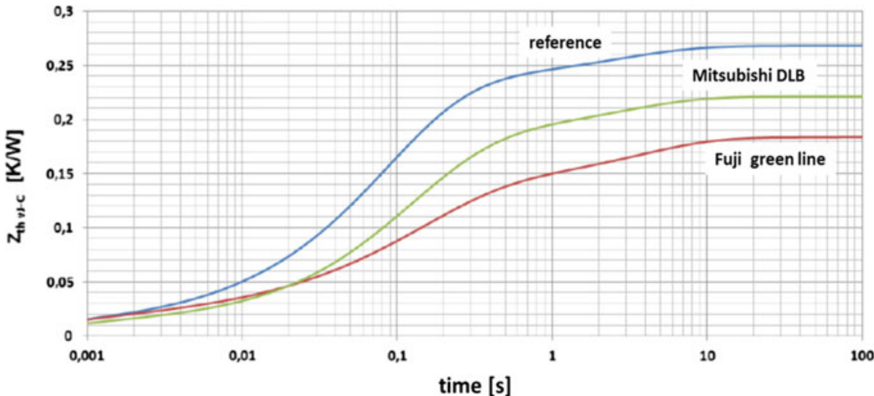


Fig. 11.51 Comparison of thermal impedance for standard modules and modules with strong heat spreaders. Figure from [Bec15]

On top side the semiconductor die is connected with Cu pins to a PCB, the whole module is encapsulated with an epoxy based mold compound.

An FEM-simulation in [Bec15] compares both package concepts using the same silicon die 10.6 mm × 10.6 mm. A thermal foil of 70 μm (2 W/m/K) to the heat sink with heat transfer coefficient of 5000 W/(m^{2}K) is assumed in all cases. The reference is the Al_2O_3 based module from Table 11.1, with 0.38 mm Al_2O_3 ceramics, chip solder of 100 μm, soldered with thick solder 500 μm to a base of 3 mm (Fig. 11.51).

At small time constants between 0.01 and 0.1 s, there is a strong improvement of the new packaging concepts. Also an improvement in the static thermal resistance is visible, most prominent in the Fuji green line concept. This is a result of the high thermal conductivity of the Si_3N_4 ceramics with the thick Cu layers on both sides. Also the strong heat spreading by two 1 mm Cu layers above the low thermal conductive layer to the heat sink is very effective to transfer the heat flow.

The SKiN Technology

The SKiN technology was presented by Semikron in [Bel11]. As shown by the schematic cross section in Fig. 11.52, the wire bonds are replaced by a flexible circuit board which is connected to the top metal layer of the chips by an Ag diffusion sinter interface. The chip-to-substrate as well as the substrate-to-heat sink interconnection is formed by diffusion sintering. This technology not only eliminates the solder layers and wire bonds known as the weak elements in the reliability chain of classical module design, it also replaces thermal interface materials (TIM) required between the module and the heat sink in the classical design by a high reliable diffusion sinter connection. All interconnections are realized by silver sinter technology: die topside, die bottom side and substrate to cooling plate.

On the bottom side a pin fin cooling plate is used, named as “heatsink” in Fig. 11.52. The topside chip contact layer is part of a flexible circuit board

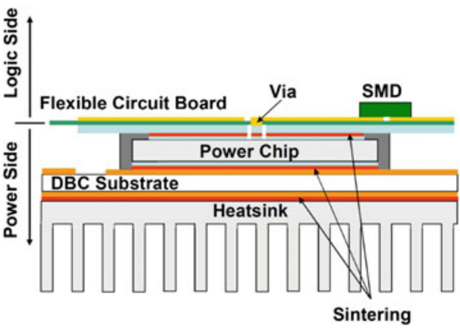


Fig. 11.52 SKiN technology. Figure from [Scn12]

consisting of two metallization layers, one on each side of a polyimide foil. The bottom metallization transports the load current and is called 'power side'. The selected layer thickness is in the range of 100 μm. For the top side metallization referred to as 'logic side', a layer thickness of 35 μm is sufficient. This layer distributes the control and sensor signals. Vias in the flexible circuit board are connecting the gate contacts on the power side to the logic side as shown in Fig. 11.52. Additional SMD components can be assembled on the logic side in close proximity of the chips.

The SKiN technology has demonstrated a high power cycling capability. With more than 3 million cycles at $\Delta T_j = 70$ K, it enhanced the power cycling lifetime by a factor of 40 related to state-of-the-art industrial modules.

While the afore presented packaging concepts are designed to improve the packaging of Si devices, they are not sufficient to exploit the capabilities of progressive wide-bandgap devices. Fast switching SiC and GaN devices allow to considerably increase the switching frequency, but this requires a substantial reduction of the package parasitic induction. Investigations with 1200 V SiC-MOSFETs implemented in industrial power module packages have shown, that either pronounced voltage oscillations reduce the maximum DC link voltage or that higher gate resistors are required to reduce the switching speed of the devices.

A proposal for a package design with lower parasitic inductance is the strip line concept from Infineon [Bor13]. The arrangement of multiple pairs of DC+ and DC− contact in close proximity results in a module stray inductance below 7nH (Fig. 11.53).

An advancement derived from the SKiN technology was applied to build a prototype module with a commutation stray inductance of about 1.4 nH [Bel16]. As shown in Fig. 11.54, the flex layer concept was extended to conduct DC− on the topside and DC+ on the bottom. Spring elements are implemented to establish a low profile contact between the flex layer and a PCB or laminated bus bar.

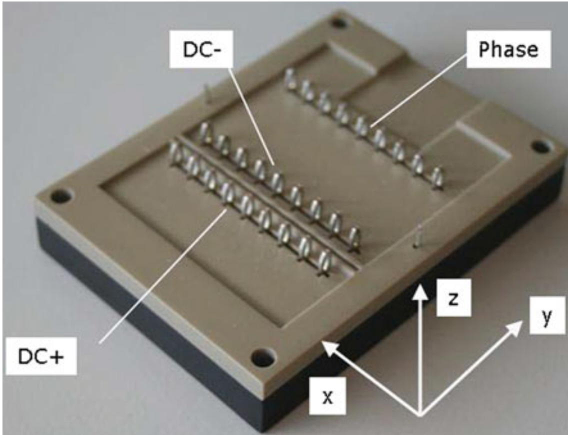


Fig. 11.53 Low inductive package with strip line concept. Figure from [Bor13]

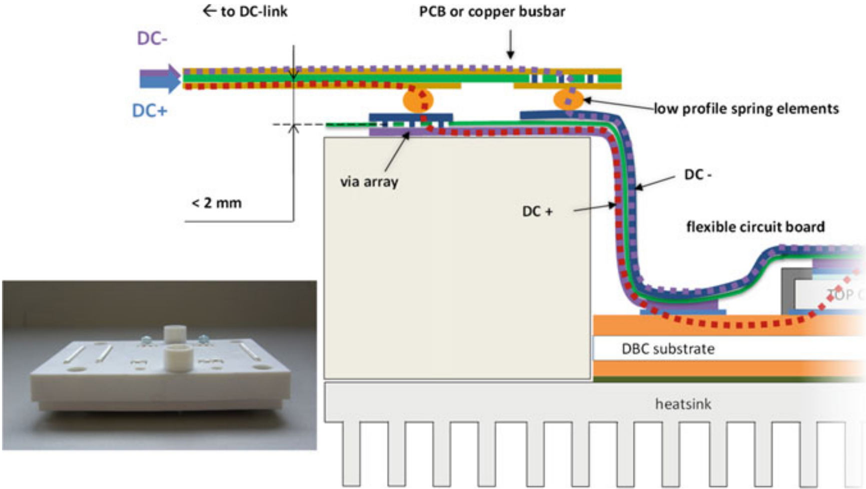


Fig. 11.54 Low inductive package derived from SKiN technology. Figure from [Bel16]

As mentioned at the beginning of Sect. 11.6, there are four basic challenges to be met:

- Low electrical resistance
- Efficient heat extraction
- High-reliable interconnections
- Minimized parasitic inductances and capacities.

New architectures face these challenges and show significant progress. The mold modules with strong heat spreading (Figs. 11.49 and 11.50) significantly improve thermal impedance and thermal resistance junction to case. The modules with pin-fin base plate (Fig. 11.52) significantly improve the thermal resistance junction to ambient [Hen10]. In the .XT technology from Infineon, Cu bond wires and diffusion soldering is applied. Further solutions are in research and development. The presented systems in this overview should be taken as examples and not as a complete list of innovative concepts.

For evaluating potential improvements in progressive packaging technologies, it is inevitable to consider the complex interdependency of single optimizations with respect to the performance of the complete system. Especially the impact of parasitic influences cannot be omitted. This aspect will be addressed in more detail in Chap. 15.

References

[Amr05] Amro, R., Lutz, J., Rudzki, J., Thoben, M., Lindemann, A.: Double-sided low-temperature joining technique for power cycling capability at high temperature. In: Proceedings of EPE, Dresden (2005)

[Amr06] Amro, R., Lutz, J., Rudzki, J., Sittig, R., Thoben, M.: Power cycling at high temperature swings of modules with low temperature joining technique. In: Proceedings of ISPSD, pp. 1–4, Naples (2006)

[Bec15] Becker, M.: Neue Technologien für hochzuverlässige Aufbau- und Verbindungstechniken leistungselektronischer Bauteile. PhD-Thesis, Chemnitz (2015)

[Bel11] Beckedahl, P., Hermann, M., Kind, M., Knebel, M., Nascimento, J., Wintrich, A.: Performance comparison of traditional packaging technologies to a novel bond wire less all sintered power module. In: Proceedings of PCIM Europe, pp. 247–252, Nuremberg (2011)

[Bel16] Beckedahl, P., Buetow, S., Maul, A., Roeblitz, M., Spang, M.: 400 A, 1200 V SiC power module with 1nH commutation inductance. In: Proceedings of CIPS, pp. 365–370, Nuremberg (2016)

[Bor13] Borghoff, G.: Implementation of low inductive strip line concept for symmetric switching in a new high power module. In: Proceedings of PCIM Europe, pp. 185–191, Nuremberg (2013)

[But14] Butron Ccoa, J.A., Strauß, B., Mitic, G., Lindemann, A.: Investigation of temperature sensitive electrical parameters for power semiconductors (IGBT) in real-time applications. In: Proceedings of PCIM Europe, pp. 456–464, Nuremberg (2014)

[Dal06] Dalin, J., Knauber, A., Reiter, R., Wesling, V., Wilde, J.: Novel aluminium/copper fiber-reinforced bonding wires for power electronics. In: Proceedings of Electronics System integration Technology Conference (ESTC), pp. 1274–1278, Dresden (2006)

[Dup06] Dupont, L., Lefebvre, S., Khatir, Z., Bontemps, S.: Evaluation of substrate technologies under high temperature cycling. In: Proceedings of CIPS, pp. 63–68, Naples (2006)

[EFU99] eFunda engineering fundamentals, http://www.efunda.com/materials/

[ELE09] Electrovac/Curamik: ZrO_2 doped alumina DBC substrates – Cur HPS, Datenblatt, unreleased draft (2009)

[Fei15] Feix, G., Hoene, E., Zeiter, O., Pedersen, K.: Embedded very fast switching module for SiC power MOSFETs. In: Proceedings of PCIM Europe, pp. 104–110, Nuremberg (2015)

[Fel09] Felsl, H.P.: Silizium-und SiC-Leistungsdioden unter besonderer Berücksichtigung von elektrisch-thermischen Kopplungseffekten und nichtlinearer Dynamik. PhD-Thesis, Chemnitz (2009)

[Goe12] Goetz, M., Lehmeier, B., Kuhn, N., Meyer, A.: Silicon nitride substrates for power electronics. In: Proceedings of PCIM Europe, pp. 672–679, Nuremberg (2012)

[Gut01] Gutsmann, B., Silber, D., Mourick, P.: Kolloquium Halbleiter-Leistungsbauelemente und ihre systemtechnische Integration, Freiburg (2001)

[Gut10] Guth, K., Siepe, D., Görlich, J., Torwesten, H., Roth, R., Hille, F., Umbach, F.: New assembly and interconnects beyond sintering methods. In: Proceedings of PCIM Europe, pp. 232–237, Nuremberg (2010)

[Gut12] Guth, K., Oeschler, N., Boewer, L., Speckels, R., Strotmann, G., Heuck, N., Krasel, S., Ciliox, A.: New interconnect technologies for power modules. In: Proceedings of CIPS, pp. 380–384, Nuremberg (2012)

[Ham98] Hamidi, A.: Contribution à l'étude des phénomènes de fatigue thermique des modules IGBT de forte puissance destines aux application de traction. PhD-Thesis, Grenoble (1998)

[Hec01] Hecht, U., Scheuermann, U.: Static and transient thermal resistance of advanced power modules. In: Proceedings of PCIM Europe, pp. 299–305, Nuremberg (2001)

[Hen10] Hensler, A., Lutz, J., Thoben, M., Guth, K.: First power cycling results of improved packaging technologies for hybrid electrical vehicle applications. In: Proceedings of CIPS, pp. 85–90, Nuremberg (2010)

[Her12] Herold, C., Hensler, A., Lutz, J., Thoben, M., Guth, K.: Power cycling capability of new technologies in power modules for hybrid electric vehicles. In: Proceedings of PCIM Europe, pp. 486–493, Nuremberg (2012)

[HER15] Heraeus: Thick Copper Bonding Wire of Extreme Softness. online https://www.heraeus.com/media/media/het/doc_het/products_and_solutions_het_documents/bonding_wires_documents/fact_sheets/Factsheet_PowerCuSoft.pdf, published July 20, 2015, visited Ju1y 10, 2017

[Hof13] Hofmann, K., Herold, C., Beier, M., Lutz, J., Friebe, J.: Reliability of Discrete Power Semiconductor Packages and Systems – D^2Pak and CanPAK in Comparison. In: Proceedings of EPE, Lille (2013)

[Hor14] Hori, M., Saito, M., Hinata, Y., Nashida, N., Ikeda, Y., Mochizuki, E.: Compact, low loss and high reliable next generation Si-IGBT module with advanced structure. In: Proceedings of PCIM Europe, pp. 472–477, Nuremberg (2014)

[Jor09] Jordà, X., Perpiñà, X., Vellvehi, M., Millán, J., Ferriz, A.: Thermal characterization of insulated metal substrates with a power test chip. In: Proceedings of ISPSD, pp. 172–175, Barcelona (2009)

[Kam93] Kamon, M., Tsuk, M.J., White, J.: FastHenry: A multipole-accelerated 3-D inductance extraction program. In: Proceedings of ACM/IEEE Design Automation Conference, pp. 678–683 (1993)

[Kla96] Klaka, S.: Eine Niedertemperatur-Verbindungstechnik zum Aufbau von Leistungshalbleitermodulen, Dissertation, Braunschweig (1996)

[Kuh91] Kuhnert, R., Schwarzbauer, H.: A novel large area joining technique for improved power device performance. IEEE Trans. Ind. Appl. **27**, pp. 93–95 (1991)

[KYO05] Kyocera: Si Si_3N_4 AMB products, product presentation (2005) http://www.ivf.se/upload/pdf-filer/Arbetsomr%C3%A5den/Elektronikutveckling/KYOCERA%20Si3N4%20AMB%20Products%202005%20(R0052D).pdf

[Lap91] Lappe, R., Conrad, H., Kronberg, M.: Leistungselektronik, 2nd edn. Verlag Technik, Berlin (1991)

[Lin01] Lindemann, A.: Kolloquium Halbleiter-Leistungsbauelemente und ihre systemtechnische Integration, Freiburg (2001)

[Lin02] Lindemann, A., Friedrichs, P., Rupp, R.: New semiconductor material power components for high end power supplies. In: Proceedings of PCIM Europe, pp. 149–154, Nuremberg (2002)

[Mac92] MacDonald, W.D., Eagar, T.W.: Transient liquid phase bonding. Annu. Rev. Mater. Sci. **22**, pp. 23–46 (1992)

[Mer02] Mertens, C., Sittig, R.: Low temperature joining technique for improved reliability. In: Proceedings of CIPS, pp. 95–100, Nuremberg (2002)

[Miy16] Miyazaki, H., Iwakiri, S., Hirotsuru, H., Fukuda, S., Hirao, K., Hyuga, H.: Effect of mechanical properties of the ceramic substrate on the thermal fatigue of Cu metallized ceramic substrates. In: IEEE 18th Electronics Packaging Technology Conference (EPTC) (2016)

[Mot12] Motto, E.R., Donlon, J.F.: IGBT module with user accessible on-chip current and temperature. In: Proceedings of Applied Power Electronics Conference and Exposition (APEC), pp. 176–181, Orlando (2012)

[Mou02] Mourick, P., Steger, J., Tursky, W.: 750A 75 V MOSFET power module with sub-nH inductance. In: Proceedings of ISPSD, pp. 205–208 (2002)

[Poe04] Poech, M.H., Fraunhofer-Institut Siliziumtechnologie, Itzehoe, private communication (2004)

[Pol13a] Poller, T., D'Arco, S., Hernes, M., Lutz, J.: Determination of the thermal and electrical contact resistance of press pack housings. In: Proceedings of EPE, Lille (2013)

[Pol13b] Poller, T., D'Arco, S., Hernes, M., Ardal, A.R., Lutz, J.: Influence of the clamping pressure on the electrical, thermal and mechanical behaviour of press-pack IGBTs. Microelectron. Reliab. **53**, pp. 1755–1759 (2013)

[Rud12] Rudzki, J., Osterwald, F., Becker, M., Eisele, R.: Novel Cu-bond contacts on sintered metal buffer for power module with extended capabilities. In: Proceedings of PCIM Europe, pp. 784–791, Nuremberg (2012)

[Saw00] Sawle, A., Woodworth, A.: Innovative developments in power packaging technology improve overall device performance. In: Proceedings of PCIM Europe, pp. 333–339, Nuremberg (2000)

[Saw01] Sawle, A., Standing, M., Sammon, T., Woodworth, A.: Directfet™ – a proprietary new source mounted power package for board mounted power. In: Proceedings of PCIM Europe, pp. 473–477, Nuremberg (2001)

[Scn97] Scheuermann, U., Wiedl, P.: Low temperature joining technology – a high reliability alternative to solder contacts. In: Workshop on Metal Ceramic Composites for Functional Applications, pp. 181–192, Wien (1997)

[Scn99] Scheuermann, U.: Power module design for HV-IGBTs with extended reliability. In: Proceedings of PCIM Europe, pp. 49–54, Nuremberg (1999)

[Scn06] Scheuermann, U.: Aufbau- und Verbindungstechnik in der Leistungselektronik, in Schröder D, Elektrische Antriebe Bd. 3 – Leistungselektronische Bauelemente, 2. Auflage, Springer Berlin (2006)

[Scn08] Scheuermann, U., Beckedahl, P.: The road to the next generation power module – 100% solder free design. In: Proceedings of CIPS, pp. 111–120, Nuremberg, (2008)

[Scn09] Scheuermann, U., Schmidt, R.: Investigations on the V_{CE}(T) method to determine the junction temperature by using the chip itself as sensor. In: Proceedings of PCIM Europe, pp. 802–807, Nuremberg (2009)

[Scn12] Scheuermann, U.: Reliability of planar SKiN interconnect technology. In: Proceedings of CIPS, pp. 464–471, Nuremberg (2012)

[Scn15] Scheuermann, U.: Packaging and reliability of power modules – principles, achievements and future challenges. In: Proceedings of PCIM Europe, pp. 35–50, Nuremberg (2015)

[Sct12] Schmidt, R., Scheuermann, U., Milke, E.: Al-Clad Cu wire bonds multiply power cycling lifetime of advanced power modules. In: Proceedings of PCIM Europe, pp. 776–783, Nuremberg (2012)

[Scz00] Schulz-Harder, T., Exel, J., Meyer, A., Licht, K., Loddenkötter, M.: Micro channel water cooled power modules. In: Proceedings of PCIM Europe, pp. 9–14, Nuremberg (2000)

[Sie10] Siepe, D., Bayerer, R., Roth, R.: The future of wire bonding is? Wire Bonding! In: Proceedings of CIPS, pp. 115–118, Nuremberg (2010)

[Smi76] Smithells, C.J.: Metals Reference Book, 5th edn. Butterworths, London & Boston (1976)

[Tin15] Tinschert, L., Årdal, A.R., Poller, T., Bohländer, M., Hernes, M., Lutz, J.: Possible failure modes in Press-Pack IGBTs. Microelectron. Reliab. **55**(6), pp. 903–911 (2015)

[Wen01] Wen, S., Huff, D., Lu, G.Q., Cash, M., Lorenz, R.D.: Dimple-array interconnect technique for interconnecting power devices and power modules. In: Proceedings of CPES Seminar, pp. 75–80, Blacksburg (2001)

[Yam03] Yamada, J., Simizu, T., Kawaguchi, M., Nakamura, M., Kikuchi, M., Thal, E.: The latest high performance and high reliability IGBT technology in new packages with conventional pin layout. In: Proceedings of PCIM Europe, pp. 329–333, Nuremberg (2003)

[Zhg04] Zhang, J., Choosing the right MOSFET package. IR application note Feb 2004, http://www.eepn.com/Locator/Products/ArticleID/29270/29270.html

Chapter 12
Reliability and Reliability Testing

The reliability of power electronic devices and components has been mentioned several times in the previous chapters. It is so important, because it is a prerequisite for the performance in applications: Reliability is the ability of a system or component to perform its required functions under stated conditions for a specified period of time [SAE08]. The requested lifetime of power electronics systems is seldom below 10 years and can reach up to 30 years.

The reliability tests that will be discussed in detail in the following sections have been developed and established during more than 30 years of experience on power packages with implemented Si device technology. With the growing maturity of wide bandgap devices like SiC and GaN, it seems obvious that the same test procedures should be applied to modules equipped with these novel device technologies, since applications will require a comparable reliability level. However, wide bandgap devices exhibit specific differences with respect to the established Si devices, which have to be taken into account in reliability testing. These special requirements will be discussed in the respective sections.

12.1 The Demand for Increasing Reliability

Applications of power electronic devices face an increasing requirement for high reliability for several reasons:

- Power electronics faces a continuous demand for an increase in power density – often expressed in terms of controlled power per unit volume. This demand results in an increasing current density in power chips and in an increasing package density in power modules, with the consequence of higher temperatures and temperature gradients in the package.
- New fields of applications define more severe ambient conditions for the power packages, i.e. automotive hybrid traction systems, in which the combustion

© Springer International Publishing AG 2018
J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8_12

engine cooling system, featuring cooling liquid temperatures up to 120 °C, extracts the losses of power electronic components. This requirement has led to an extension of the operation temperature range from a maximum T_j = 150 °C to a specification limit of 175 °C.
- The number of interdependent frequency inverters is continuously growing in some areas of industrial automation. In automobile assembly for example, several hundred process steps are linked together in a single production line, each has to remain functional to keep the line running. For the same operational availability of the assembly line as for a single inverter, this results in mean time between failure (MTBF) values for each inverter divided by the number of interlinked inverters, which reduces the acceptable failure rate easily by orders of magnitude.

These general trends have been the boundary conditions for the progress of power electronics packages in the past and will continue to do so in the future.

It is obviously impossible to test the reliability of power modules under field conformal stress conditions, because these tests would last as long as the expected service lifetime in the field – 10 to 30 years. Thus, manufacturers of power modules have developed a canon of accelerated test procedures during the last 30 years, which are derived from experience and which are considered as a base line for the product qualification to verify the expected functionality over the total field lifetime.

This historic dimension might make it more comprehensible, that the general test categories seem identical for all power module manufacturers at the first glance, but exhibit considerable differences at a closer look. International standards are defining the general test setup, but the procedural details remain ambiguous. Therefore, every manufacturer of power modules has established its internal test philosophy, which is capable of defining, maintaining and improving the internal quality level, but which makes it difficult to compare qualification test results between different manufacturers.

Table 12.1 collects a common set of (accelerated) qualification tests for IGBT and MOSFET power modules with test conditions from [LV324]. The standards mentioned in the table are background for the specific conditions.

This compilation also indicates the progress of modern IGBT- and MOSFET-modules compared to conventional diode and thyristor modules: By way of example is the high temperature reverse bias test performed at 100% of the nominal blocking voltage and the power cycling requirement is increased from 10,000 to 20,000 cycles.

Before we will discuss each test in more detail, we have to add an important definition for every test: the failure criteria. It must be emphasized, that the exact definition of failure criteria is essential for the evaluation of any test. Table 12.2 states the common failure criteria for qualification and endurance tests, as they are specified by international standards.

The failure criteria in Table 12.2 allow a certain increase relative to specification limits or initial measured values. Since the accelerated test conditions aim to

Table 12.1 Reliability tests for qualification of IGBT/MOSFET-modules for industrial applications with reference to conventional modules

	Name	Conditions	Standards
HTRB	High temperature Reverse bias test	MOS/IGBT: 1000 h, T_{vjmax}, 0.8*V_{Cmax} partially V_{Cmax} ($\leq$2.0 kV), Conv.: 1000 h, T_{vjmax} −20 K, $V_R/V_D = 0.8*V_{RRM}/V_{DRM}$ resp. $0.66*V_{RRM}/V_{DRM}$[a]	IEC60747-9:2007 IEC 60747-2/6
HTGS (HTGB)	High temperature Gate stress test	1000 h, V_{Gmax}, T_{vjmax}	IEC60747-9:2007
H3TRB (THB)	High humidity High temperature Reverse Bias test	1000 h, 85 °C, 85% RH, $V_C = 0.8*V_{Cmax}$, however max. 80 V, $V_G = 0$ V	IEC60749 -5:2003
LTS	Low temperature Storage test	$T = T_{stgmin}$, 1000 h	JESD-22 A119:2009
HTS	High temperature Storage test	$T = T_{stgmax}$, 1000 h	IEC60749-6:2002
TST	Thermal shock	$T_{stgmin} - T_{stgmax}$, typ. −40 °C to +125 °C, $t_{storage} \geq$ 15 min, $t_{change} \leq$ 30 s 1000 cycles Conv.: 25 cycles[a]	IEC60749-25:2003
PC_{sec}	Power cycling	Internal heating and external cooling $t_{on} < 5$ s; $I_L > 0.85 \cdot I_{nom}$[b]	IEC60749-34:2011
PC_{min}	Power cycling	Internal heating and external cooling $t_{on} > 15$ s; $I_L > 0.85 \cdot I_{nom}$[b]	IEC60749-34:2011
V	Vibration	Sinusoidal sweep, 5 g, 10–1000 Hz, 2 h per axis, (x, y, z)	IEC60068-2-6 Test Fc
MS	Mechanical shock	Half sine pulse, 30 g, 18 ms, 3 times each direction ($\pm x$, $\pm y$, $\pm z$)	IEC60068-2-27 Test Ea

[a]Conventional devices—thyristors, diodes
[b]For min. one data point in [LV324]

simulate the stress applied in the total service lifetime, the specification limits are permitted to be exceeded within certain limits. For parameters less critical for the performance of the module, a greater increase can be allowed as shown for the leakage current. Other more critical parameters for the performance, as the forward voltage drop or the thermal resistance have to remain within closer limits, because they have a direct impact on the chip temperature.

The qualification tests in Table 12.1 can be classified into three groups. The first three tests are chip related qualification tests, which are also part of every chip

Table 12.2 Failure criteria for acceptance after endurance tests

Failure criteria IEC60747-9(2001):	
Gate leakage current I_G	+100% USL
Collector/Drain leakage current I_D/I_C	+100% USL
On-state voltage $V_D/V_{C(\mathrm{sat})}/V_F$[a]	+20% IMV or 0% USL
Threshold voltage V_T	+20% USL
	−20% LSL
Thermal resistance R_{thjc}/R_{thjh}[b]	+20% IMV or 0% USL
Isolation voltage V_{ISOL}	Not below specification limit

USL Upper specification limit, *LSL* lower specification limit, *IMV* initial measured value
LV324 for power cycling tests [a]+5% IMV [b]ΔTj <+20% IMV

qualification. But since the chips are exposed to different substances during the module assembly process (i.e. solder flux, cleaning solvents and silicone soft mold), a confirmation of the chip reliability in the assembled module is inevitable. This set of chip related tests is followed by a group of seven tests related to stability of the package in the specified operation and storage temperature range and under external and internal temperature swings. Especially the power cycling test is important for the lifetime of power modules in application. The last two tests are confirming the mechanical integrity of the package.

12.2 High Temperature Reverse Bias Test

The high temperature reverse bias test (HTRB) – sometimes also referred to as hot reverse test – verifies the long term stability of the chip leakage currents. During the HTRB-test, the semiconductor chips are stressed with a reverse voltage at or slightly below the blocking capability of the device at an ambient temperature close to the operational limit. No degradation can be expected in the bulk silicon of the devices at these temperatures, but the test is able to reveal weaknesses or degradation effects in the field depletion structures at the device edges and in the passivation.

The electrical field has to be expanded at the edges of a power device to reduce the tangential field at the chip surface by a field depletion structure. This can be achieved by a field ring structure, by a variation of lateral doping or by a suitable geometric contour. Nevertheless, in silicon the electrical fields at the surface are in a range of 100–150 kV/cm. Movable ions can accumulate in these high field areas and can generate a surface charge. The source of these ions can be contaminations during the assembly process or residues of process agents, for example solder flux. The high temperature accelerates the process. The surface charge can alter the electrical field in the device and generate additional leakage currents. It can even produce inversion channels in device regions with low doping profiles and produce short circuits paths across the pn-junction.

The failure criteria limit the allowed leakage current increase after the test – when the device is disconnected from the voltage supply and cooled down – to identify such degradation effects. Additionally, most manufacturers of semiconductor devices also continuously monitor the leakage current during the 1000 h test and require a stable leakage current throughout the test.

Figure 12.1 shows an example of the recorded leakage current in a high temperature reverse bias test. Eight devices were monitored for the test duration. The devices are initially stable but after approx. 200 h the leakage current starts to increase. The test was aborted after 920 h due to the massive increase of leakage current of some devices. The test failed for these devices, because the applied junction passivation was not capable of fulfilling the requirement.

The test conditions apply a considerably higher stress than the typical application. The nominal DC-link voltage will be in the range of 50–67% of the specified device blocking voltage in real systems; it will be exceeded only by temporary voltage peaks. Furthermore, the device will reach the maximum operational temperature only occasionally in normal applications. Thus, the test is a highly accelerated procedure to generate stress within a test duration of 6 weeks for applications designed for a lifetime of 20 years and more.

But even if the junction passivation is compatible with the requirements and the assembly process, the test can reveal flaws in the package design. Figure 12.2 shows an example of a temporarily increased leakage current for a single device after nearly 100 h of test time. After reaching its maximum, the leakage current decreased and showed only a small increase at the end of the test. The following investigation showed, that a single wire bond of this device had a wrong geometry: it had no loop and was directly laying on the guard ring passivation of the IGBT.

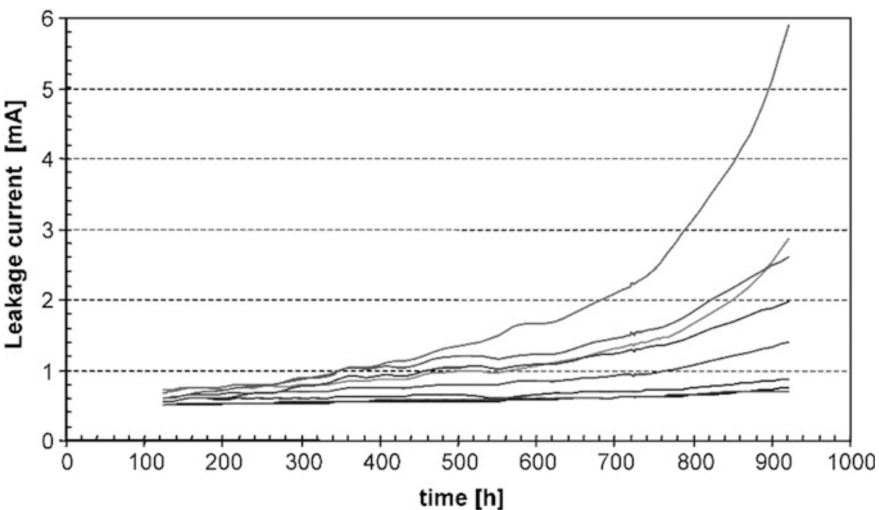


Fig. 12.1 Recorded leakage current during a high temperature reverse bias test – an example for a failed test

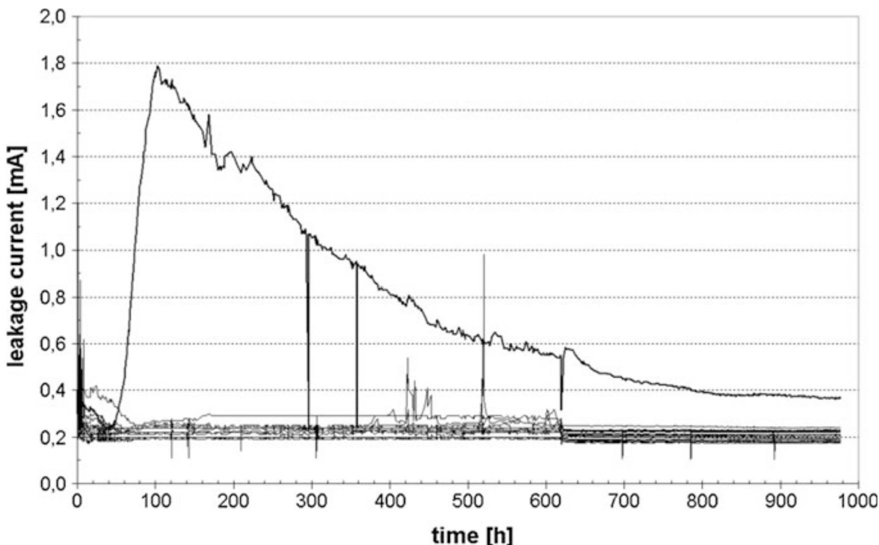


Fig. 12.2 Recorded leakage current during a high temperature reverse bias test – the test was passed but revealed a design flaw of the wire bond geometry

The test was repeated after the correction of the bond wire layout and the formally observed leakage current increase was eliminated. This result confirms the relevance of the HTRB-test for the package development.

In SiC, in most cases this junction termination is a "reduced surface field" (RESURF) structure which is fabricated as junction termination extension by a low doped p-layer, similar to Fig. 4.25. Due to the material properties of SiC, the critical electric field strength is in the range of 2.5 MV/cm, even up to 3 MV/cm at the edge of cells [Rup14]. The high critical field strength is not fully used in SiC nowadays and, compared to Si, higher safety margins are used in state-of-the-art devices. Nevertheless, electrical fields at the SiC surface in the range of >1 MV/cm are expected.

Furthermore, the acceptors implanted in the junction termination region are typically partially compensated (or enhanced) by surface charges generated by the polarity of the hexagonal axis of SiC (Si-face or C-face behave differently) and by processing techniques (e.g. oxidation or dry etch processes, see [Yan00]). Those surface charges typically are in the range of 10^{12} cm^{-2}, i.e., they have an influence on the optimum junction termination design in order to allow a maximum breakdown voltage. As discussed above, mobile ions can accumulate and can modify the initial surface charge. The surface charge alters the electrical field in the device and changes the breakdown voltage of the junction termination. Depending on the initial state both an increase and a decrease is possible. This is illustrated in Fig. 12.3 which shows the dependency of the breakdown voltage on the junction termination extension (JTE) dose as used for design of SiC devices.

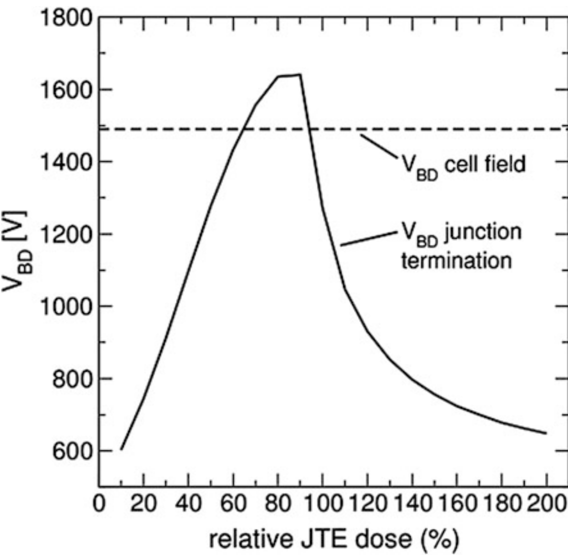


Fig. 12.3 Simulated breakdown voltage versus JTE edge dose in comparison to cell field breakdown **for a 1200 V device** indicating the process window for a single zone JTE implantation. Figure from R. Elpelt, Infineon

Even if a device is designed to have the avalanche breakdown in the cell field [Rup14], a change of the JTE charge via mobile ions may even lead to a shift of the avalanche onset from the cell field to the periphery, accompanied by a decrease of the avalanche current withstand capability. This becomes even worse if this effect does not occur homogenously on the chip circumference, but is concentrated on specific areas leading to punctiform avalanche locations. Such surface charges in the edge termination region can even produce inversion channels in device regions with low doping density and produce short circuit paths across the pn-junction. Since the base doping in SiC is about 100 times higher than in Si for the same blocking voltage, 100 times more surface charges are also required to create an inversion channel. In general, hard failures in HTRB testing are therefore not very likely in case of mature SiC device designs, but it is also important to carefully monitor any shift of breakdown voltage and leakage current triggered by the stress test. Such a drift is a good indicator for a JTE design which is not rugged enough. This holds not only for HTRB but – even more pronounced – for high voltage H3TRB stress tests (see Sect. 12.4).

12.3 High Temperature Gate Stress Test

The high temperature gate stress test (HTGS) or high temperature gate bias test (HTGB) confirms the stability of the gate leakage current. Even though the maximum allowed gate voltage is limited to ± 20 V, this voltage is applied to a not more than 100 nm thick gate oxide layer in state-of-the-art IGBTs and MOSFETs.

This results in an electrical field of 2 MV/cm across the gate oxide. For a stable leakage current, the gate oxide must be free of defects and only a low density of surface charges is tolerable. The boundary condition of maximum operational temperature for the devices again accelerates the test.

Since the leakage currents are very small (< 10 nA), this test is also extremely sensitive to surface contaminations on the chip. In test modules, where thermocouples were glued to the emitter contact of an IGBT for measurement purposes, the gate leakage was found to be considerably increased. This increase was caused by the residues of the solvent of the glue, which remained on the chip surface between the gate and the emitter and caused a measurable increase of the leakage current. Therefore, the gate leakage test also ensures the cleanness during the assembly process of a module.

While Si MOSFETs and IGBTs are usually very stable in the gate stress test, the gate oxide reliability has always been a great challenge for MOS (metal-oxide-semiconductor) structures fabricated on SiC substrate materials [Lip99]. As SiC technology has matured, SiC MOS devices have exhibited gradual improvements in time-dependent dielectric breakdown (TDDB) characteristics. However, even until now the gate oxide reliability of large devices (5–50 mm^2) did not succeed in achieving the same low rates of early failures as Si devices with comparable gate oxide area [Lut17].

Especially the trade-off between R_{on} and allowed electric field in the gate oxide is more challenging with SiC. The channel resistance is a major part of R_{on}. R_{ch} of the MOSFET is given by Eq. (9.3), with Eq. (9.1) it can be written as

$$R_{ch} = \frac{L}{W \cdot \mu_n \cdot C_{ox} \cdot (V_G - V_T)} = \frac{L \cdot d_{ox}}{W \cdot \mu_n \cdot \varepsilon_0 \cdot \varepsilon_r \cdot (V_G - V_T)} \qquad (12.1)$$

with the entire width W and the length L of the channel; μ_n is the mobility of free electrons, V_G is the applied gate voltage, V_T the threshold voltage and d_{ox} the gate oxide thickness. The channel mobility in SiC is, despite some progress, much lower than in Si. Low R_{ch} might be achieved by reducing d_{ox} in SiC-MOSFETs in comparison to Si-MOSFETs and Si-IGBTs due to the impact of oxide thickness on the channel resistance. Increasing the recommended V_G is another possibility for low R_{ch}, see Eq. (12.1). Both measures increase the electric field in the gate oxide and are a risk for its durability.

The qualification conditions in Table 12.1 require a test of 1000 h at V_{Gmax} and 125 °C, the recent specification [LV324] requires T_{vjmax} which can be up to 175 or 200 °C. However, for a high-quality oxide layer it will not be possible to reach end-of-life within reasonable time. Intrinsic failures are caused by broken SiO_2-bonds or due to Fowler-Nordheim tunneling and are expected at 10 MV/cm for SiO_2. More detailed models [McP85, Kim97] allow the calculation of oxide breakdown depending on electric field, temperature and time. Assuming oxide thickness d_{ox} = 80 nm, V_{Gmax} = 20 V, $E = V_{Gmax}/d_{ox}$ = 2.5 × 10^6 V/cm with the parameters of [Kim97] leads, even for T_{vjmax} = 200 °C, to an intrinsic lifetime of about 32 years, which is not applicable as end-of-life test method.

A new test method is suggested in [Bei16] and [Bei17]. It is based on the distinction between extrinsic and intrinsic failures. Extrinsic failures are attributed to defects or weak points in the oxide, whereas intrinsic failures represent the true capability of the oxide itself [Coo97]. Simplified, any extrinsic failure is considered as gate oxide thinning [Lee88]. Figure 12.4 shows gate oxide with d_{ox} as nominal thickness. d''_{ox} and d'_{ox} are the reduced gate oxide thicknesses for the extrinsic failures. For the lowest thickness d''_{ox} the breakdown of gate oxide thickness will occur first due to the highest electrical field at a given gate voltage.

In the test at a temperature of 150 °C the applied gate voltage V_G of the devices is increased during the test every 168 h (one week). After each interval of 168 h, the test was interrupted for a measurement of the threshold voltage at room temperature. The gate voltage in the first step is the rated gate voltage V_{GUSE} according to the data sheet. In the second step, the voltage was set to the maximum use gate voltage V_{GMAX}. After this step, the gate voltage was increased in defined steps.

A graphical demonstration of the test strategy is shown in Fig. 12.5. In this image, the voltage steps increasing each interval of 168 h is shown. To be able to compare different devices, the difference between applied gate voltage V_G and the use-gate-voltage V_{GUSE} recommended by the manufacturer, $V_G - V_{GUSE}$, is chosen for characterization. By applying the $V_G - V_{GUSE}$-method, the test can be

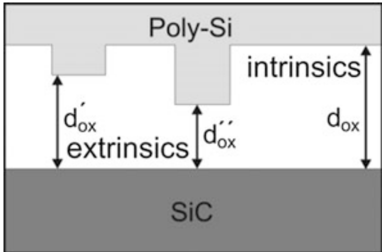


Fig. 12.4 Model for extrinsic and intrinsic failures. d''_{ox} and d'_{ox} symbolize gate oxide thinning for extrinsic failures

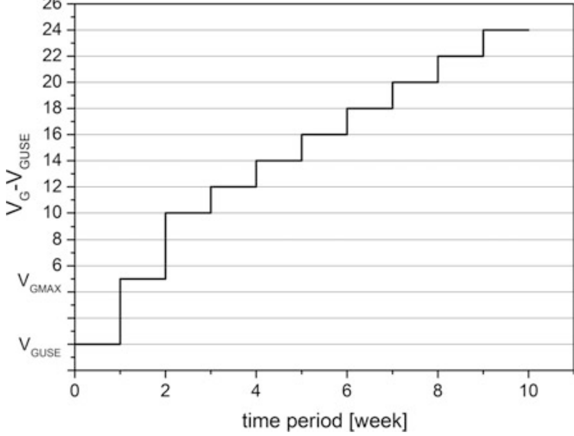
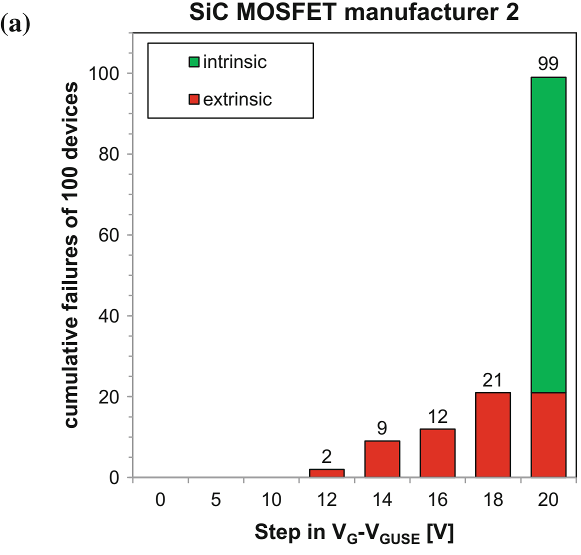


Fig. 12.5 Example for the test process with stepwise increased $V_G - V_{GUSE}$

reproduced in every laboratory. $V_G - V_{GUSE}$ is relevant for evaluating the impact of gate overvoltages in applications.

The test results depicted in Fig. 12.6 show a considerable difference in the gate stability for both manufacturers. In both test, a pronounced increase of the failure rate is observed. This increase could be attributed to the change of failure mode from extrinsic to intrinsic failures. The devices from manufacturer 2 reached the intrinsic limit at the step $V_G - V_{GUSE} = 20$ V (Fig. 12.6a), while no failures were



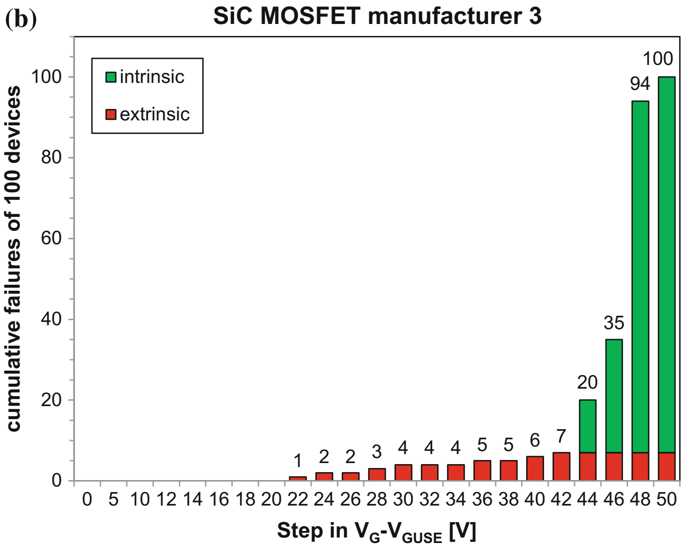


Fig. 12.6 Test results with stepwise increased $V_G - V_{GUSE}$ for two SiC MOSFETs from different manufacturers

observed for manufacturer 3 up to this stress level. The devices from manufacturer 3 reached the intrinsic limit at $V_G - V_{GUSE} = 44$ V (Fig. 12.6b).

Si-IGBTs of two professional manufacturers have been exposed to a similar test method in [Bei17]. They show a smaller rate of extrinsic failures. The intrinsic limit for the first IGBT manufacturer is found between 45 and 55 V, for the second IGBT manufacturer it is detected at 75 V. The differences in intrinsic failure limit can be explained by a different thickness of the gate oxide, see discussion of Eq. (12.1). From an application viewpoint, no evidence for significant gate oxide failures of IGBTs from established manufacturers in the field is available, even though there are many millions of IGBTs working in various applications. It must be considered that the applied gate voltage in the tests is much higher than used in practical applications.

Compared with IGBTs, only the SiC MOSFET of manufacturer 3 shows a similar behavior. The first intrinsic failures are observed in the same range of the gate voltage with a marginally higher level of extrinsic failures compared to Si IGBTs. This could be taken as evidence, that SiC MOSFETs are capable to reach a gate oxide reliability comparable to Si IGBTs. However, in state-of-the-art SiC MOSFETs huge differences between the investigated manufacturers were confirmed by the tests.

It was pointed out in [Lut17] that the difference between SiC and Si MOSFETs is a three to four orders of magnitude higher defect density of SiC MOS structures at the end of the process. This much higher defect density is most likely linked to substrate defects, metallic contaminations and particles. If further investigation would confirm that the extrinsic failures in the presented accumulated gate stress test are early life failures, then efficient screening tests could be developed on this basis. The data on failures as a function of test time reported in [Sie17] exhibit a decreasing failure rate over time and therefore identify the failure category as early life failures. The test procedure could also be applied to define the necessary gate oxide thickness to achieve a comparable gate oxide reliability for SiC MOSFETs as is established today in commercial Si switches.

12.4 Temperature Humidity Bias Test

The temperature humidity bias test (THB), also known as high humidity high temperature reverse bias test (H3TRB), is focusing on the impact of humidity on the long term performance of a power component.

Capsules are – when defect-free assembled – hermetically sealed against the environment. This is not the case for the majority of power module packages. Although bond wires and chips are completely embedded in silicone soft mold, this material is highly permeable for humidity. Therefore, humidity can intrude the package and can reach the chip surface and junction passivation. This test aims to detect weaknesses in the chip passivation and to initiate humidity related degradation processes in the packaging materials.

Sometimes, proposals are made to suppress the humidity access by an additional protective layer. But two strong arguments have to be taken into account: First, the silicone soft mold exhibits a high linear coefficient of thermal expansion ($\sim$ 300 ppm/K). This yields a considerable increase in volume during temperature swings and makes it difficult to apply a hermetically tight layer on top of the soft mold. Moreover, if a hermetically sealing is not achievable, then an additional layer can only reduce the diffusion rate of the humidity. However, a reduced diffusion rate works in both ways: It increases the time for the humidity to soak in, but it will also increase the time for the intruded humidity to be driven out. Therefore, a non-hermitical protection layer would increase the duration for the humidity to influence the device in operative condition. This is not desirable. If the humidity cannot be kept from intruding the package, then it should be driven out fast when the module is going in operational mode. For this reason, a highly penetrable embedding compound should be preferred.

The applied electrical field during the test acts as a driving force to accumulate ions or polar molecules at the semiconductor surface. On the other hand, the power losses generated by the leakage current must not heat up the chip and its environment and thus reduce the relative humidity. Therefore, standards require a limitation of the self-heating of the chip to not more than 2 °C. Consequently, the reverse voltage had been limited to 80% of the blocking voltage for low blocking voltage MOSFETs and was restricted to a maximum of 80 V for higher blocking capabilities in the past.

A number of field experiences in the past years have shown that this test condition is not sufficient for all application conditions. Field failures, which could clearly be attributed to the influence of humidity, have raised a discussion about this 80 V maximum applied voltage. Since the leakage currents of modern semiconductor chips are low enough to maintain the temperature increase within 2 °C even at 80% of the nominal blocking voltage for blocking voltages of 1200 V and more, the restriction to 80 V seems to be outdated.

Accelerated humidity bias tests [Zor14] have confirmed that the failure rate is considerably increased at elevated voltage levels. Furthermore, degradation of the blocking voltage was observed at bias levels of 65 and 90% of the nominal blocking voltage (Fig. 12.7), which was not observed at 80 V bias voltage.

For power semiconductors, the most prominent corrosion mechanisms induced by an exposition to humidity are the electrochemical migration and the aluminum corrosion [Zor15]. On the other hand, humidity can also enhance the mobility of mobile ions resulting from processing, polyimide passivation or mold compound [Beu89]. This can cause an alteration of the junction termination blocking capability faster than in the HTRB test.

A model to simulate the moisture inside the converters and power modules is presented in [Bay16]. This model describes the diffusivity, permeability and storage of humidity in polymers by equivalent R-C-networks. It was found that the moisture inside the converter cabinets can exceed the ambient conditions significantly and that condensation of water can occur under unfavorable conditions of operation.

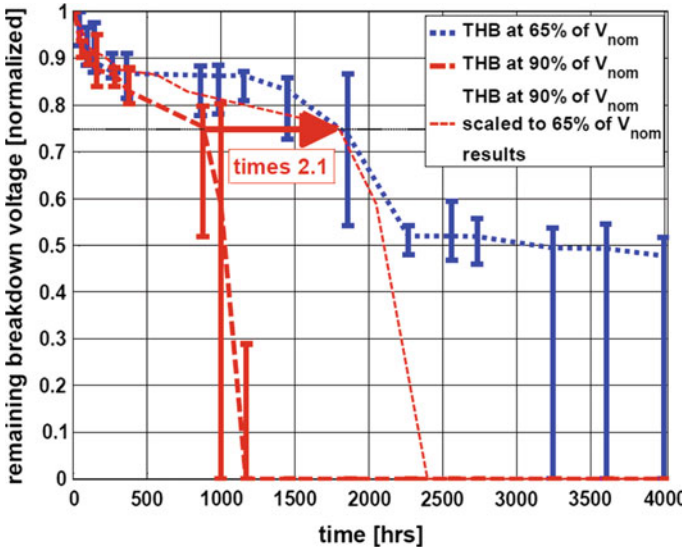


Fig. 12.7 Degradation of blocking voltage in humidity test biased at 65 and 90% of the nominal blocking voltage [Zor14]. Reprint with permission of VDE Verlag

SiC devices were found to be in the same way sensitive to humidity. Figure 12.8 shows a test of a phase leg of a 1200 V module with SiC MOSFETs executed at 1080 V, which is 90% of the rated voltage. Up to a time of 122 h the leakage current at 85 °C was below the measurement resolution. Then an increase becomes visible, after 134 h follows a fast increase and the module fails.

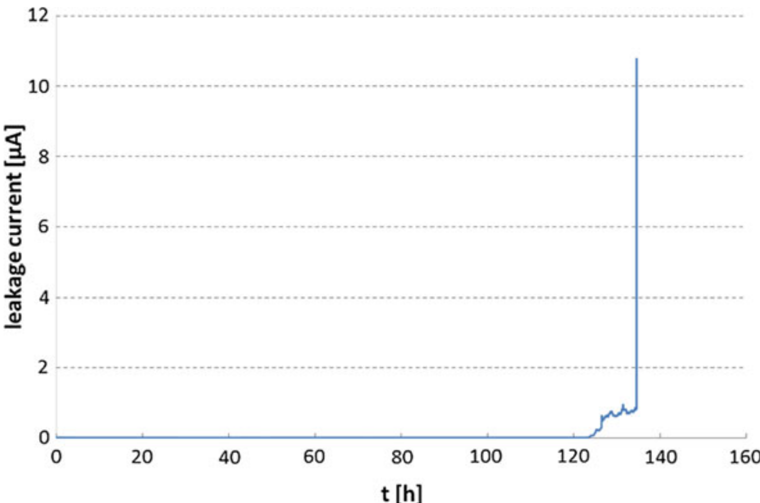


Fig. 12.8 Leakage current of one phase leg of a 1200 V SiC MOSFET module. Relative humidity 85%, temperature 85 °C, DC voltage 1080 V. Test executed by Chemnitz University of Technology

Many applications of power devices are outdoor applications, like photovoltaic inverters. The humidity might be more critical when the temperature inside the converter enclosure and modules housing is varying with daily temperature variations and weather constraints [Sad16]. A test where half of the power modules are placed indoor and half outdoor and both are operated in inverter mode is executed in [Sad16]. The disadvantage is that there is no acceleration of failure mechanisms.

Currently, activities have been started for improving the humidity sensitivity of power devices at high bias levels. In parallel, high voltage humidity testing (HV-H3TRB) is implemented in power module qualification programs. This progress will result in a higher reliability of non-hermetical power modules in high humidity application environments.

12.5 High Temperature and Low Temperature Storage Tests

The storage test at the maximum and minimum storage temperatures have been implemented to verify the integrity of the plastic materials, rubber materials, organic chip passivation materials, glues and silicone soft molds utilized in most state-of-the-art power module packages. These materials must maintain their characteristics in the complete specified storage temperature range.

At this point, a remark is necessary to prevent a misunderstanding of the term 'storage temperature'. The denomination 'storage temperature' has been established in the early days of semiconductor power modules. It refers to the non-operational temperature limits for power modules assembled in a power electronic system. It does not refer to storage conditions of the unassembled power module, as the denomination might suggest. The description 'non-operational temperature limits' would be more appropriate, but the heritage of the early days of power modules impedes this transition.

Long term storage at high temperatures is critical for the mechanical strength of all thermoplastic housing materials. It is also minatory for the flame retardant additives to the thermoplastic materials required for the resilience against fire hazards. Conventional silicone soft mold starts to degenerate at temperatures above 180 °C, so that specially developed high temperature silicone gels must be implemented for junction temperatures above 175 °C.

Long term storage at low temperatures is critical for the softening agents in plastic materials and rubber materials, it can destroy the elastic capabilities of these materials and thus impair their functionality. Silicone soft mold compound is also limited in the minimum storage temperature; most standard soft molds are limited to −50 °C. Below this temperature, cracks might appear in the soft mold, which will not be cured by increasing temperatures and thus will compromise the insulating environment for high blocking voltage devices.

The standard storage temperature limits are −40 °C/+125 °C and a variety of materials are available for reliable performance in this temperature range. Extensions of the temperature range, which are demanded by many applications today, will make these qualification tests more challenging in the future.

12.6 Temperature Cycling and Temperature Shock Test

Temperature swings are an essential stress condition for every power electronic component in application. The temperature cycling test and the temperature shock test are two test methods to simulate ambient temperature swings during the field lifetime.

The test conditions are discriminated by the change rate of the externally imprinted temperature. If the rate of temperature change is slow in the range of 10–40 °C/min, the test is called temperature cycling test. In a temperature shock test, the ambient temperature is changed typically in less than 1 min. For power modules, this is typically achieved by a two chamber equipment, in with the air is permanently heated or cooled to the maximum or minimum test temperature, while an elevator carrying the devices under test moves between the two chamber in a time interval below one minute. Since the heat exchange rate is rather slow for a gas environment, the duration for reaching an equilibrium temperature distribution inside the module can vary from 30 min to 2 h, depending on the total thermal capacity of the devices under test.

A more extreme version of the temperature shock test is the liquid-to-liquid thermal shock test. In this test, the ambient is formed by appropriate liquids, heated or cooled to the desired temperature limits, for example oil at 150 °C or more and liquid nitrogen at −196 °C. Such test conditions are not common for modules, but are often performed for package elements as DBC substrates. In a liquid ambient, the heat transfer is much faster than in a gaseous ambient, so that an equilibrium temperature distribution can be achieved in minutes rather than hours.

Due to the wide range of heat transfer rates, the denomination of temperature swing test is somewhat ambiguous. While some manufacturers refer to the two chamber air ambient test as temperature shock test to distinguish it from the rather slow single chamber test with ambient temperature change rates in the range of 20 °C/min, other supplier label the two chamber air environment test as Temperature Cycling Test to discriminate it from the much faster liquid-to-liquid test. This ambiguity of denomination must be kept in mind when comparing qualification requirements and test results from different manufacturers.

A common boundary condition for all types of temperature cycling tests is the requirement, that the cycle time must be long enough, so that all parts of the assembly reach the maximum or minimum temperature – which are typically the storage temperature limits – so that the assembly is in a thermal equilibrium condition. Since the test simulates the impact of temperature changes by external sources, for example the change of ambient temperatures or an increase of heat sink

temperature by other heat sources, the power modules are not actively stressed by current or voltage. The changes in parameters are checked by an initial and final measurement and have to comply with the failure criteria.

The combination of different materials with different coefficients of thermal expansion results in high mechanical stress in the system. More so, the bimetal effect causes a cyclic deformation of the module. Simulations of the thermo-mechanical behavior of a power module have shown [Mik01], that if this bimetal bending is reduced – for example by mounting the module on a heat sink – the stress is reduced and the lifetime is extended. Therefore, modules should be mounted on assembly plates during the test to simulate as close as possible the application conditions.

The cyclic mechanical deformation generated by temperature cycles due to the difference in coefficients of expansion of the material layers causes stress in the functional layers themselves and in the interconnection layers. This will lead over time to the initiation of cracks and cause growing delaminations in these layers. Scanning Acoustic Microscopy (SAM) is the appropriate detection method of identifying delaminations in power semiconductor modules.

An example of the damage caused during temperature cycling with classical 34 mm base plate modules show the SAM images in Fig. 12.9. Two different solder materials for the substrate-to-base plate interface were compared in this test: a RoHS compatible SnAg(3.5) solder and the classical SnPb(37) solder. Both solder interfaces – the interface between the substrate and the base plate and the interface between chip and substrate – were investigated by choosing the appropriate time-of-flight windows in the SAM signal for each solder type. The comparison between the initial measurement and the SAM image after 200 temperature cycles (−40/+125 °C) in a two-chamber test equipment reveals growing delaminations in the substrate-to-base plate solder layer for both solder versions, indicated by the white areas (= regions of high reflection) which move inward from the corners and short edges of the substrates. The classical eutectic SnPb solder shows more damage along the outside short sides of the substrates, where the terminals are connected at the topside of the substrate.

The SAM images of the chip-to-substrate interface show no indications of any fatigue in the chip solder interfaces, but it presents black areas in the regions, were solder delaminations are found in the substrate-to-base plate solder layer. This artefact is produced by a lack of acoustic energy in these regions, because most of the signal was already reflected in the delaminations found in the substrate-to-base plate solder layer. Since the SAM signal is injected form the base plate surface, reflections nearer to the base plate reduce the signal propagating into deeper layers. However, this common artefact allows conveniently evaluating how close the delamination has come to the chip position. As Fig. 12.9 clearly shows, the delaminations have propagated much deeper under the chip areas for the SnPb solder system as for the SnAg solder system with a stronger unfavorable impact on the thermal resistance of the chips.

The lifetime under temperature cycles is determined by the combinations of different materials (with different coefficients of thermal expansion) and the stability

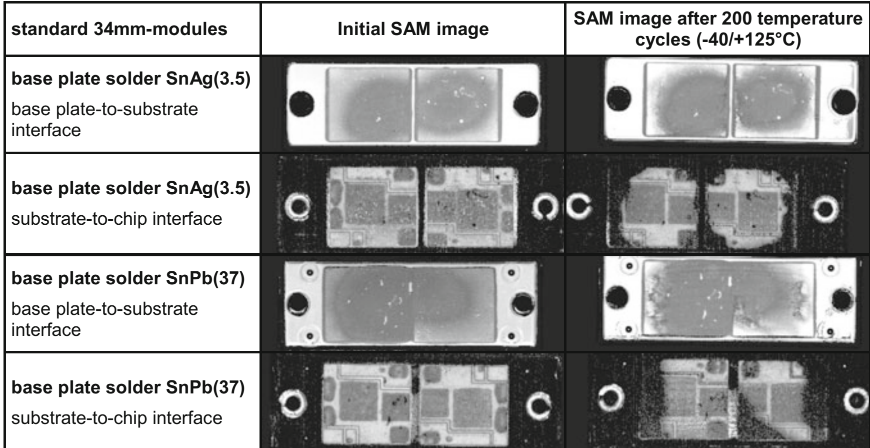


Fig. 12.9 Scanning Acoustic Microscope (SAM) images of standard 34 mm modules before and after 200 temperature cycles (−40/+ 125 °C) – different delamination patterns are found for different base plate solder materials

of the interconnect layers. Due to the mechanical deformation, smaller packages as the TO-family are more stable than larger modules that are more complex. Since the source for the passive temperature cycles is located outside of the module, only the used materials and interconnection layer decide about the reliability of a package.

12.7 Power Cycling Test

In contrast to the temperature cycling test, the power chips are actively heated by the losses generated in the power devices themselves in a power cycling test. This accounts for a fundamental difference between the two tests: The durability of a power module in temperature cycling tests is determined by the CTE of the implemented materials and the geometry of the layers. In a power cycling test, the amount of losses can be affected by the chip technology and by the silicon area implemented in the module. Therefore, any power cycling lifetime requirement can be met just by implementing sufficient silicon area to reduce the temperature swing generated by the chip losses. However, commercial aspects limit this option in practical applications.

12.7.1 Power Cycling Test Execution

During power cycling test, the device under test is mounted on a heat sink as in a real application. A constant DC load current is conducted by the power chips and

the power losses are heating up the chip. When the maximum target temperature in the chip is reached, the load current is switched off and the system cools down to a minimum temperature. The reaching of the minimum temperature completes the cycle and the next cycle begins by starting the load current again. During each cycle, considerable temperature gradients are generated inside the module.

An exemplary test setup as established for Si IGBTs is shown in Fig. 12.10. The device is heated up by a DC constant current load pulse. At the end of the load pulse the upper temperature T_{vjhigh} is reached and the device cools down. The lower limit of the junction temperature T_{vjlow} is reached at the instant when the load current is turned on again, and the cycle is repeated. The characteristic parameter for power cycling tests, the temperature swing ΔT, is given by the temperature difference between maximal junction temperature T_{vjhigh} at the end of the heating phase and the minimal junction temperature T_{vjlow} at the end of the cooling interval:

$$\Delta T = T_{vjhigh} - T_{vjlow} \tag{12.2}$$

In Fig. 12.10, ΔT_j can be determined as 78 K.

Another important parameter for the power cycling test is the medium temperature T_m of the swing:

$$T_m = T_{vjlow} + \frac{T_{vjhigh} - T_{vjlow}}{2} \tag{12.3}$$

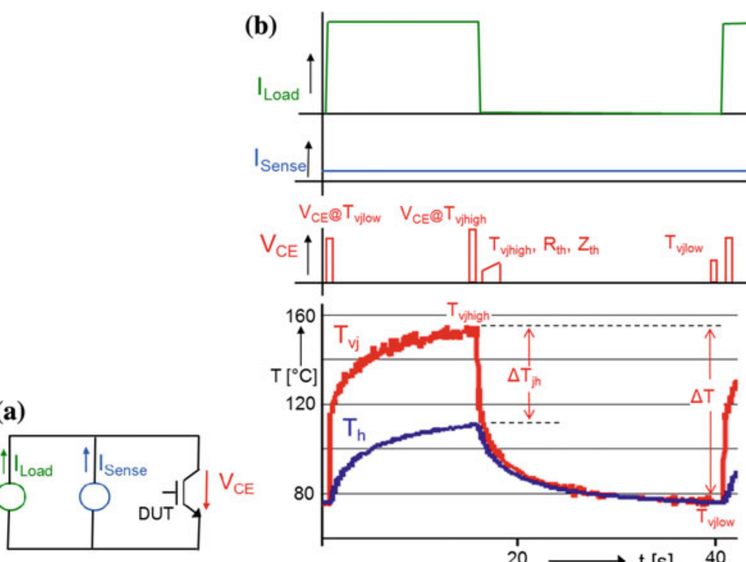


Fig. 12.10 **a** Basic test setup, **b** course of time for load current, sense current, executed voltage measurements, virtual junction temperature and heat sink temperature for one cycle of a power cycling test

Because of these relations, any pair of the parameters ΔT_j, T_m, T_{vjhigh}, and T_{vjlow} can be used as characteristic parameters for the temperature swing. Further parameters, e.g. the duration of the load pulse, are of importance as shown below. A long power pulse duration (15 s in Fig. 12.10) usually represents a higher stress for the devices.

The measurement of the junction temperature is executed with the constant sense current I_{Sense} which must be small enough to create only negligible losses, $I_{Sense} \approx 0.001\ I_{Load}$. A pn-junction is used as temperature sensor, its junction voltage is the temperature-sensitive electrical parameter (TSEP). This method is known as the determination of the virtual junction temperature T_{vj}. It is based on the physics of a pn-junction whose characteristic is described in Chap. 3 with Eqs. (3.50) and (3.51). It leads to a voltage V_j which is strongly depending on temperature given in Eq. (3.52). An example is given in Fig. 3.12.

The voltage V_j is decreasing with T due to the domination of n_i in Eq. (3.51). The linear dependency is lost if the voltage drop in the base is to be taken into account.

First, a calibration function $V_j(T)$ is determined, see Fig. 11.19. When a current I_{Sense} is applied, the temperature can be read out. It is done in Fig. 12.10b just after turn-off of I_{load} to determine T_{vjmax} and just before the next turn-on to determine T_{vjmin}. The method to determine T_{vj} is established since the beginning of power device development. The pn-junction of a diode or the base-emitter junction of a bipolar transistor is used in [Oet73]. Thermal resistance in data sheets of European manufacturers is determined with the $V_j(T)$ method.

The so-determined virtual junction temperature T_{vj} deviates significantly from the maximum and minimum temperatures on the chip surface. The lateral temperature gradient across the die (see Fig. 11.20) is especially pronounced for chips > 1 cm^2 and under the condition of forced water cooling. The temperature at the center position of the die can be 20 K higher and at the edges 40 K lower than the temperature measured with the $V_j(T)$ method. A detailed investigation of the geometrical interpretation of the virtual junction temperature value for an IGBT is reported in [Scn09]. It shows that T_{vj} correlates to the area related average of the chip surface temperature. This is important when comparing temperature measurements to simulation results. The comparison with an IR camera measurement in this investigation revealed, that the shading of the temperature radiation of the IGBT surface by wire bonds affects the temperature measurement and leads to a lower value of the area related average of the junction temperature.

Details of measurements with an IGBT at turn-off of I_{load} are shown in Fig. 12.11. Just before turn-off, I_{load} (= I_{CE}) and V_{CE} are measured, leading to $P_v = I_{CE} \cdot V_{CE}$. After turn-off and a time interval t_d, V_{CE} at I_{sense} (= V_j) is measured. With the help of the calibration function, it gives $T_{vjhigh} = f(V_{CE})$. The case temperature T_{case} is measured by a thermocouple, hence obtaining $\Delta T = T_{vjhigh} - T_{case}$, see Fig. 12.10. The thermal resistance follows with

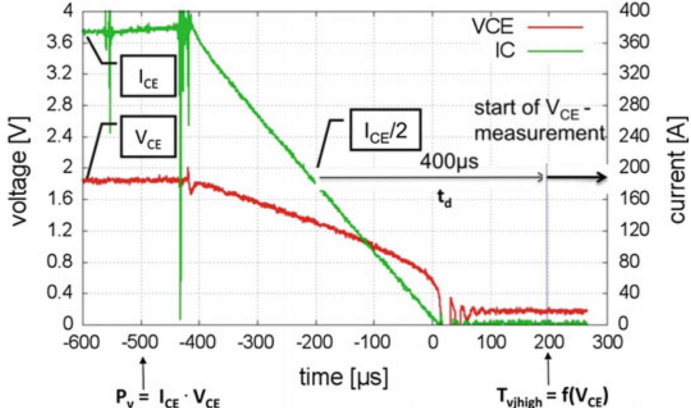


Fig. 12.11 Details of measurements at turn-off of the load current in Fig. 12.10

$$R_{thjc} = \frac{\Delta T}{P_V} \tag{12.4}$$

Collecting more measurement points after t_d, the thermal impedance can be calculated from the cool-down curve.

The time interval t_d must be set because of possible oscillations at turn-off, and also because of internal recombination processes if bipolar devices are used. The cooling-down in this interval for Si devices lies in the range of 2 K for usual power density, and up to 4 K for modules with high power density and advanced cooling systems. Usually, this cooling is neglected. A correction can be made with a simulation model, however, it should be mentioned in test evaluations, and the t_d should be given if a detailed evaluation is done.

The German automotive standard [LV324] is fixing much more details on the measurement process than the international JEDEC standard. It requires:

- the virtual junction temperature T_{vj} of the device under test must be determined with the $V_{CE}(T)$ method according to [Scn09].
- the supervision of the failure criteria has to be done with the parameters voltage drop (IGBT: V_{CE}, MOSFET V_{DS}, Diode V_F) and the temperature swing of T_{vj}. Both parameters must be monitored during the whole test for each cycle and have to be documented accordingly.
- the EOL criteria are to be checked by continuous supervision. Care has to be taken that the measurement data are with sufficient granularity according to the expected lifetime to ensure a valuable and exact determination of EOL.

The failure criteria are

- an increase of $V_{CE}/V_{DS}/V_F$ by 5%.
- an increase of R_{th} by 20% resp.ΔT_j by 20%.

- failure of one of the functions of the device, e.g. failure of the blocking capability or of the gate to emitter (gate to source) insulation capability for IGBTs and MOSFETs or by gate bond lift-off.

After the desired test conditions are adjusted in an initial setup phase, the heating time t_{on} and the cooling time t_{off} are used as constant control parameters for the total test duration. This is the preferred control strategy, however, different control strategies are still applied today with a considerable impact on the test results, as will be discussed in more detail at the end of this section.

The different coefficients of thermal expansion of materials together with the vertical and lateral temperature gradients generated in this layer system create mechanical stress at the interfaces during the temperature swing. This thermal stress leads in the long run to fatigue of materials and interconnections. Figure 12.12 shows the result of a power cycling test with a standard module. During the test, the forward voltage V_C of an IGBT is monitored. Additionally, it is possible to feed a defined sense current of some mA through the device after the turn-off of the load current, which allows to determine the upper temperature T_{vjhigh} with use of a calibration function (compare Fig. 11.19). The power losses P_V are also measured online. From junction temperature T_{vjhigh}, upper heat sink temperature T_h and P_V the thermal resistance is calculated using Eq. (11.4). Since the measurement of the fast changing heat sink temperature is difficult due to the response time of the applied sensor, the thus measured thermal resistance can deviate from the true stationery value, especially for short cycles below 10 s. However, even then this relative value can be used to monitor relative changes in R_{thjh}.

In Fig. 12.12 one can recognize that the on-state voltage drop at the IGBT remains almost constant over a large number of cycles, and the thermal resistance begins to increase slowly after approx. 6000 cycles. This is an indication for an increase of thermal resistance in the thermal path, mostly attributed to solder

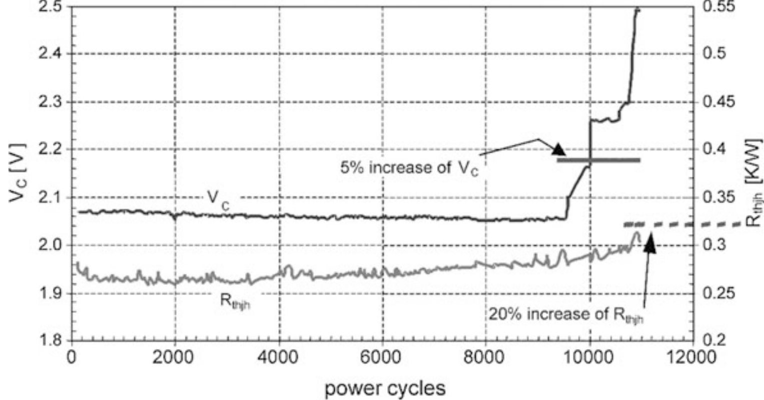


Fig. 12.12 Behavior of on-state voltage drop V_C and thermal resistance R_{thjh} at a power cycling test with ΔT_j = 123 K

fatigue. After more than 9000 cycles a first step in the V_C characteristic is observed, which is an indication for wire bond degradation. Shortly after that, the further steps are observed until finally all bond wires failed, the power circuit is open and the test can no longer be continued.

Bond wire degradation and solder fatigue are the main failure mechanisms in standard power modules. But from the shape Fig. 12.12 it is difficult to determine the primary failure mechanism. The failure limit of R_{thjh} would be reached at after approx. 11,000 cycles. But increase of R_{thjh} results in increasing temperature T_{high} and this will escalate the thermal stress for the bond wires. Therefore, solder fatigue is a significant failure mechanism in this test; it could be even the main failure mechanism. On the other hand, bond wire lift-off leads to increased V_C, which together with the constant current causes increasing losses and raises the upper junction temperature T_{high}, resulting in more thermal stress in solder layers. Due to the interdependency of the failure modes, power cycling tests require a careful failure analysis.

As mentioned before, different control strategies can be applied in power cycling tests. To evaluate the impact of different control strategies on the test result, a special test setup was designed which allowed the adjustment of test conditions during the power cycling test [Scr10]. In this special test setup, a single IGBT chip was subjected to power cycling with identical start conditions, but with 4 different control strategies (see Fig. 12.13):

1. Constant timing (t_{on} and t_{off} both constant): This strategy means that the operating conditions are kept constant throughout the test without reaction to an increase in thermal resistance or dissipated losses. This strategy provides no compensation for degradation effects and yields the lowest number of cycles to failure.

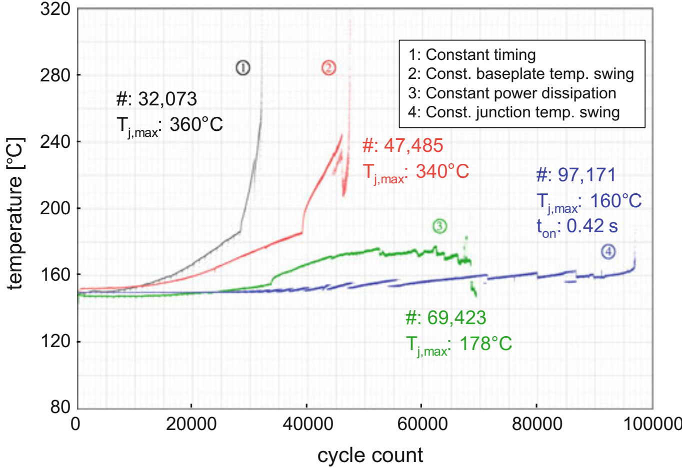


Fig. 12.13 Evolution of the maximum chip temperature during power cycling tests with 4 different control strategies from [Scr10]

2. Constant base plate temperature swing: In this strategy the signal of a thermocouple pressed to the baseplate underneath the center of the chip was used to define switching levels for turn-on and turn-off of the load current. This strategy was applied in the early days of power cycling testing, when several PC test setups shared the same cooling loop. If for example the cooling liquid temperature would increase due to losses generated in other test setups, the load pulse duration would be decreased and the cooling phase would be increased to maintain the same temperature swing. On the other hand, degradation in the thermal path between the baseplate and the heat sink is also compensated by this method resulting in an increased numbers of cycles to failure as shown in Fig. 12.13. Since today each PC test setup is equipped with an individual cooling system, this strategy has become obsolete.
3. Constant power dissipation: The aim of this strategy is to keep the power losses constant throughout the test. This can be achieved by controlling the gate voltage. The test is started with a reduced gate voltage and when losses increase due to degradation during the test, this is compensated by increasing the gate voltage. This compensation of degradation increased the number of cycles to failure for the investigated test vehicle by a factor of 2.
4. Constant junction temperature swing: This strategy aims for a constant temperature swing throughout the test. In [Scr10] this was achieved by reducing the load pulse duration when the temperature increased during the power cycling test. At the end of the test, the initial load pulse duration of 2 s was reduced to 0.42 s and the lifetime was a factor 3 higher than without compensation of degradation.

It should be mentioned that the temperature displayed in Fig. 12.13 is the maximum chip temperature measured in the chip center with a pyrometer in contrast to all other power cycling results showing T_{vjhigh}. This change was necessary to allow an active control of the load pulse duration based on the junction temperature value.

This example shows the significant impact of the control strategy on the power cycling lifetime. This must be kept in mind when comparing results obtained with different test strategies.

12.7.2 Power Cycling Induced Failure Mechanisms

12.7.2.1 Bond Wire Degradation

A typical fault image after power cycling of a standard module with base plate is shown in Fig. 12.14. All bond wires are lifted off from the IGBT chip. The bond wire in the background was the last one to fail and the current was flowing in a short time via an arc flash-over which caused a crater below the bond wire stitch. The bond wire failure in the foreground shows a characteristic feature of the lift-off

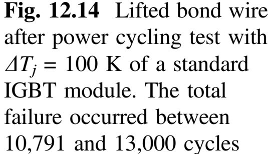


Fig. 12.14 Lifted bond wires after power cycling test with ΔT_j = 100 K of a standard IGBT module. The total failure occurred between 10,791 and 13,000 cycles

failure mode. The dissection did not occur at the interface between bond wire and chip metallization, but it emerges partially in the volume of the bond wire. Residues of bond wire material can still be detected on the surface of the chip metallization.

Figure 12.15 shows a magnified image of the lift-off area on the chip metallization. This example shows no adherence in the center of bond area. Improvements of the wire bond process can enhance the quality of the adherence significantly. In [Amr06] it was shown, that an improved wire bond process in combination with a replacement of the chip solder by a silver sinter technology can achieve a very high power cycling capability for power cycles up to T_{high} = 200 °C. In the viewpoint of high temperature applications, bond wires seem not to be the main limiting factor.

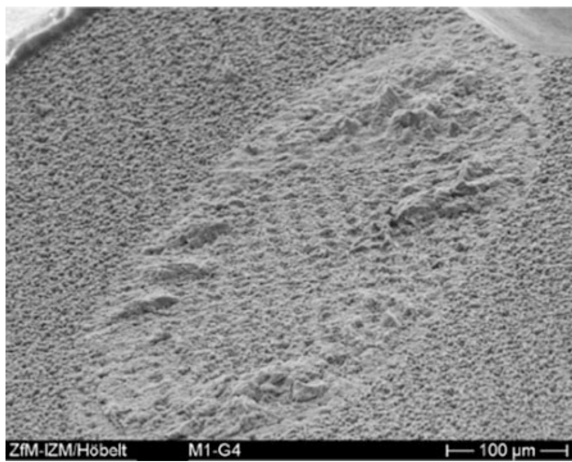


Fig. 12.15 Lift-off pattern as a result of power cycling

Special attention has to be paid to gate wire bonds on chips with a center gate contact. For power devices with a field effect gate structure, the leakage current of the gate is so small, that is takes days to discharge the gate via the leakage current. Therefore, a wire bond lift-off of the gate wire bond will not be noticed if the gate is continuously switched on and the load current is controlled by an external switch. In this case, a special functional test must be performed during the cooling phase of each cycle to verify the gate functionality. This can be done by switching of the gate voltage during the cooling phase. The constant current source, that supplies the sense current during this phase, must then go into voltage limitation mode. This technique ensures that a gate bond wire lift-off will not remain undetected.

It was found that bond wires have a higher lifetime if they are coated with a polyimide cover layer [Cia01]. Another proposal to increase the lifetime of Al wire bonds was the improvement of the bond loop geometry. Investigations on Al wire bonds under cyclic mechanical stress had shown that the lifetime increases with increasing loop height [Ram00]. Thus, increasing the aspect ratio, i.e., the ratio of loop height divided by the distance between the bond stitches, should increase the mechanical stability and therefore the lifetime in power cycling tests. However, first power cycling tests with increased aspect ratio of the Al wire bonds performed soon after publication of the mechanical tests did not confirm this expectation (Fig. 12.16).

It took another ten years until this discrepancy could be explained. Active power cycling tests were performed on baseplate power modules with Al wire bonds with different aspect ratios [Scn11]. Two groups of modules were produced for each aspect ratio, one group with soldered chips and one group with Ag diffusion sintered chips. A comparison of the test results for ΔT_{vj} = 70 K (T_{vjhigh} = 150 °C, T_{vjlow} = 80 °C, t_{on} = 1.2 s) displayed in Fig. 12.16 shows, that the chip solder fatigue limits the increase of lifetime in power cycling for higher aspect ratios. If

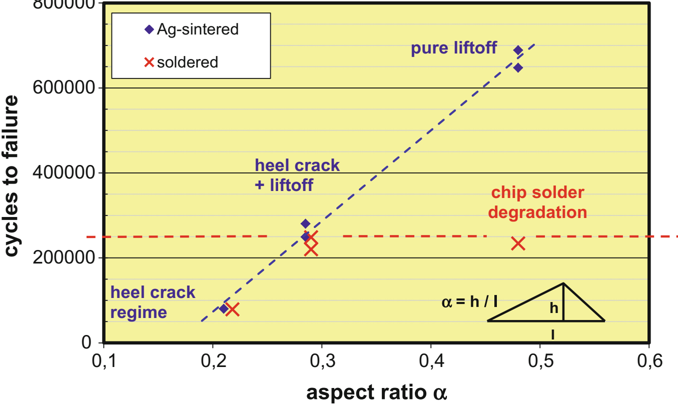
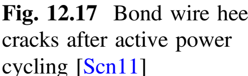


Fig. 12.16 Power cycling lifetime as a function of the Al wire bond aspect ratio at ΔT_j = 70 K for soldered chips and Ag diffusion sintered chips [Scn11]



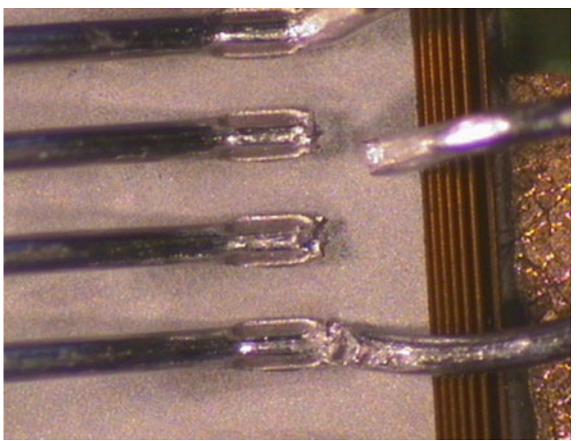


Fig. 12.17 Bond wire heel cracks after active power cycling [Scn11]

solder fatigue is eliminated by implementing Ag diffusion sintering for the chip-to-substrate interconnection, considerable gain in lifetime can be achieved with higher bond loops.

The test results revealed also that the failure mode of the Al wire bond is affected by the aspect ratio; while wire bond lift-off is the dominating failure mode for high aspect ratios, heel cracks become the predominant failure mode for low aspect ratios. As illustrated by Fig. 12.17, these heel cracks occur at the first stitch on the chip of the wire bond connection between chip and substrate.

The replacement of Al by Cu as wire bond material as described in Sect. 11.6.3 further enhances the power cycling lifetime of the topside chip interconnection. It has the capability to eliminate the failure mode of wire bond degradation in power modules completely and reveals a fatigue of the copper metallization of the DBC substrates as a new failure mode [Heu14].

12.7.2.2 Reconstruction of metallization

A phenomenon observed during active power cycles with a high temperature swing is the reconstruction of the topside chip metallization. This contact metallization is conventionally made of a vacuum-metallized aluminum layer, which is formed in a grain structure. Due to the difference in thermal expansion between silicon and aluminum, this layer suffers from a considerable stress during repeated temperature swings. While the silicon chip is only marginally expanding with increased temperature (2–4 ppm/K), the grains of the Al metallization expand considerably (23.5 ppm/K). Thus, the metallization layer is subjected to a compressive stress during the heating phase of temperature cycles.

The surface reconstruction of aluminum films on silicon was first reported in the late 1960s [Pad68] followed by detailed investigations of this degradation effect. Comparison between temperature cycled samples with annealed (uncycled) samples

for equivalent time-at-temperature revealed, that the surface reconstruction is increased by thermal cycling by a factor 2–5 depending on temperature and grain size of the aluminum film [San69]. Analysis of reconstruction phenomena at high temperatures (above 175 °C) and low temperatures (below 175 °C) suggest different fatigue mechanisms for these temperature ranges. Diffusional creep and plastic deformation involving conservative motion of dislocations are assumed as dominant contributions for high temperatures, while for low temperatures the only possible mechanism of mass transport is plastic deformation caused by compressional fatigue [Phi71].

While the former investigations were triggered by surface reconstruction observed in integrated circuits, the same effect was found in power electronic devices [Cia96]. The periodical compressive stress during the heating phase in temperature cycles results in plastic deformation of grains when the elastic limits are exceeded. According to [Cia01] this is the case for junction temperature above 110 °C. The plastic deformation can lead to the extrusion of single grains. This process leads to an increasing surface roughness of the metallization with the macroscopic observable effect of a dull non-reflective surface appearance. In the cooling phase of the temperature cycle, tensile stress can lead to cavitation effects at the grain boundaries if the elastic regime is exceeded. This can explain the observed increase in electrical resistance of the surface metallization [Lut08].

Reconstruction of the aluminum contact layer was also found in failed devices after repetitive short circuit operation of IGBTs [Ara08]. Recently, a similar effect of increasing surface roughness was observed also in thick aluminum layers of 'Direct Bonded Aluminum' (DBA) substrates after temperature cycling between −55 and 250 °C. After 300 cycles the surface roughness of 300 µm thick aluminum layers on AlN substrates increased by more than a factor of 10, while a multitude of voids where observed in a cross section. The authors attribute this effect to grain boundary sliding [Lei09].

Figure 12.18 shows the optical image of such a metallization reconstruction after the power cycling test. The reconstruction appears as a milky white discoloration of the diode metallization. It is concentrated at the center of the chip. Especially interesting is the area around the solder void, which can be seen in the X-ray image. The reconstruction is also very pronounced in this region, corroborating that the reconstruction is generated by the temperature swing. The maximum temperature has its peak in the center of the die, while the thermal resistance is locally increased in the area of the solder void. The dissymmetry of the reconstruction clearly follows this temperature profile [Scn99].

The impact of the maximum temperature during power cycling is illustrated in Fig. 12.19. Figure 12.19a shows an IGBT metallization after 3.2 millions of power cycles between 85 and 125 °C, the figure is taken from [Cia02]. Figure 12.19b shows an IGBT metallization after 7250 power cycles with $\Delta T = 131$ K and $T_{high} = 171$ °C. Figure 12.19c finally shows the metallization of a diode after 16,800 cycles at $\Delta T = 160$ K with $T_{high} = 200$ °C, where significant grains of approx. 5 µm diameter appear on the surface leaving voids in the metallization layer.

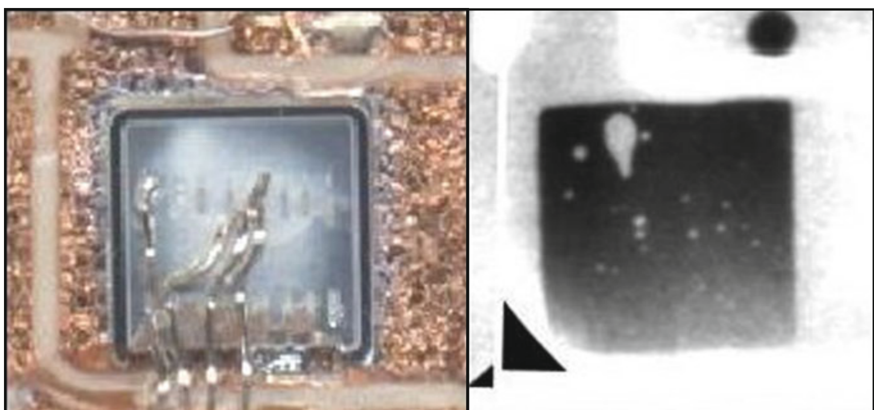


Fig. 12.18 Optical image (left) and X-ray image (right) of a diode after active power cycling – the reconstruction of the metallization appears as a milky non-reflecting discoloration, which is most pronounced in the chip center and in the area of the solder void [Scn99]

The reconstruction effect is suppressed beneath wire bond stitches. Figure 12.20 shows a detail of the metallization at the edge of a bond stitch after wire bond lift-off. This diode survived 44,500 cycles with $\Delta T_j = 130$ K, $T_{high} = 170$ °C. The high number of cycles was achieved because of single side silver sinter technology [Amr06]. Other investigations have shown that a polyimide cover layer suppresses the reconstruction effect as well [Ham01]. This is an expected phenomenon, because any cover layer will restrict the movement of the grains out of the contact layer. Nonetheless remains the high stress in the layer and it can be expected, that the initiation and the growth of fractures in the interface of bond stitches is driven by the same CTE mismatch that generates the metallization reconstruction.

The reconstruction of the chip metallization reduces the density of the contact layer and therefore increases the specific resistivity of the contact. Since the layer thickness is typically in range of 3–4 µm, the movement of grains sized in the range of the layer thickness as in Fig. 12.19c can be expected to change the conductivity of the layer considerably.

The layer resistivity can be measured according to the method of van der Pauw [Pfu76]. For the example of the power cycled diode in Fig. 12.19c, a specific resistance of 0.0456 mΩ m was measured. The resistivity of an unstressed diode of the same type was found to be 0.0321 mΩ m. This is close to the literature value for pure bulk Al of 0.0266 mΩ m, a slightly higher value can be expected because the chip metallization has a grain structure and contains some small admixture of Si. The observed resistivity increase of 42% during power cycling [Lut08] is a severe change of the device characteristics. Even though only a very small increase of V_F was detected during the power cycling tests of the devices in Figs. 12.19c and 12.20, an impact on the current distribution in the device must be expected which will reduce the lifetime of the power device under high stress conditions.

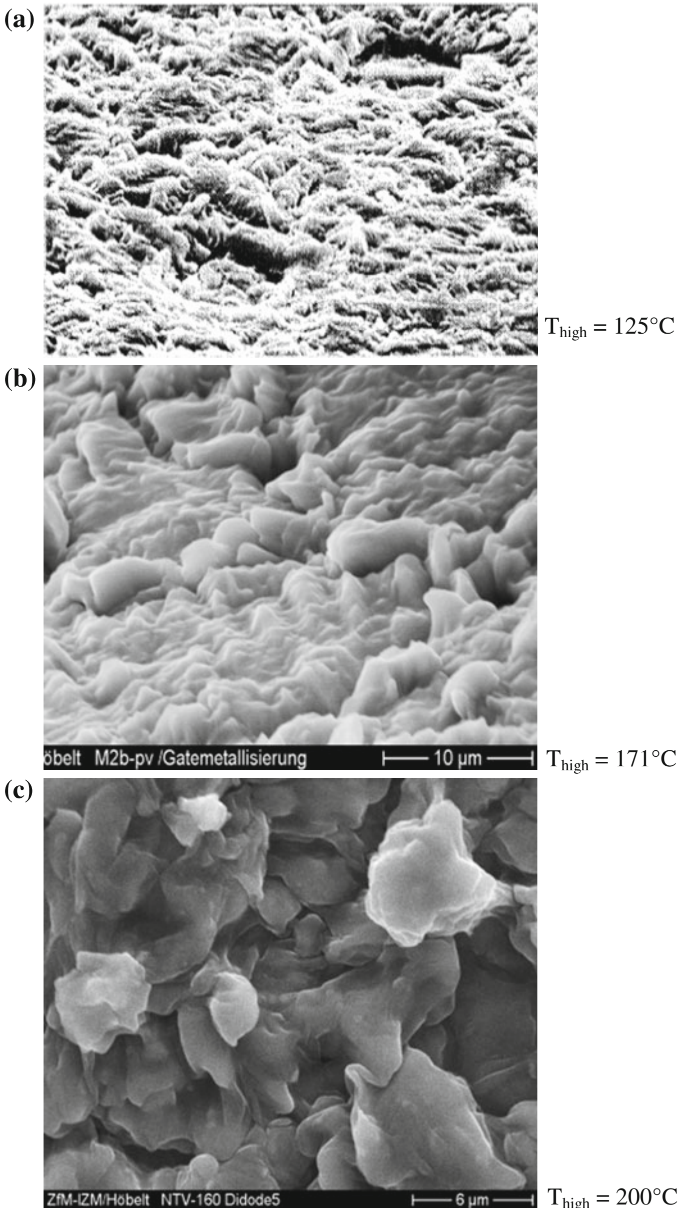


Fig. 12.19 REM images show the augmentation of contact reconstruction with increasing maximum cycle temperature T_{high} during power cycling

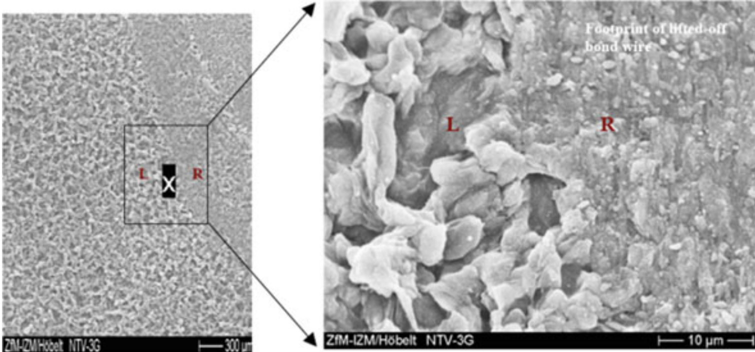


Fig. 12.20 REM image of the edge of bond stitch area after a wire bond lifted-off – the reconstruction is more pronounced outside of the stitch area. The diode failed after approx. 44,500 cycles with ΔT_j = 130 K and T_{high} = 170 °C

Latest investigations on device failure under repetitive short circuit exposure below the critical thermal destruction level have shown that the increase in contact resistance caused by reconstruction of the contact layer is the root cause of the failure [Ara08]. It must be expected, that this reconstruction also has an impact on the surge current capability of freewheeling diodes.

12.7.2.3 Solder Fatigue

The degradation of the solder interface is a fundamental failure mode during active power cycling. The so-called solder fatigue is caused by the formation of fractures in the solder interface, which leads to an increase in thermal resistance and is thus accelerating the total failure of the device. In devices with a positive temperature coefficient of the forward voltage drop (i.e. IGBT and MOSFET), the increasing temperature leads to increasing power losses and therefore expedites the degradation in a positive feedback loop. However, the final end-of-life failure of the conventional module design will typically be the breakdown of the wire bond contacts. If no investigation of the solder interface is performed by Scanning Acoustic Microscopy (SAM) after a power cycling test, the impact of solder fatigue cannot be evaluated and the wire bond breakdown is often mistakenly assumed to be the root cause of failure.

Figure 12.21 shows the evolution of the forward voltage drop and the maximum temperature reached at the end of each heating phase of the IGBTs in the center phase leg of a sixpack module without base plate. The maximum temperature starts to increase after 65,000 cycles and generates increasing power losses in the device due to the positive temperature coefficient. After 85,000 cycles, the temperature swing has already increased to ΔT_j = 125 K and it would continue to grow fast until the liquidus temperature of the solder is reached, if the wire bond contacts would not have failed first. This example illustrates, that only marginal lifetime can

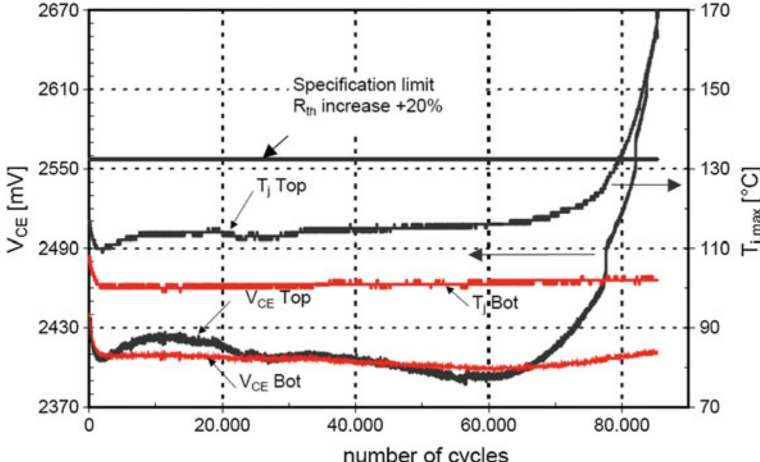


Fig. 12.21 Power cycling results of an IGBT phase leg in a module without base plate with ΔT_j = 80 K [Scn02b]. The temperature evolution is a strong indication of solder fatigue

be gained by improving the top side chip contact (i.e. the wire bonds) without improvement of the bottom side contact (i.e. the solder layer), because the increasing temperature amplitude would destroy any top side chip contact.

The geometry of the solder fatigue depicted in the SAM image in Fig. 12.22 follows the expected deterioration pattern. The discontinuity at the chip edges is responsible for a stress peak at the edges and especially at the chip corners. Therefore, the fractures start at the outside corners and edges and propagate towards the chip center. In this case, the arrangement of the four chips in parallel generates a temperature distribution with the maximum temperature being located at the chip corners close to the center of this quadruple chip assembly. Thus the degradation of the solder layers starts at the corners pointing toward the center of the group and moves outward.

The continuous improvement of the package thermal resistance and the enhanced heat extraction capability of advanced cooling systems have considerably increased the power density in modern IGBT chips. Assisted by the trend to minimized chip thicknesses (down to 70 µm for latest generation 600 V trench chips), which reduces the lateral thermal conductivity and thus diminishes possible heat spreading effects, the lateral temperature gradient in the chip has become more and more pronounced. Figure 12.23 illustrates this effect, where the diagonal temperature gradient exceeds 40 °C for a 12.5 × 12.5 mm^2 IGBT on a water cooled copper heat sink with 9 °C cooling liquid temperature [Scn09].

For such pronounced lateral temperature gradients, the stress in the chip center generated by the high temperature exceeds the stress induced by the edge discontinuity and the degradation begins in the center instead of the edges and corners. This is confirmed by the SAM image of a state-of-the-art power module after the power cycling test in Fig. 12.24. Here the damaged region is located in the chip

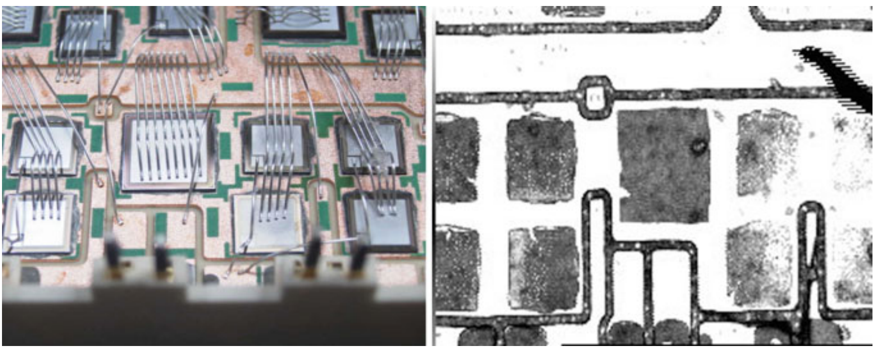


Fig. 12.22 Optical image and SAM image of the IGBT module in power cycling test shown in Fig. 12.21. The four IGBTs in parallel of the failed TOP switch (right side of the images) show the effect of solder fatigue

center, indicated by the light areas of high reflection, while the chip edges seem unaffected.

This observation was also reported from other authors. While some publications attribute this phenomenon to special solder types used [Moz01], other authors have observed this effect with lead free and lead-rich solder systems [Her07]. The characteristics of the solder interface layer might have a small impact on the solder fatigue progress; however the driving force for the deterioration of the solder interface results from the high stress induced by the differences in thermal expansion.

This change in the failure mechanism could have a considerable impact on the evolution of the degradation process. Fractures in solder interface increase the local thermal resistance of the affected chip region and thus raise locally the chip temperature. If the fractures start at the edges, the temperature of relatively cool chip regions is increasing, while the maximum temperature remains unchanged in the chip center. The situation is different, when the fractures start at the center, which has the highest temperature to begin with. Therefore, with fractures in the center of the chip, the maximum chip temperature is immediately increased and it can be expected, that this positive feedback loop will accelerate the fatigue progress and thus reduce the power module lifetime.

The considerable temperature gradients on the chip level also generate temperature gradients in the solder layer between the substrate and the base plate. Test results from a power cycling test at $\Delta T_j = 67$ K and $T_{high} = 150$ °C on a module with AlSiC base plate show degradation effects in the substrate to base plate interface, that start below the chip position as shown in Fig. 12.25.

These results illustrate the complexity of solder fatigue effects due to the interaction of thermal and mechanical characteristics of the interconnection interface. The problem of solder fatigue must be solved to exceed the limits in reliability of the classical module design.

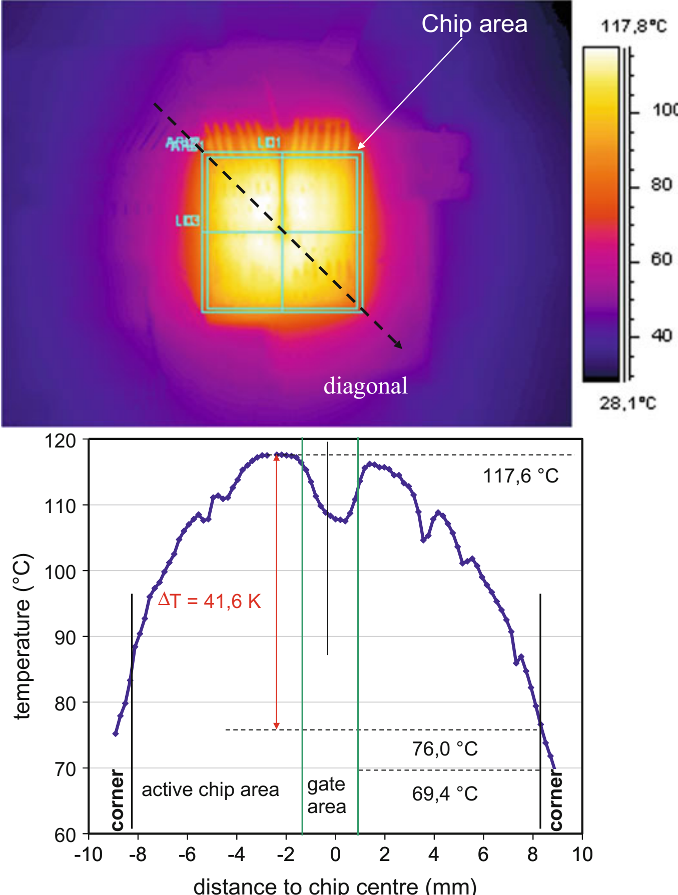


Fig. 12.23 Infrared image of a 12.5 × 12.5 mm^2 1200 V IGBT chip heated by a continuous DC current of 150 A. The module was mounted on a Cu water cooler with 9 °C cooling liquid temperature. The equilibrium temperature profile along a diagonal line through the chip center (bottom diagram) shows a temperature gradient between center and corner of more than 40 °C

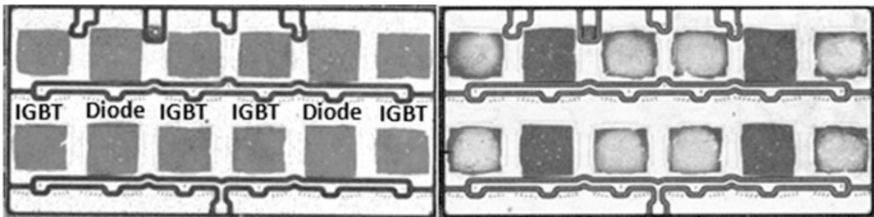
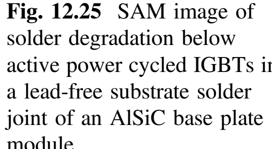


Fig. 12.24 SAM images of the initial solder layer (left) in comparison to the end-of-life image (right) after > 1 10^6(more than 1 million) power cycles of the IGBT with T_{high} = 150 °C, ΔT_j = 70 K, heating phase 0.2 s – the unstressed diode solder remains unchanged



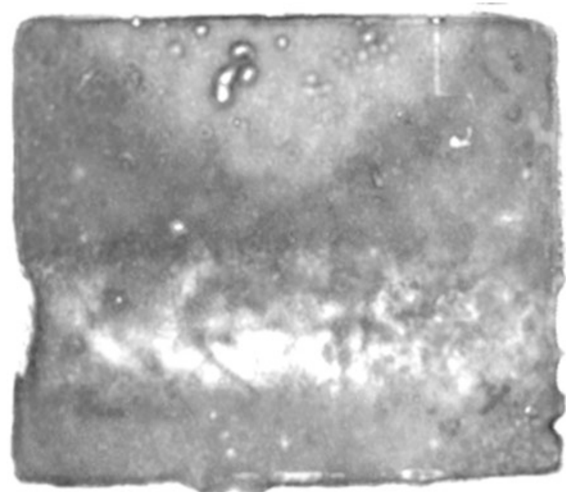


Fig. 12.25 SAM image of solder degradation below active power cycled IGBTs in a lead-free substrate solder joint of an AlSiC base plate module

12.7.3 Models for Lifetime Prediction

Since a standard technology is established for the construction of power modules with base plate and technologies and materials are very similar even for different suppliers, a research program for determination of lifetime for standard power modules was implemented in the early 90s. In this project named LESIT, modules from different suppliers from Europe and Japan have been tested; a common feature was the standard package according to Fig. 11.13 with use of an Al_2O_3 ceramics according to Table 11.1, left row. Tests were executed at different ΔT_j and different medium temperatures T_m, the results have been summarized in [Hel97] and are shown in Fig. 12.26 in form of a characteristics of cycles to failure N_f depending on ΔT_j and for different medium temperatures T_m.

The lines in Fig. 12.26 represent a fit from [Scn02b] to the experimental data: The expected number of cycles to failure N_f at a given temperature swing ΔT_j and an average temperature T_m, the absolute medium temperature in K, can be approximated with the equation

$$N_f = A \cdot \Delta T_j^{\alpha} \cdot \exp\left(\frac{E_a}{k_B \cdot T_m}\right) \tag{12.5}$$

with the Boltzmann-constant $k_B = 1.380 \times 10^{-23}$ J K^{-1}, activation energy $E_a = 9.89 \times 10^{-20}$ J and the parameters $A = 302{,}500$ and $\alpha = -5.039$.

Equation (12.5) consists of a Coffin-Manson law, i.e. the number of cycles to failure (N_f) is assumed to be proportional to $\Delta T_j^{-\alpha}$ [Hel97]. It appears as a straight line when plotting $\log(N_f)$ over $\log(\Delta T_j)$. In addition, an Arrhenius factor containing

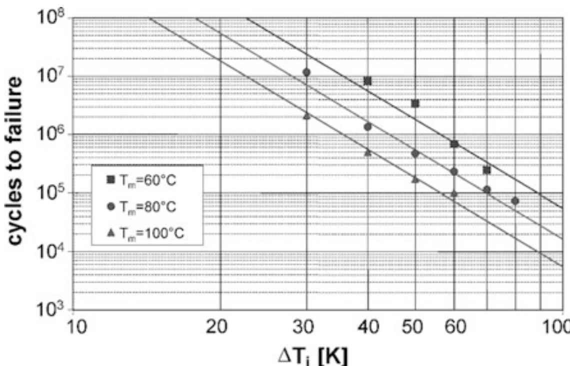


Fig. 12.26 LESIT-results

an exponential dependency on an activation energy is added to the Coffin-Manson law [Hel97].

This LESIT model was the first lifetime model which contained more parameters than the temperature swing ΔT_j. It can be considered as an empirical lifetime model, since it is based on experimental power cycling results on the one hand and it contains only parameters that are directly related to the test conditions on the other hand. The values for ΔT_j and T_m for the experimental data points in Fig. 12.26 are the initial test conditions of the associated power cycling tests, i.e., they represent the initial test condition before any impact of degradation can be observed.

Using Eq. (12.5) it is possible to calculate the number of cycles to failure for given ΔT_j and T_m according to the LESIT model. If the typical cycles in the application are known, it is possible to calculate the expectation for the lifetime of a module under these conditions.

Technologies for standard modules have been improved since 1997. Power cycling results for more recent power modules of two different suppliers are shown in Fig. 12.27. They are compared with the Eq. (12.5) for the condition T_{low} = 40 °C, extrapolated to higher temperature swings outside of the data range in Fig. 12.26. It is visible that the number of cycles to failure is increased by a factor of 3 to 5 in the range of ΔT_j > 100 K compared to the prediction of Eq. (12.5).

However, the temperature swing ΔT_j and the medium temperature T_m alone are not sufficient to uniquely characterize the test conditions of a power cycling test. The target ΔT_j is a function of the dissipated energy – which for a given chip technology is determined by the forward current and the t_{on} duration in the test – and of the thermal resistance of the test setup. Different combinations of load current, pulse duration and cooling condition can be found that generate the same temperature swing and medium temperature in a given power module. If the specific selection of test conditions have an impact on the test results they have to be included in a power cycling lifetime model.

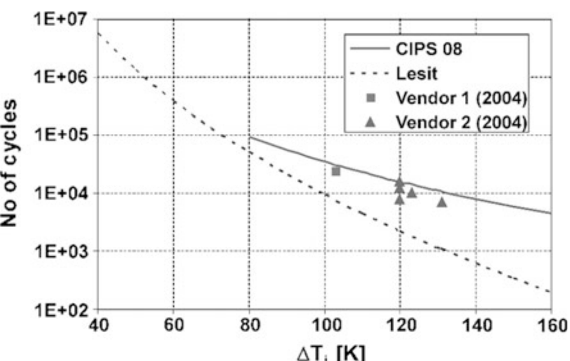


Fig. 12.27 Comparison of experimental power cycling results of state-of-the-art 2004 modules with predictions by the extrapolated LESIT model Eq. (12.5) and CIPS 08 model Eq. (12.6). T_{low} = 40 °C

This was the motivation to present an extended model for the lifetime of standard power modules [Bay08]. Based on a large number of power cycling results of modules, the following equation was derived:

$$N_f = K \cdot \Delta T_j^{\beta 1} \cdot \exp\left(\frac{\beta 2}{T_{low}}\right) \cdot t_{on}^{\beta 3} \cdot I^{\beta 4} \cdot V^{\beta 5} \cdot D^{\beta 6} \tag{12.6}$$

As parameter K we use the value 9.30×10^{14}, the other parameters ß2 … ß6 are given in Table 12.3 [Bay08]. Equation (12.6), which we denote as CIPS 08 model, contains additionally the dependence on the heat-up time t_{on} in seconds, the current per bond stitch on the chip I in A, the voltage range of the device V in V/100 (reflecting the impact of the semiconductor die thickness), and the bond wire diameter D in µm. The prediction by the new CIPS 08 model is also shown in Fig. 12.27 for t_{on} = 15 s. The CIPS 08 model holds for standard modules with Al_2O_3 substrates, it is not valid for high-power traction modules which are built with the materials AlN and AlSiC, see Table 11.1.

Equation (12.6) was a result of purely statistical analysis and is not a result of physics-based models [Bay08]. The dependency of the cycling time t_{on} in Eq. (12.6) – higher number of cycles to failure for short cycling times – may be explained by the fact that degradation mechanisms like viscoplastic deformation or fracture growth are time dependent processes, i.e., a short duration of the stress will generate less damage than a long duration. Furthermore, the temperature gradients in the module will depend on the load pulse duration. For short load pulses thermo-mechanical stress occurs mainly at the interface between semiconductor and

Table 12.3 Parameters for calculation of power cycling capability according to Eq. (12.6)

ß1	−4.416
ß2	1285
ß3	−0.463
ß4	−0.716
ß5	−0.761
ß6	−0.5

bond wire, while in layers closer to the heat sink the temperature increase is marginal and less stress is generated there. The dependency on current per bond stitch on the chip I can be attributed to improved current distribution on the chip with more bond stitches and presumably to a positive impact of the thermal capacity of the bond stitches. The dependency on the Al bond wire diameter D is related to the greater mechanical stress applied to the bond stitch by thicker bond wires. The dependency on the voltage class V (blocking voltage divided by 100) is in fact a dependency on device thickness, which increases from 600 to 1700 V. With thinner devices, the mechanical stress induced in the solder interface by the Si-material will be reduced. Note, that the used devices are produced in a thin-wafer technology for 1200 V ($V = 12$) and 600 V ($V = 6$). For devices fabricated from epitaxial wafers, e.g. PT-IGBTs or Epi-diodes, Eq. (12.6) is not applicable.

As a consequence of including additional test parameters (load pulse duration t_{on} and load current per bond stitch I) in the CIPS 08 model, the model parameters are no longer statistically independent, which was pointed out and discussed by the authors themselves [Bay08]. This can be illustrated by the fact, that it is in general not possible to adjust different values of the temperature swing ΔT_j in a series of power cycling tests while keeping the current magnitude and the load pulse duration constant. Although the variation of gate voltage allows to modify the dissipated losses for a given current in a limited range for IGBTs (see Fig. 10.3) or MOSFETs, this option is not available for power cycling tests with diodes. The influence of different heating times t_{on} is shown in Fig. 12.28. Figure 12.27 gives the impression that for a $\Delta T_j < 60$ K the new model predicts less cycles to failure N_f than the former LESIT model. But the dependency on the heating time shows that the lifetime N_f is higher for state-of-the-art 2008 modules, if a short heating time t_{on} for low ΔT_j is assumed.

Despite the fact that data for Eq. (12.6) were only generated with modules of one manufacturer, the equation seems also useful for lifetime calculation of modules of other manufacturers. If lifetime calculations are of vital importance in an application with high reliability requirements, the manufacturing company should always be consulted.

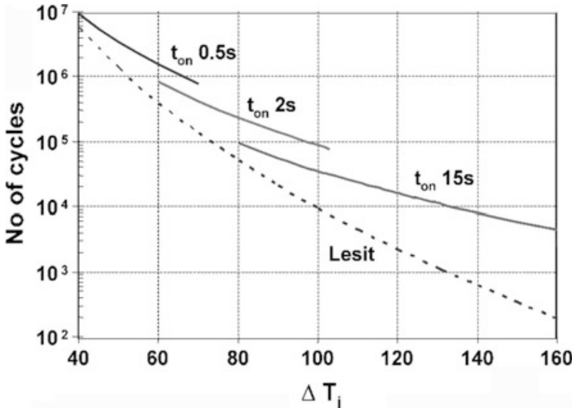


Fig. 12.28 CIPS 08 model according Eq. (12.6) for different heating times t_{on}, compared to the LESIT model. $T_{low} = 40$ °C

12.7.4 Separation of Failure Modes

The empirical lifetime models presented in Sect. 12.7.3 do not differentiate between the different failure modes discussed in Sect. 12.7.2. Since the materials involved in the degradation process of each failure mode are different, have distinct physical properties and feature divergent thermal behavior, each failure mechanism exhibits individual degradation driving factors resulting in different lifetime limitations. Thus, the extrapolation of the lifetime data gained by accelerated testing to the range relevant to applications may lead to wrong lifetime estimations, if the lifetime-limiting failure mechanism in the accelerated test differs from the failure mechanism dominant in the application. As a consequence, solder fatigue and wire bond degradation should be investigated independently to isolate the individual impact on empirical lifetime models. However, this was not feasible until advanced more reliable alternatives for these interconnections became available.

Today, high reliable interconnection technologies allow to isolate the degradation of different interconnections: by combining these advanced technologies with classical technologies, lifetime of classical interconnections can be studied without interaction. The SKiM63 lifetime model is based on several power cycling test results performed with the SKiM63 module from Semikron [Scn13]. This module comes without baseplate and is assembled in pressure contact technology [Scn08]. The inside connection technology is characterized by sintered chips and the topside die attachment is realized by aluminum wire bonds with an improved loop geometry. Since the load current terminals are implemented by pressure contacts and the auxiliary contacts by spring technology, the standard SKiM63 is a 100% solder-free module. Thus, the degradation of the wire bond connection to the chip determines lifetime. The SKiM63 lifetime model can therefore be regarded as an Al wire bond lifetime model. Following function describes the dependencies of the lifetime on the load conditions:

$$
\begin{aligned}
&n_f\left(\Delta T_j, T_{jm}, ar, t_{on}\right) = \\
&A \cdot \Delta T_j^{\alpha} \cdot ar^{\beta_1 \cdot \Delta T_j + \beta_0} \cdot \left(\frac{C + t_{on}^{\gamma}}{C+1}\right) \cdot exp\left(\frac{E_a}{k_B \cdot T_{jm}}\right) \cdot f_{diode}
\end{aligned}
\tag{12.7}
$$

A is a general scaling factor. The impact of the temperature swing ΔT_j in K and the medium junction temperature T_{jm} in K is represented by a Coffin-Manson relationship and an Arrhenius term with the activation energy E_a and the Boltzmann constant k_B, respectively. The wire bond aspect ratio ar, i.e. the ratio of loop height to distance between the stitches of the bond wire from DBC to the chip, was found to have an influence on lifetime as discussed in Sect. 12.7.2. The SKiM63 lifetime model also considers the impact of load pulse durations t_{on} in s. The factor f_{diode} reflects the influence of the chip thickness of the tested 1200 V diodes since the diodes were thicker and exhibited a lower lifetime than the tested 1200 V IGBTs. The derived equation parameters are presented in Table 12.4.

Table 12.4 Parameters of SKiM63 lifetime model in Eq. (12.7)

α	−4.923
β_1	-9.012×10^{-3}
β_0	1.942
C	1.434
γ	−1.208
E_a [eV]	0.06606
f_{diode}	0.6204

To investigate the lifetime of chip solder interconnections the method of separation of failure modes has been proposed [Sct12a]. Additional investigations [Jun15, Jun17, Sct13] focused on the impact of test conditions on the lifetime of the chip solder interfaces in comparison to the lifetime of the Al wire bond interconnections. For this purpose, two groups of modules were subjected to power cycling tests simultaneously; one group eliminates solder degradation by using the commercial SKiM63 module with sintered chips, the other group eliminates wire bond lift-off in a modified version of the SKiM63 module with SnAg3.5 soldered chips and high reliable Al-cladded copper wires [Sct12b]. Alternatively, other high reliable interconnection technologies like diffusion soldering for the chip interconnection to the DBC or copper wire bonds for the topside chip connection could be implemented, though the latter would require a copper metallization on the chip topside surface.

In [Sct13] load time t_{on} and temperature swing ΔT_j were adjusted identically in each test in order to investigate the impact of the medium junction temperature T_{jm} on the power cycling lifetime of the interconnections near to the chip. In [Jun15] t_{on} and either T_{jmax} or T_{jmin} were equal in all tests (T_{jmax} = 150 °C or T_{jmin} = 40 °C, t_{on} = 2 s) and solely ΔT_j and consequently T_{jm} were varied. The results (Fig. 12.29) revealed that sintered modules with standard aluminum bonds exhibit a lower lifetime at high ΔT_j than the soldered modules with Al-cladded copper bonds. Solder fatigue in contrast becomes the lifetime determining mechanism at low ΔT_j. Furthermore it was demonstrated that the mechanism of solder fatigue is much more affected by the medium junction temperature than the wire bond degradation. This is also reflected in the higher activation energy of E_A = 0.097 eV for the solder fatigue compared to E_A = 0.042 eV for the wire bond degradation. The divergent temperature dependency of both failure mechanisms can be explained by the thermo-mechanical behavior of the involved materials. During testing, the solder layers are subjected to temperatures near the melting point of the solder. Crack formation and propagation is accelerated with increasing temperature level. The melting point of aluminum in contrast is much higher. However, due to the large difference in the CTEs of the silicon of the chip and the aluminum of the wire bond, the interconnection between chip and wire bond is more sensitive to temperature swings as illustrated by the results.

In a further step the impact of load pulse duration was analyzed [Jun17]. The load pulse duration was varied in a wide range and the other test conditions were identical for all tests (T_{jmax} = 150 °C and ΔT_j = 70 or 110 K, t_{on} = 0.07 s…0.60 s). Figure 12.30a illustrates the lifetime results for load pulses $t_{on} \leq 2$ s and lifetime

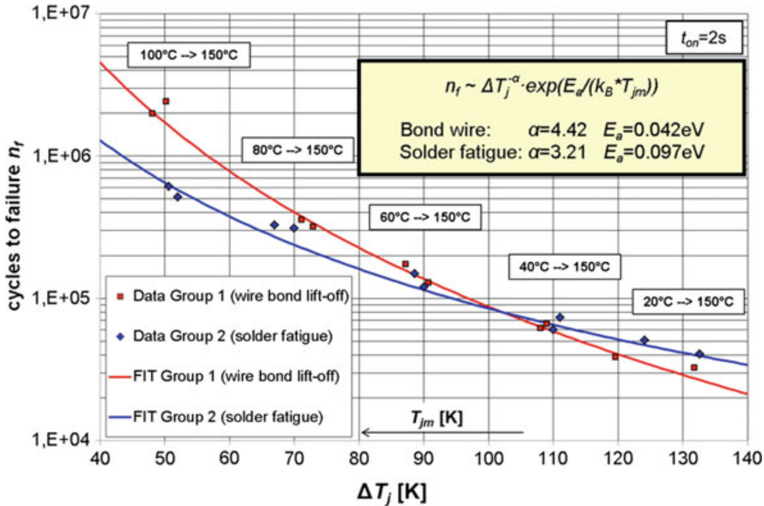


Fig. 12.29 Isolated lifetime of Al wire bonds and SnAg3.5 chip solder over ΔT_j (t_{on} = 2 s, T_{jmax} = 150 K) [Jun15]

results for $t_{on} \geq 2$ s are presented in Fig. 12.30b, along with the lifetime estimation by the SKiM63 lifetime model. The solder lifetime increases with decreasing load pulse durations, though not as pronounced as the lifetime of the wire bond interconnection. Both, the lifetime of solder and wire bond interconnection, decreases with extended load pulses but with a declining rate.

The collected power cycling data separated for Al wire bond degradation and SnAg3.5 chip solder fatigue can be considered as a data base for future improved lifetime models for classical power modules.

The experimental power cycling test results presented here were derived based on a specific architecture: the 1200 V SKiM63 module. It must be considered when applying this model to other module designs with baseplate, that potential baseplate related failure modes (e.g. fatigue of the substrate-to-baseplate solder interconnection) are not contained in these investigations.

However, the extended data base collected by the separation of failure mode investigations can be useful for scaling and testing physics-of-failure lifetime models derived from simulations based on constitutive material models.

Currently, many research groups are working on so-called physics-of-failure lifetime models. These lifetime models are implemented in a simulation procedure and are based on constitutive material models. While empirical lifetime models are mainly built on experimental results and statistics, the goal of simulation models is to reproduce the deformation of the material as an answer to the stress conditions and to describe the failure mechanism on material level in order to deduce impact on lifetime. The thermally induced mechanical stress is hereby considered as an indicator to activate and foster a degradation mechanism. As will be discussed in

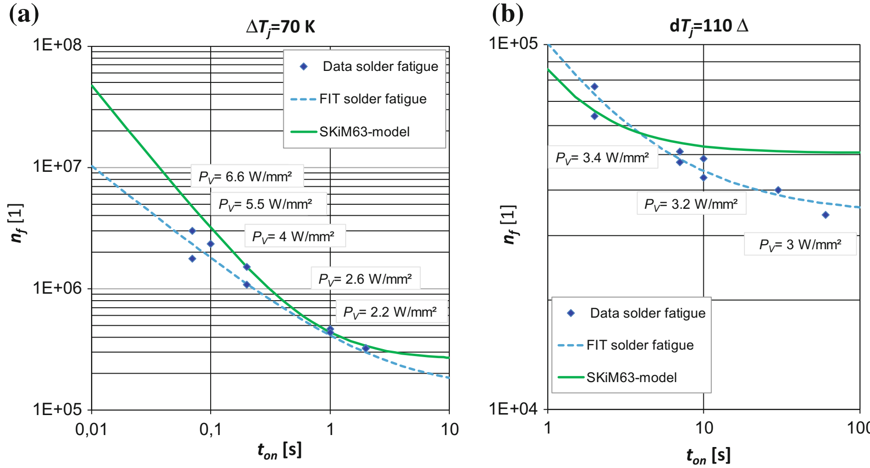


Fig. 12.30 Isolated lifetime of Al wire bonds and SnAg3.5 chip solder over t_{on} **a** ΔT_j = 70 K, T_{jmax} = 150 K, $t_{on} \leq 2$ s and **b** ΔT_j = 110 K, T_{jmax} = 150 K, $t_{on} \geq 2$ s)

the following examples, the simulation models need parametrization by experimental test results.

In [Kov15] a simulation model was proposed to describe lifetime of planar solder joints within power semiconductor modules. The accumulated deformation energy Δw_{hys} in the solder interconnection per cycle is calculated by a numerical approach based on Clech's algorithm [Cle97], which calculates the stress-strain hysteresis as material response to a given thermal load profile. For that, suitable solder constitutive equations describing the time-independent elastic and plastic deformation and the time-dependent viscoplastic deformation of the solder are required. With Morrow's type of fatigue law [Cle02], which defines a failure as exceeding a certain critical accumulated deformation energy W_{crit}, the lifetime for a given temperature profile in the solder layer can be estimated. Power cycling test results are required to parametrize the simulation model. The change of properties of the solder interconnection due to fatigue and the impact on the stress-strain hysteresis is not considered in the modeling process.

By contrast, crack propagation in solder joints was directly determined in [Dep06]. Using the crack growth model by Paris [Par63] crack initiation and propagation in solder joints were simulated in dependence on the accumulated creep strain along defined crack propagation lines. Latter was calculated by constitutive solder equations proposed by Déplanque in [Dep07]. The crack model was finally scaled on an empirical database which was generated using SAM-measurements of the crack length generated by thermal cycling.

Crack propagation models are also used to simulate fatigue in Al wire bond interconnections. In [Yag13] damage of the bonding interface is emulated by a damage-based crack propagation model. The analysis is performed in the time domain instead of a cycle-dependent description of the damage in order to capture

time sensitive effects. Furthermore, both damage accumulation and damage removal resulting from thermally activated processes are taken into account.

The presented constitutive lifetime model approaches require the knowledge of the temperature evolution inside the investigated interconnection. This can be obtained by finite element simulation for the case of repetitive cyclic testing. However, it is not possible to determine this internal temperature evolution for a complex mission profile of a real application, which might contain millions of non-repetitive load changes. Therefore, empirical models will be indispensable to evaluate complex mission profiles. On the other hand, it is impossible to validate empirical lifetime models derived from accelerated tests for very low stresses in applications operating for 20 years or more. In this regard, constitutive lifetime models could be very helpful to consolidate extrapolations to application specific stress levels from a material point of view.

12.7.5 Mission Profiles and Superposition of Power Cycles

Additionally to the already discussed restrictions, lifetime models are derived from the repetition of identical power cycles, but in real applications various different temperature swings are superimposed.

To calculate the lifetime under realistic application conformal conditions with superimposed power cycles, a linear accumulation of damage is assumed as discussed for example in [Cia08]

$$Q_i(\Delta T_j, T_m, \ldots) = \frac{N_i(\Delta T_j, T_m, \ldots)}{N_{fi}(\Delta T_j, T_m, \ldots)} \tag{12.8}$$

The parameters are given by the applied lifetime model. If the number of cycles N_i for a specific parameter set reaches N_{fi} – the number of cycles to failure for this parameter combination calculated by the lifetime model – Q_i is equal to 1. Based on this assumption – which is often called ‘Miner’s Rule’ [Min45] – contributions from different cycles in a complex mission profile can be summed up to obtain the accumulated damage. The inverse of the accumulated damage is the estimated repetition rate N_r for the considered mission profile (Eq. 12.9). If this repetition rate N_r is equal to 1 the lifetime is consumed by a single run of this mission profile. If it is larger than 1 the mission profile can be repeated N_r times; if it is smaller than 1 the power module will not survive a single run.

$$N_r = \frac{1}{\sum\limits_{i=1}^{n} Q_i} = \frac{1}{\frac{N_1}{N_{f1}} + \frac{N_2}{N_{f2}} + \cdots + \frac{N_n}{N_{fn}}} \tag{12.9}$$

An investigation with two superimposed power cycles is reported in [Fer08]. A cycle with high ΔT_j of 140 °C is superimposed with short cycles with low ΔT_j.

However, the cycle with high ΔT_j was dominant for the failure under the chosen superposition. It is generally difficult to define two superimposed power pulses in such a way, that each of them contributes with exactly 50% to the accumulated damage.

A different approach was reported in [Scn02b], where first conventional power cycling tests with uniform cycles were performed at two different temperature swings and then a test with both cycling condition interleaved was conducted. The test results showed that in contradiction to the assumption of linear fatigue accumulation, the number of cycles to failure was equal to the sum of cycles for each single test condition. A possible explanation for this result is that the test conditions are initiating different failure mechanisms, which do not interact with each other.

The method applied for cycle counting in application conformal temperature evolutions known as mission profiles is of fundamental importance for the lifetime estimation. The trivial approach of simply collecting the maximum-minimum temperature swings is underestimating the thermo-mechanical stress and it is very sensitive to the resolution of the analysis. A more robust approach, which is also consistent with the strain-stress characteristic in physics-of-failure oriented models, is the widely accepted Rainflow counting method [Dow82]. It is less sensitive to resolution changes and evaluates the fundamental frequencies with a higher weight.

The importance of an adequate cycle counting method can be illustrated by a simple example. Let us suppose a reliability engineer has a problem with the lifetime of a power module in his application. For a cycle with T_{low} = 40 °C and T_{high} = 120 °C a lifetime of 5.1×10^4 cycles can be expected from the LESIT model in Eq. (12.5). The application requires 8 times the number of cycles, so the engineer has an idea: He adds a small temperature swing in the slopes of his large temperature swing so that he divides it in smaller swings as shown in Fig. 12.31.

Applying the trivial counting method and using Eqs. (12.5), (12.8) and (12.9), the engineer calculates for his modified cycle in Fig. 12.31b the repetition rate as displayed in Table 12.5. Since he reduced the maximum temperature swing from ΔT_j = 80 K to ΔT_j = 42 K by his small modification, he now calculates a repetition

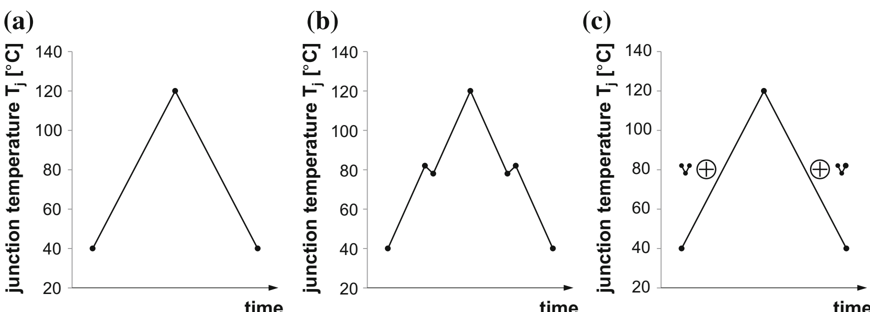


Fig. 12.31 Simple temperature cycle (**a**), modified cycle (**b**) and decomposition of modified cycle with the Rainflow algorithm (**c**) (lines are only guide to the eye)

Table 12.5 Lifetime calculation for the example of superimposed cycles in Fig. 12.31 with the LESIT lifetime model in Eq. (12.5)

Profile	T_{low} (°C)	T_{high} (°C)	N_f	Q_i	$\sum Q_i$	N_r
Simple cycle	40	120	5.1×10^4	2.0×10^{-5}	2.0×10^{-5}	5.1×10^4
Modified cycle trivial counting	40	82	4.1×10^6	2.4×10^{-7}	2.4×10^{-6}	4.2×10^5
	78	120	4.6×10^5	2.2×10^{-6}		
	78	82	1.8×10^{11}	5.5×10^{-12}		
Modified cycle rainflow counting	78	82	1.8×10^{11}	5.5×10^{-12}	2.0×10^{-5}	5.1×10^4
	40	120	5.1×10^4	2.0×10^{-5}		
	78	82	1.8×10^{11}	5.5×10^{-12}		

rate $N_r = 4.2 \times 10^5$ which would meet his requirements. However, by comparing the temperature profiles in Fig. 12.31a, b it seems not plausible that such a small modification would increase the lifetime by more than a factor 8.

This problem is resolved by applying the Rainflow algorithm to the modified profile in Fig. 12.31b. The algorithm decomposes the modified profile as shown in Fig. 12.31c resulting in a fundamental cycle with $\Delta T_j = 80$ K and two superimposed small cycles with $\Delta T_j = 4$ K. Since the damage contributed by these small cycles can be neglected, the repetition rate of the modified cycle is the same as the original simple cycle.

Based on the assumption of linear damage accumulation resulting from thermal cycles, together with the extraction of temperature cycles by a Rainflow algorithm and the lifetime model, application specific mission profiles can be evaluated as shown in Fig. 12.32. The process begins with the specific requirements of the application, e.g. the evolution of rotational speed and the torque for a motor drive or the wind distribution as a function of time for wind turbines. Then the dynamical characteristics of the motor or generator has to be taken into account, resulting in a sequence of values for the time step, the output voltage, the output current, the output frequency, the phase shift between voltage and current, the switching frequency and the DC link voltage. This sequence is the electrical mission profile for a specific application. The generated losses can be calculated with the power module data sheet for a known temperature. This requires an iteration loop, because the dynamical characteristics of the cooling system has to be taken into account for the calculation of temperatures. From the resulting temperature profile the cycles of junction temperature can be extracted by the Rainflow algorithm. The accumulated damage of the different cycles is then evaluated by the lifetime model and a repetition rate for the characteristic mission profile can be determined as an indicator for the estimated lifetime.

When the thermal characteristics of the devices in a power module and heat sink including interface are estimated by one dimensional Foster networks, mission profiles of a million time steps can be calculated within minutes on a state-of-the-art laptop.

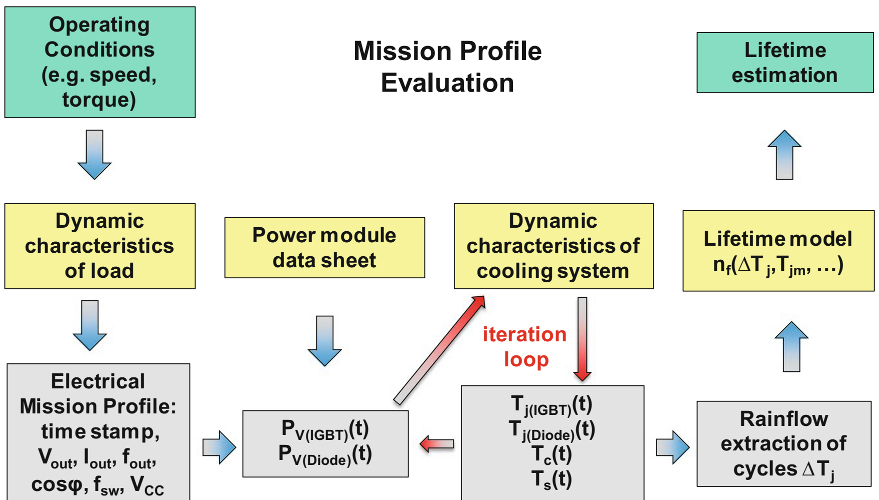


Fig. 12.32 Mission profile evaluation process [Scn15a]

However, a consistent definition of the junction temperature is required for this estimation process. This requirement can be fulfilled by applying the *Vj*(*T*) method (see Sect. 12.7.1) for the thermal impedance measurement as well as for the junction temperature measurement in power cycling tests. Furthermore, the control strategy for power cycling tests should provide no compensation for degradation. By defining the initial temperature swing ΔT_{vj} (after stable thermal conditions are reached) and initial losses as the characteristic test parameters, all degradation effects are inherently included in the lifetime model. If this is not the case, power losses and thermal impedance of the system will become a function of consumed lifetime, which makes the mission profile estimation process considerably more complex.

Additionally, the lifetime model must provide the number of cycles to failure as a function of operating parameters like virtual junction temperature, current, voltage and load pulse duration. It therefore has to be an empirical lifetime model. Constitutive models, which require the knowledge of the temperature in interconnections cannot be implemented in this mission profile estimation method. When FEM simulation is required to obtain input parameters at each time step, just the iteration loop for the temperature calculation would consume minutes to hours of calculation time so that the evaluation of extended mission profiles is not feasible. However, a serious problem of empirical lifetime models remains: they cannot be validated for very low temperature swings. Even if a test would be conducted with a predicted duration of 20 years, the technological progress would render any result obsolete. If constitutive lifetime models could be developed, which correctly predict the experimental results of accelerated cyclic tests for a wide range of test conditions, they could supply supporting evidence for the extrapolation of empirical lifetime models.

When the boundary conditions are fulfilled, lifetime estimation based on mission profiles is a powerful method to validate that a system design meets the lifetime requirements. It should be noted, that this estimated lifetime only considers the impact of application specific thermo-mechanical stress. Other lifetime relevant factors like high humidity, corrosive environments and cosmic ray failures can lead to a much lower lifetime. Therefore, this estimated lifetime defines the upper limit of the lifetime of power modules in a specific application.

12.7.6 Power Cycling Capability of Molded TO Packages

A high power cycling capability was found for DBC-based transfer molded TO-housings, which were described in the context of Fig. 11.6 [Amr04]. The power cycling results in Fig. 12.33 were gained in power cycling test with this package type for ΔT = 105 °C and T_m = 92.5 °C. The number of cycles to failure (Weibull 50% accumulated probability) is about a factor of 10 higher than predicted by the LESIT model (Eq. 12.5) and still significantly above the CIPS 08 model (Eq. 12.6). The statistical methods will be discussed in detail in Sect. 12.9.

Figure 12.34 shows a picture of bond wires in a device which survived 75,000 power cycles under said conditions. On the dark area in the foreground, a bond foot was initially attached. In the first of the still attached bond wires, a heel crack is visible.

The stiff mold-material has a similar effect as a bond wire coating; it additionally hinders mechanically the detachment of bond wires from the chip surface. Even though the bond wire shows signs of heavy deterioration, the electrical contact is still maintained. In contrast to standard TO packages with copper lead frames, the implementation of Al_2O_3 substrates accounts for a lower thermal mismatch between the semiconductor material Si and the assembly layer.

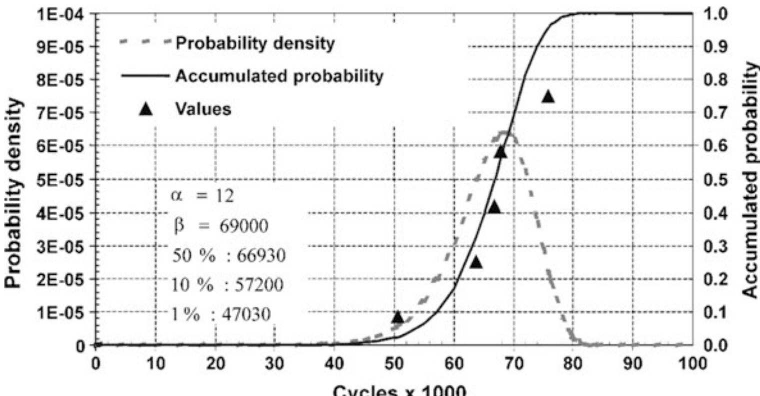


Fig. 12.33 Weibull analysis of a set of power cycling tests on a series TO package

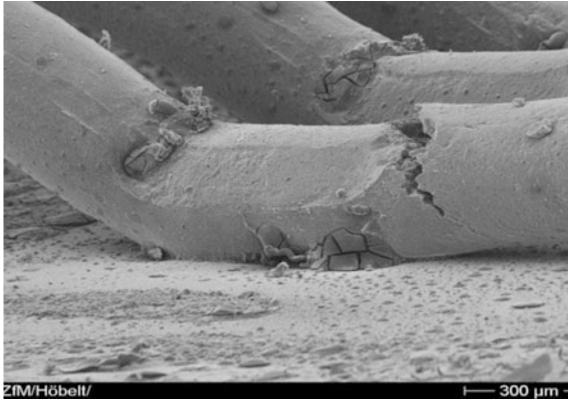


Fig. 12.34 Failure analysis of a DCB-based soft-molded TO package after 75,000 power cycles with $\Delta T = 105$ K, showing a footprint of a detached bond wire in the foreground (left). and a heel crack in a still attached bond wire

The classical TO package with Cu lead frames (Fig. 11.5) exhibits a great mismatch in thermal expansion between copper and silicon. For large chip sizes, the power cycling capability was found to be clearly inferior to packages with ceramic substrates. In a power cycling test of standard TOs with chips of 63 mm^2 area, two of six chips lost their blocking capability after only 3800 cycles with $\Delta T_j = 110$ °C and $T_m = 95$ °C. The failure analysis of these TOs revealed cracks in the silicon device as the root cause. Figure 12.35 shows such a fracture of a silicon chip.

Even though, no more failures occurred up to more than 38,000 cycles when the test was continued with the four remaining samples, the early failures cause by fractures in the silicon device are alarming. Presumably, the relatively large area of the silicon devices is responsible for these early failures, because this effect was not observed for power cycling tests of smaller chips (< 30 mm^2) in the same package type.

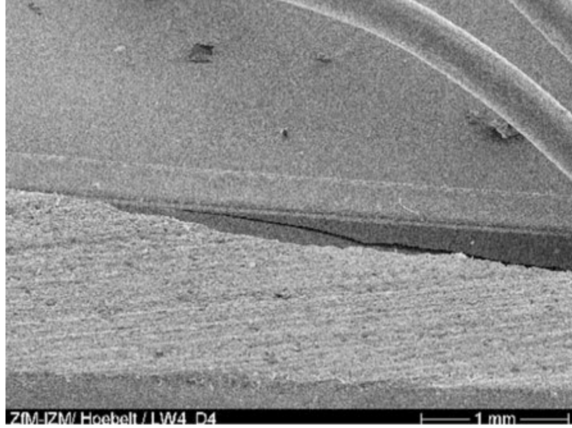


Fig. 12.35 Crack in the silicon diode chip in a transfer mold TO package on a Cu leadframe after 3800 cycles with $\Delta T_j = 110$ K and $T_m = 95$ °C

12.7.7 Power Cycling of SiC Devices

When we compare the material parameters of Si and SiC, a significant difference in the mechanical characteristics emerges. The Young's modulus, which represents the stiffness of the material under mechanical stress is about three times higher in SiC (501 GPa in SiC compared to 162 GPa in Si). The temperature swings and thermal mismatch induce mechanical stress into the package which leads to fatigue. For evaluation of the expected effect in a solder layer, the plastic strain energy density ΔW according to the method of Darveaux [Dar02] is often used. An investigation of Si and SiC chips in power modules under thermo-mechanical stress was performed by FEM simulation [Pol10]. The results give evidence that crack propagation in a solder layer with SiC chips will happen about 3.5 times faster than with Si chips, and – if we use an inverse proportionality as first order approximation – the power cycling lifetime will be shortened by this factor.

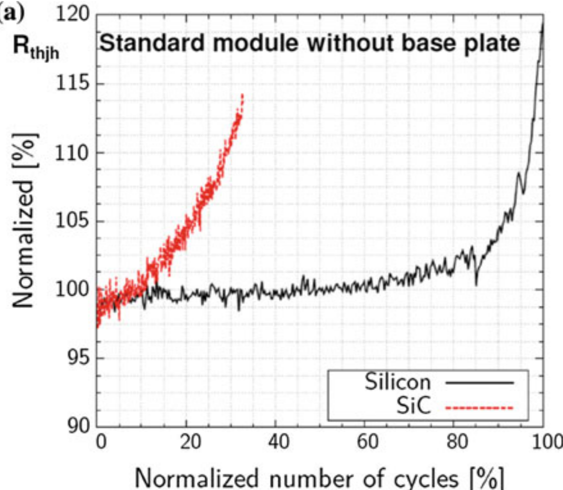
For execution of power cycling tests with Schottky diodes, the junction voltage can be used as temperature sensitive parameter (TSEP) for junction temperature determination, since the current voltage characteristic Eq. (6.3) discussed in Chap. 6 has close similarities to that of a pn-junction. Thus, the same test procedure as described in Sect. 12.7.1 can be used.

Figure 12.36 shows a comparison of Si IGBTs and SiC Schottky diodes at similar power cycling conditions using the same power module package without baseplate [Hed14]. Figure 12.36a compares the course of thermal resistance for one selected IGBT and one Schottky diode. The SiC device reaches 1/3 of cycles to failure compared to the Si device. Figure 12.36b shows a metallographic preparation of a solder layer in a failed SiC device. A crack starting at the edge of the solder is propagating towards the center, this is the root cause for the result in Fig. 12.36a. Figure 12.36c displays finally a summary of all tests, which again indicates that SiC devices reach approx. 1/3 No. of cycles to failure compared to Si devices in the used standard packaging technology. This confirms the simulation results of [Pol10] and shows: To reach the same power cycling capability, SiC devices need a packaging technology with more effort. For SiC MOSFETs the definition of a suitable TSEP is difficult because some parameters show a drift which makes them unsuited as a reliable temperature indicator [Ibr16]. The on-state resistance $R_{DS,on}(T)$, which is strongly temperature dependent, is not suitable since it will increase at bond wire failures and separation of degradation effects will be complex. Further, $V_G - V_T$ which enters $R_{DS,on}$ according to Eq. (12.1) is influenced by a gate threshold voltage $V_T(T)$ drift.

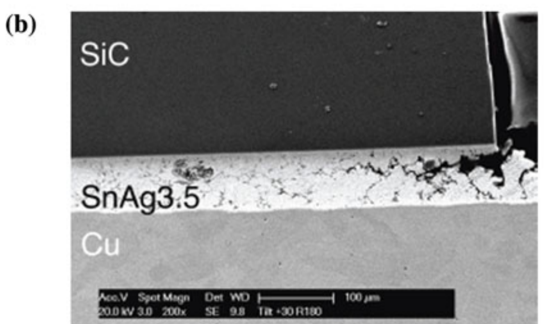
The gate threshold voltage V_T is strongly temperature dependent but affected by a trapping phenomenon: After a power cycle, it is found that SiC MOSFETs need up to seconds to recover to the initial V_T value. The effect is shown in Fig. 12.37. For the MOSFET of manufacturer #1, an increase of 30 mV (measured 1 ms after power pulse) is found, for manufacturer #2 a shift of 140 mV is detected and the decay back to the initial value takes up to seconds. Used as TSEP, a measurement error of 26 K would result for this device. Additionally, a power cycling induced V_T

Fig. 12.36 Power cycling of a soldered 600 V SiC-Schottky diode, compared to a 1200 V IGBT with the same technology, ΔT_j = 81 K ±3 K and T_{jmax} = 145 °C **(a)** Thermal resistance R_{thjc} in dependence on number of cycles **(b)** Metallographic preparation of a solder layer in a failed SiC device. Preparation by Infineon Warstein. **(c)** Comparison of cycles to end-of-life. Figs. from [Hed14]

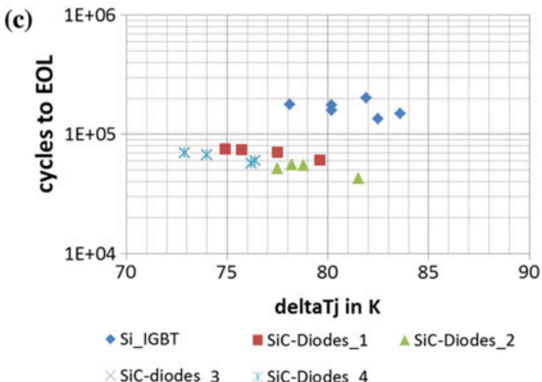
(a)



(b)



(c)



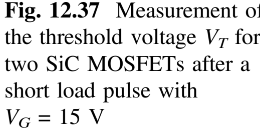
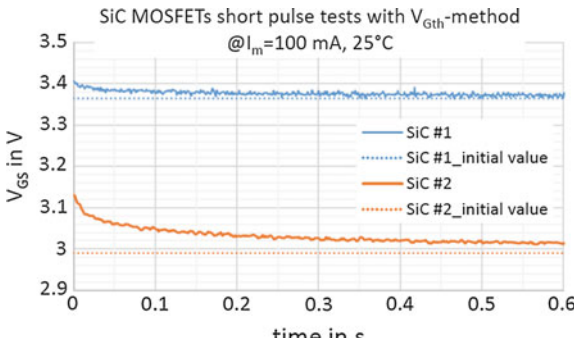


Fig. 12.37 Measurement of the threshold voltage V_T for two SiC MOSFETs after a short load pulse with $V_G = 15$ V



shift with SiC MOSFETs is possible. Based on these findings, V_T is regarded not suitable as TSEP.

Next, there is the voltage drop of the inverse diode $V_{SD}(T)$ at low measurement current. However, the SiC MOSFET's gate-channel is not completely off at $V_G = 0$ V. The voltage drop up to the built-in voltage of the pn-junction opens the channel partially and enables part of the current to pass the marginally inverted channel. The current-voltage characteristic of the inverse diode is depending on gate voltage as shown in Sect. 9.12.3, Fig. 9.32. It indicates that a gate voltage of −6 V or lower is required to obtain a stable characteristics of the diode and the junction voltage $V_j(T)$ can be measured.

Since the current through the MOS channel will depend on the threshold voltage, it must be ensured that the channel is safely turned-off. Therefore, a negative gate voltage below −6 V is recommended. Figure 12.38 shows $V_j(T)$ for a SiC MOSFET measured with $V_G = -10$ V.

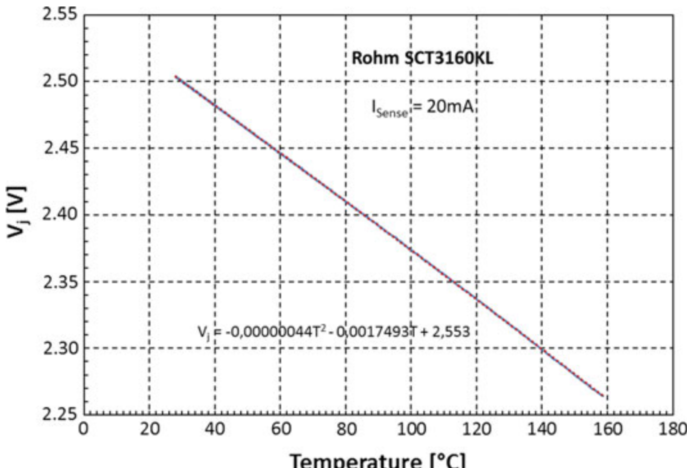


Fig. 12.38 Calibration function $V_j(T)$ for a 17 A 1200 V SiC MOSFET (Rohm) at $V_G = -10$ V

For power cycling of MOSFETs, the automotive standard [LV324] allows to use the body diode to generate the power losses. However, in MOSFETs of Si as well as of SiC the on-state losses of the inverse diode decrease with increasing temperature, while in forward operation the losses increase with temperature. Using the inverse diode results in a reduction of power losses when the temperature increases. This does not resemble the typical application. Therefore, it is recommended for SiC MOSFETs to use the drain-source forward current to create the power load.

The recommended test method for MOSFETs is creating power losses in forward mode and measuring $V_j(T)$ in reverse mode. Figure 12.39a shows the setup for one device under test (DUT), there may be several devices arranged in series connection. In series with the current source for I_{sense} are some diodes for protection. Figure 12.39b shows the control signals applied during a cycle. First, V_G is set on with the specified voltage V_{Guse} of the manufacturer. Next, the auxiliary switch is closed. Now the load current flows. At the end of the load pulse the auxiliary switch is turned off. After a short delay of 1–50 µs, a negative voltage is applied (here −7 V) for the temperature measurement via V_j of the inverse diode.

For measuring V_j, there is again a delay time t_d between turning-off the load current and the moment of measurement, compare Fig. 12.11 The cooling-down during t_d is higher for SiC devices compared to Si devices. For SiC, it was found in the range of 4–5 K, even up to 6 K for a t_d of 1 ms, due to the usually higher power densities [Hed16]. This is significant now, especially if packages with SiC and Si are compared. A correction is possible with the square-root-t method [Bla75]

$$T_{vj}(t) - T_{vj}(0) = \frac{2 \cdot P_v}{A \cdot \sqrt{\pi \cdot \rho \cdot \lambda \cdot c_{spec}}} \cdot \sqrt{t} \tag{12.10}$$

Equation (12.10) holds for the boundary condition of a planar heat source at the surface of a semi-infinite thick cylinder, assuming one-dimensional heat flow. Since the dominant heat source in SiC devices is in a narrow region close to the device surface, Eq. (12.10) was confirmed to be applicable for SiC devices with good accuracy [Hed16].

For an exact evaluation of power cycling tests with SiC devices, the selected delay time t_d between load current turn-off and T_{vj} measurement should be added in the documentation. It has to be mentioned when the measured values have been corrected with a Z_{th}-model or another method.

For SiC MOSFETs, a test with the method described in Fig. 12.39 with prototypes of 250 A 1200 V power modules is reported in [Hed17].

The $V_j(T)$ method with V_G = −6 V was used as temperature sensing method in a power cycling test in [Sct17]. For a new packaging technology with silver sintering and 125 µm aluminum bond wires topside, a very high power cycling capability can be achieved. The evolution of T_{vjhigh} in a power cycling test with constant t_{on} and t_{off} and ΔT_j = 110 K is shown in Fig. 12.40. The module finally fails after more than 1.1 million cycles due to substrate failure.

The results of [Sct17] show that an excellent power cycling capability can be achieved with SiC devices, if improved packaging technologies are applied.

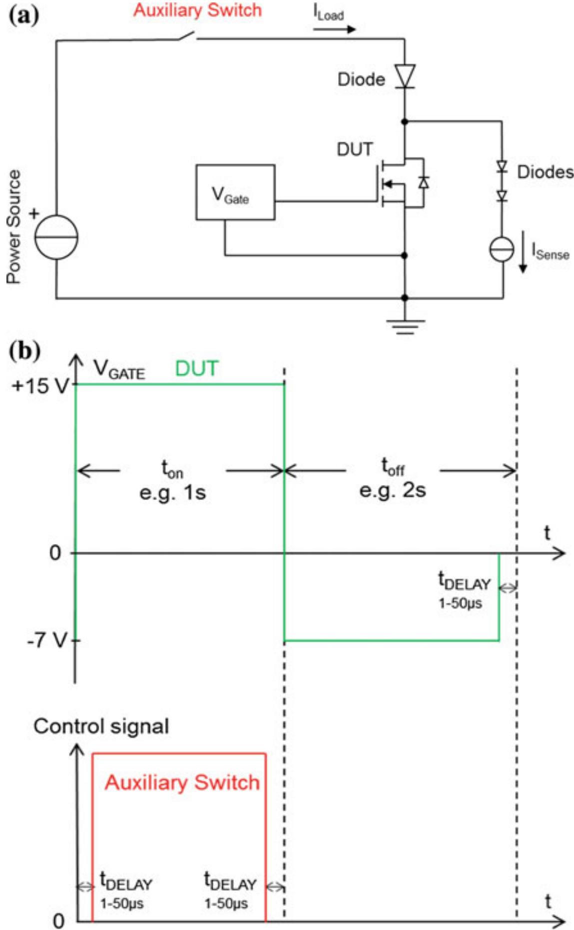


Fig. 12.39 Power cycling of SiC MOSFETs. **a** Test setup, **b** pulse pattern at the MOSFET and at the auxiliary switch

However, before design-in into a reliability-sensitive application, the power cycling capability of any new solution has to be proven.

The German automotive companies have established a mutual delivery specification for the qualification of power electronic components for automotive applications – including test methods and conditions for power cycling tests [LV324]. Even though, this delivery specification needs to be extended regarding the latest research results on Si and SiC MOSFETs discussed here, it is an important milestone on the path to a standardization of power cycling test methods and documentation. An international standard for power cycling tests, which is accepted by the major players in the field, would help to render power cycling test results more transparent and comparable.

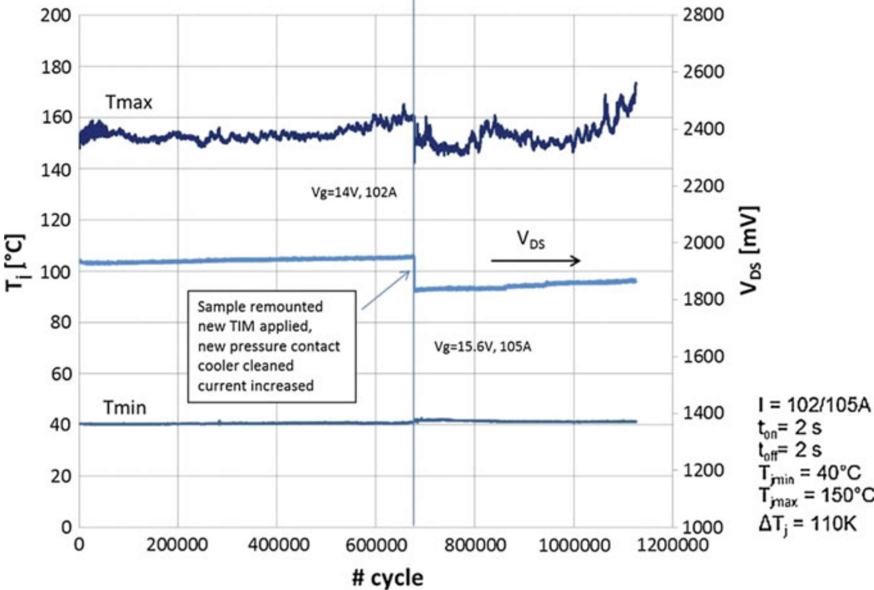


Fig. 12.40 Power cycling test with a SiC-MOSFET module: Evolution of V_{DS}, T_{vjlow} (*Tmin*) and T_{vjhigh} (*Tmax*). Figure from [Sct17], PCIM Europe 2017

12.8 Cosmic Ray Failures

12.8.1 The Salt Mine Experiment

With the introduction of high voltage semiconductors with turn-off capability in converters for electric traction in the beginning of the 1990s, failures in the application were observed which could not be explained with the available knowledge at that time. The failures occurred during the blocking mode of the devices. The application conditions were employed in the lab, and long term tests with high DC voltage in blocking direction were carried out. The tests confirmed the occurrence of spontaneous failures [Kab94]. The spontaneous character of the failures was strange, with no prior indications in device behavior, e.g. an increase of the leakage current. In some English literature, a cosmic ray failure is denoted as “Single Event Burnout” (SEB) [Alb05].

Figure 12.41 shows the results of the salt mine experiment. First, 6 failures occurred in the lab within 700 device hours. The test was interrupted and continued in a salt mine, with 140 m of solid rock above the test location. Under these conditions, no failure occurred. The test in the salt mine was interrupted again and continued in the lab, and now failures reappeared at a similar rate as in the lab before. The test setup was moved to a place in the cellar of a high-rise building with 2.5 m of concrete in total above the test location. The failures continued to occur but at a reduced rate.

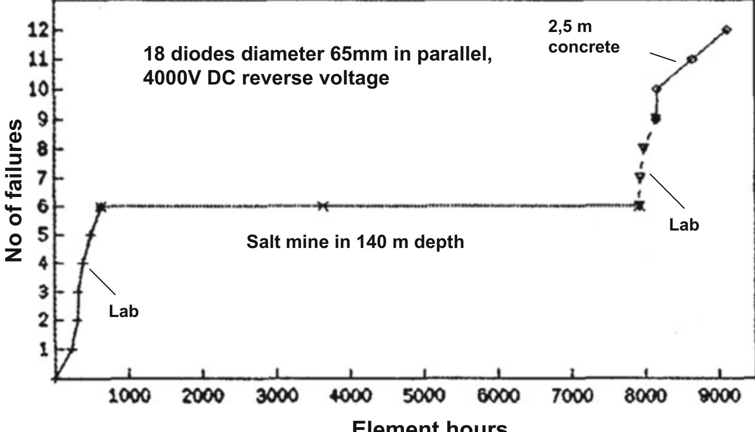


Fig. 12.41 Results of the salt mine experiment. Number of failures over the accumulated time of DC voltage stress. Figure from [Kab94] © 1994 IEEE

In addition to [Kab94], two further research groups [Mat94, Zel94] published the effect at the same conference. [Mat94] showed that the failure rate decreases by more than one decade with shielding by 2 m thick concrete. With these results, it was proven that cosmic ray was the root cause for said failures of power semiconductor devices under high voltage stress.

12.8.2 *Origin of Cosmic Rays*

The primary cosmic radiation consists of high energy particles: 87% protons, 12% alpha-particles and 1% heavy nuclei. Ultra-high energy photons are also part of the cosmic radiation. For particles up to 10 GeV (10^{10} eV), our sun is considered to be the main source. Up to 10^{16} eV, supernova eruptions are assumed as the main source, up to 10^{18} eV the source is expected to be located in our galaxy. Above 10^{18} eV, supernovae cannot explain the extreme energy. The cores of distant active galaxies are potential candidates for the origin of particles with extreme high energy.

A high-energy primary cosmic ray particle usually does not reach the surface of the earth directly; it rather collides with atom nuclei of atmospheric molecules. Thereby, it generates a variety of secondary high energy particles and this particle shower arrives at the earth surface as terrestrial cosmic radiation. A single primary high-energy particle can create a shower of even 10^{11} secondary particles. Most relevant for SEB at usual terrestrial altitudes are neutrons, the neutron flux density is in the range of 20 cm^{-2} h^{-1} at sea level [Nor96] and increases strongly with altitude. At 12.2 km (40,000 ft, the upper flight level of civil airplanes) there is a

neutron flux density of 7200 cm^{-2} h^{-1} for a latitude of 45° [IEC10]. The maximum neutron flux density is found at an altitude of 18 km. Further relevant parts are high-energy protons, they contribute with 20–30% at sea level and 50% at 12.2 km altitude to the total cosmic radiation. Creation and absorption compete in the showers in higher atmosphere. Also short-living high-energy particles like pions are generated, which decay to further particles. The pion flux is estimated as 1–3% of the flux of protons [IEC10].

Regarding space application, proton flux is the majority of energetic particle flux at low earth orbit [Das17]. Protons as charged particles can be shielded, however aluminum shielding of 2.5 mm thickness is no proper protection against proton flux, since a large fraction of the proton flux exhibit a high energy above 100 MeV range [Das17]. At sea level, usually only neutrons are considered for triggering cosmic ray failures in power devices. It has to be noted that the sun activity enhances the shielding effect of the Van-Allen belt. With a more intensive Van-Allen belt, less high-energy charged particles pass and collide with molecules of the atmosphere, finally less neutrons reach the earth surface. Figure 12.42 shows the correlation between sun activity and relative neutron flux on the earth´s surface.

The sun activity shows a somehow regular repetition in about 11 years, and after some time of higher activity an enforced belt becomes visible. The sun, in this respect, acts more protective than destructive. The shielding of the primary particles due to earth's magnetic field is more effective at the equator, where the flux of secondary particles at sea level amounts to $^{1}/_{3}$ compared to latitude of 45°, and the shielding effect is less at the poles, where the intensity is 3 times higher [IEC10].

For high altitude aircrafts and space applications, this is more complex. Radiation bursts of the sun were found to increase the particle flux in an altitude of 12 km up to a factor of 300, and can last several hours. Such events are estimated to happen 7 times in 67 years. [IEC10].

The high energy particles come from deep space. Figure 12.43 shows the flux density of primary cosmic radiation as a function of kinetic energy. Direct observations have been possible from satellites up to 10^{14} … 10^{15} eV. For higher energies, the data come from air-shower detectors on the ground, which observe the cascades of secondary particles in the atmosphere initiated by the high-energy

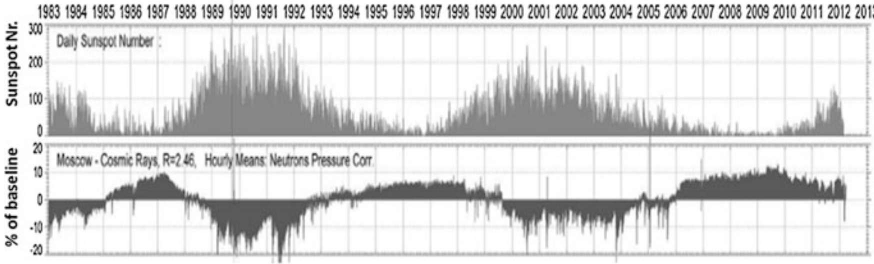


Fig. 12.42 Sun activity (daily sunspot no.) and relative neutron flux on the earth's surface. Figure from [Wil12], Figure D. Wilkinson, Wikimedia Commons license.

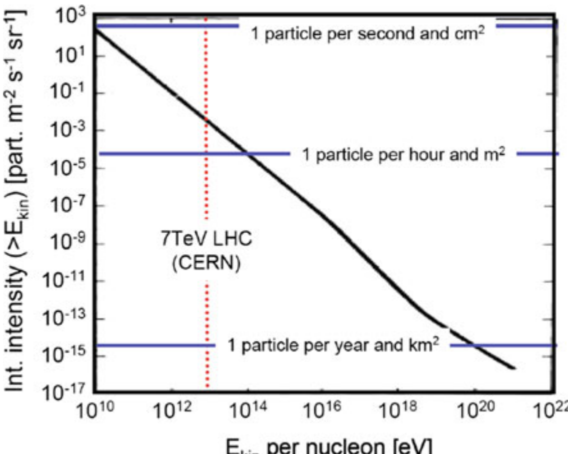


Fig. 12.43 Flux density of primary cosmic radiation as a function of kinetic energy. CERN, the largest accelerator on earth, can generate about 1×10^{13} eV. Figure from [Sti09] based on data from [Zom90]

primary particles [Gai13]. Primary particle energies above 10^{20} eV have been found. The highest energy reached by human technology at the CERN accelerator lies in the range of 10^{13} eV – far below the energy reached by cosmic particles.

For very high energy, cores of active galaxies are the only reasonable candidates as sources. Figure 12.44 shows the nearest active galaxy in our neighborhood, Centaurus A in a distance of 12 million light years.

Centraus A is a quite low-energy active galaxy compared to other active galaxies, for example Cygnus A in a distance of about 750 million light years. High energy particle impacts are monitored by the Pierre Auger observatory in Argentina which uses a large area of 3000 km^2 for detection of secondary particles and which

Fig. 12.44 Active galaxy Centaurus A. Left: optical image ESO 2.2 m telescope Silla Observatory Chile. Right: Chandra x-ray telexcope. X-ray image, showing one of two particle jets ejected from the core with relativistic velocities. Reprint of both Fig's according to Wikimedia Commons license

recalculates from these observations the path and energy of the initial particle. For some high-energy events, active galaxies are identified as origin [Blu10], but the majority of particles seem to originate from a random background.

Cosmic ray particles may be on their journey for several 10 or 100 million years before they collide with the earth atmosphere. An explanation with the simplified model of “black holes” in galactic cores has not been successful so far [Joo06], the real process creating such high energies has not been understood up to now.

12.8.3 Cosmic Ray Failure Patterns

Cosmic ray failures in power devices are single event burnout (SEB) or single event gate rupture (SEGR) in MOSFETs [IEC13] and other field controlled devices. SEGR means the gate oxide breakdown, leading first to gate leakage and finally to gate rupture. In power MOSFETs irradiated with protons, the gate ruptures are explained as caused by charge deposition from recoil ions [Tit98]. Griffoni et al. [Gri12] conducted tests with silicon IGBTs, silicon superjunction and SiC MOSFETs using neutrons, they observed SEGR only in the superjunction MOSFETs. SEGR should also be considered for power devices, however it is assumed to be a major effect for space applications only, while in all other applications SEB is the main effect. Thus, we will focus in the following on SEB effects.

Devices failed by SEB in laboratory tests exhibit a pinhole-size molten channel between the cathode and the anode side randomly distributed over the chip area. Failure pictures from device failures caused by cosmic ray in laboratory tests are shown in Fig. 12.45. On the left hand side in Fig. 12.45, a pinhole can be seen. On the right hand side, bubbles in the metallization can be seen, and below the metallization a pinhole is hidden. The failure pictures clearly show an effect occurring in a very narrow region.

Figure 12.46 shows the cosmic ray failure picture of a 3.3 kV IGBT die in a laboratory test. Again a pinhole is found, with a size in the order of the cell pitch of the IGBT die.

Figure 12.47 shows a cross section through the failure position of another IGBT die. The destroyed area is reaching through the whole drift layer of the device. The molten silicon in the *n*-drift region rapidly solidified and a crack was generated during the rapid cooling process.

Several tests of different device designs were carried out; for acceleration of the failure rate tests stations at high altitudes were arranged (Zugspitze 2964 m, Jungfraujoch 3580 m). The terrestrial cosmic particle flux increases with an altitude above sea level up to a height above 11 km [All84]. The acceleration factor with respect to sea level failure rates increases from 10 in 3000 m to approx. 45 in 5000 m [Kai05]. In accelerated experiments, tests with particle accelerators are performed. Devices at high reverse voltages are irradiated with high energy neutrons, protons and other ions accelerated with high energy. The irradiation with ions, especially the types of ions and energies which are generated at an inelastic

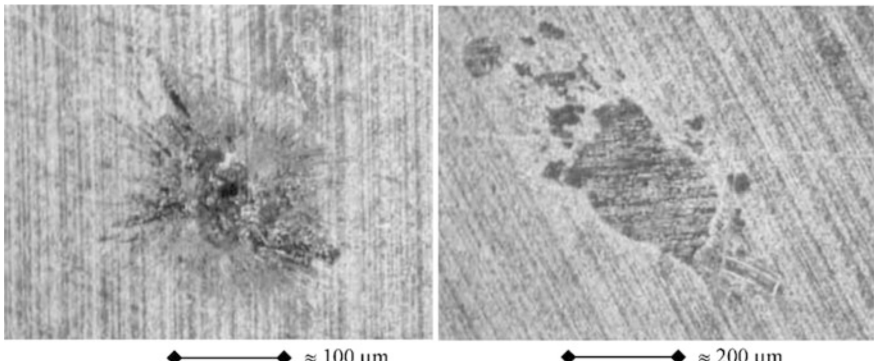


Fig. 12.45 Failure pictures after a cosmic ray destruction in a laboratory test of 4.5 kV diodes with a diameter of <50 µm. The photos are taken from the cathode side. Left hand side: small pinhole. Right hand side: Molten area in the metallization with bubbles. Pictures from Jean-Francois Serviere, Alstom

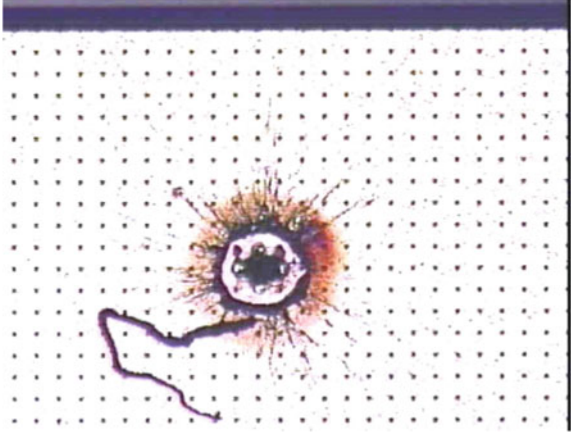


Fig. 12.46 Cosmic ray failure of a 3.3 kV IGBT in the cell area (laboratory test). Cell pitch 15 µm. Picture from G. Sölkner, Infineon

neutron-silicon-collision, is suited to study the detailed failure mechanism. A great amount of research work has been carried out in this field, see e.g. [Soe00]. Tests are executed today with "white neutron sources" having a similar energy spectrum than the cosmic neutrons or with monochromatic proton beams. It was shown that high-energy protons (>150 MeV) lead to similar results as white neutrons [IEC13].

It must be emphasized, that the failure images presented in this section are all resulting from accelerated or non-accelerated laboratory tests. In laboratory tests, only very limited energies are involved, so that the initial destruction image can be evaluated. In a real power electronic applications, the substantially higher energies in DC link capacitors involved will cause considerably more damage. It is in

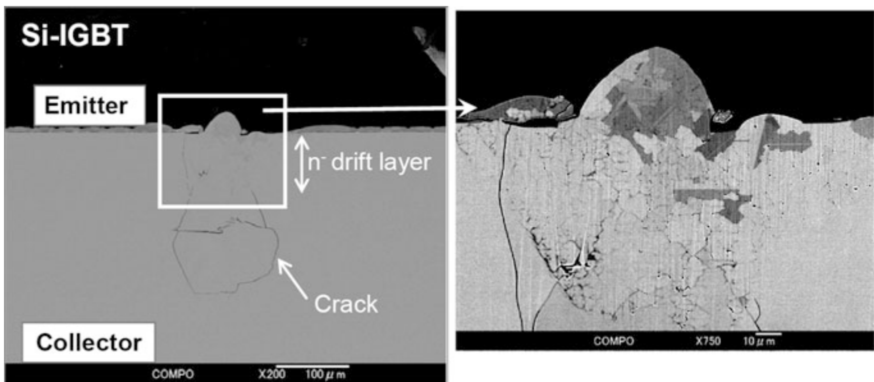


Fig. 12.47 Cross section at the failure position of a cosmic ray failure of an IGBT in a laboratory test. Figure from T.Shoji, Toyota [Sho11] © IEEJ 2011

general not possible to distinguish a cosmic ray failure from any other device failure resulting in a short circuit mode of the device.

12.8.4 Basic Failure Mechanism Model

A collision of a particle with the nucleus of a lattice atom does not destroy a power semiconductor by the SEB effect. Precondition is the presence of a high electric field and impact ionization.

Let us consider a device in the blocking state with a trapezoidal electric field as drawn in Fig. 12.48a. A neutron travelling through the space charge of a blocking semiconductor device is captured by a ${}^{28}_{14}$Si nucleus in a process of inelastic scattering, see Fig. 12.48b. The created ${}^{29}_{14}$Si nucleus is highly excited and decays fast in several light ions. The light ions have a high kinetic energy and create electron-hole pairs locally forming a dense plasma of charge carriers. All this happens on a time scale of picoseconds.

In the plasma is a high density of electrons and holes, and due to its state of almost neutrality no high fields can occur. The electric field in the plasma region is low. At the boarders between the plasma and the space charge, however, an extreme high density of charge emerges, leading to very high and steep field peaks. In silicon, it can even reach up to 1 MV/cm [Wei15]. If the electric field exceeds a certain threshold value, impact ionization creates more carriers than carriers which flow out of the plasma region by diffusion. One field peak moves fast to the anode, the other to the cathode. A so-called “streamer” is formed in analogy to a discharge in a gas. The device is flooded locally with free carriers within some hundred picoseconds, hence, a local current tube occurs. Basic failure models assumed that the very high local current density destroys the semiconductor device [Kab94, Kai04].

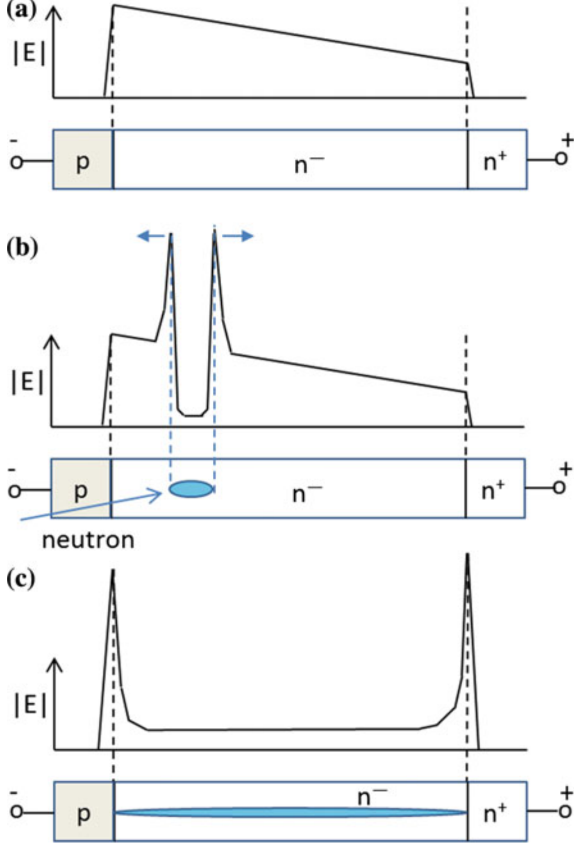


Fig. 12.48 Destruction of a power device (diode) by a neutron. **a** Device in the blocking state, **b** inelastic scattering of the incoming particle, formation of an electron-hole plasma, high field peaks at the boarders of the plasma, **c** the field peaks have run through the device, forming a streamer. At the pn^- and nn^+ junction, field maxima occur. Figure inspired by Kaindl [Kai05] and Weiss [Wei15]

If the field peaks arrive at the pn- and nn^+ junction, then an Egawa-type field (see Chap. 13) with two peaks at the pn- and nn^+ junction is given. This phenomenon is displayed in Fig. 12.48c for the case of abrupt junctions. We will come back to this process in the discussion of extended models, Sect. 12.8.6.

12.8.5 Basic Design Rules

The basic design rule considers the field strength at the position of the impact. The lower the field strength, the lower is the probability of failures.

In Fig. 12.49 PT and NPT designs are sketched for the same rated voltage. The maximal electric field is much lower for the PT-design. The shape of the electric field is drawn for two simplified diode structures with the same thickness and for the same applied blocking voltage. The area below the line $-\mathsf{E}(x)$, which corresponds to the reverse voltage, is the same for both devices. However, the value E_0 is

much lower for the PT design. At the same voltage, impact ionization is still negligible for the device with PT design. For the occurrence of impact ionization in the PT diode, the applied voltage must be increased. Then the line of $-E(x)$ in Fig. 12.49b is shifted upwards, until E_0 becomes close to a value as in Fig. 12.49a. Then impact ionization emerges in the diode with PT-dimensioning as well, but at a much higher voltage.

Experimental and simulation results on this effect are shown in Fig. 12.50, taken from [Kai04]. Diodes rated at 3.3 kV have been irradiated with ^{12}C ions, and the reverse voltage applied during irradiation was increased. Two diode designs were compared having a triangular electric field shape (NPT, see Fig. 12.49a) and another diode with a trapezoidal shape of the electric field (PT in Fig. 12.49b), which is indicated by FS (field stop). At a small voltage the charge generated by a single ^{12}C ion is small for all samples, above a defined threshold voltage a strong charge carrier multiplication arises and the created charge increases more than three decades. For the FS diode, which has a PT design, this threshold voltage is more than 700 V higher when compared to the NPT diode. A device design for increased cosmic ray stability must keep the maximum electric field E_0 at the maximum DC voltage V_{bat} required by application as low as possible. Thus, the base doping N_D

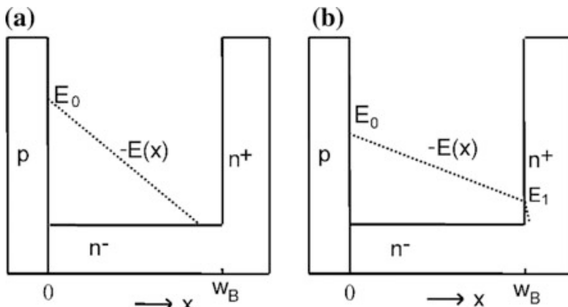


Fig. 12.49 Schematic drawing of the field shape for an NPT design (**a**) and PT-design (**b**) at the same thickness and same applied blocking voltage

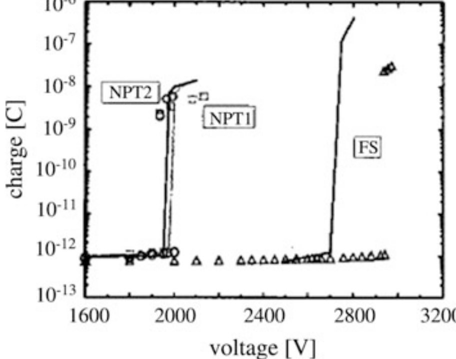


Fig. 12.50 Charge created by irradiation of a 3.3 kV diode with single ^{12}C ions of a kinetic energy of 1.7 MeV as a function of the reverse DC voltage applied during irradiation. Simulation (straight line) and experiment (symbols) for a triangular field shape NPT1 (white square), NPT2 (white circle) and a trapezoidal field shape FS (white triangle). Figure from [Kai04] © 2004 IEEE

must be lowered, so that a PT dimensioning can be achieved. Additionally, E_0 can be reduced by a wider drift region w_B. However, the latter measure is detrimental with regard to conduction and switching losses.

Furthermore, some quantitative models for this design dependency were published, the first was introduced by Zeller [Zel95]. The failure rate is given in the unit 'Failure In Time' (FIT); 1 FIT is one failure in 10^9 h. For a power electronic module in traction application for example, a failure rate below 100 FIT is required. Considering that a typical power module for such an application comprises 24 IGBT chips and 12 freewheeling diode chips, the failure rate for a single device must be more than one order of magnitude lower.

The Zeller model for the failure rate r reads as

$$r = a_1 \cdot A \cdot S^2 \cdot e^{-\frac{b_1}{S}} \quad \text{with} \quad S = 0.2786 \cdot \frac{V}{t} + 0.8972 \cdot \frac{t}{\rho} \tag{12.11}$$

With the device area A in cm^2, the specific resistivity of base doping ρ in Ω cm, the field strength factor S, the applied voltage $V = V_{DC} = V_{bat}$ in V and the thickness of the base layer t in µm. The equation has the advantage that it uses values which are known to the device designer, like thickness and resistivity. It can be shown that the results correspond to the field strength at a cosmic ray impact, and can also be written as

$$r = a \cdot e^{-\frac{b}{E_C}} \tag{12.12}$$

The model was quite in agreement with experimental results for high voltage devices >2 kV. For devices with lower voltage rating, Pfirsch und Sölkner [Pfi10] showed that for IGBTs and freewheeling diodes the failure rate is overestimated. They extended it by a correction factor $f(V_{bat})$ for V_{bat} below 2 kV

$$r = f(V_{bat}) \cdot a \cdot e^{-\frac{b}{E_C}} \tag{12.13}$$

The factor $f(V_{bat})$ is displayed as given in [Pfi10] in Fig. 12.51. For V_{bat} > 2 kV it approaches 1, for V_{bat} < 500 V the failure rate r becomes insignificant.

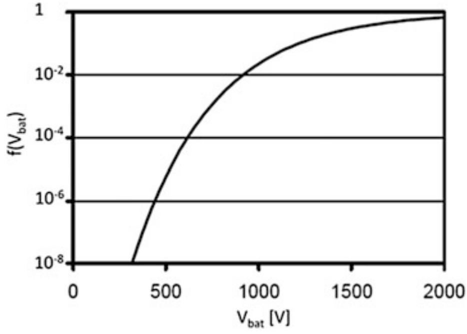


Fig. 12.51 Correction factor $f(V_{bat})$ for cosmic ray failure rate. Figure adapted from [Pfi10]

Moreover, a model by Kaminski [Kam04] was published, which is based on a fit of experimental results from IGBT power modules of one manufacturer. These equations are just the best adaption to the experiment. For T = 25 °C and sea level altitude, the results are given by the equation

$$r = C_3 \cdot a \cdot e^{\frac{C_2}{C_1 - V_{bat}}} \tag{12.14}$$

valid for $V_{bat} > C_1$.

Even though the models of Zeller, Pfirsch and Kaminski are designed for devices from different manufacturers, a quantitative comparison is given in Fig. 12.52 for a 1700 V device with an area of 0.44 cm^2.

The model of Pfirsch predicts a lower failure rate compared to that of Zeller. For a wide range, the model of Kaminski predicts an even higher failure rate than the Zeller model. The shape of the function, however, is different for Kaminski: It has a pole at the voltage C_1 in Eq. (12.14). Below C_1 the model is not valid and the failure rate is zero.

In addition, the model of Kaminski contains the dependency on the temperature and the altitude above sea level. The full equation is

$$r = C_3 \cdot a \cdot e^{\frac{C_2}{C_1 - V_{bat}}} \cdot e^{\frac{25 - T_{vj}}{47.6}} \cdot e^{\frac{1 - \left(1 - \frac{h}{44,300}\right)^{5.26}}{0.143}} \tag{12.15}$$

with the temperature T_{vj} in °C and the elevation above sea level h in m. Terms 2 and 3 of this equation are very useful to transfer results from one condition to other conditions and they are, if no further details are known, also used for devices of other manufacturers. The failure rate is highest at low temperatures and it decreases with temperature due to the decreasing avalanche ionization rates at increased temperature, see Chap. 2. The dependency on altitude is due to the increasing neutron flux density with height and is in reasonable agreement with the data in [IEC10] for a height up to 12,000 m. For altitudes above 18 km, the neutron flux is

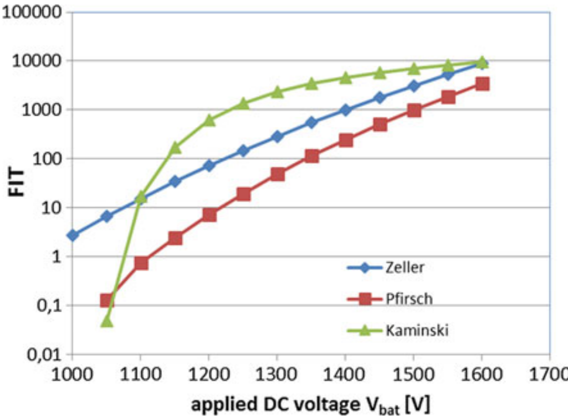


Fig. 12.52 Prediction of different models for a 1700 V device, area 0.44 cm^2

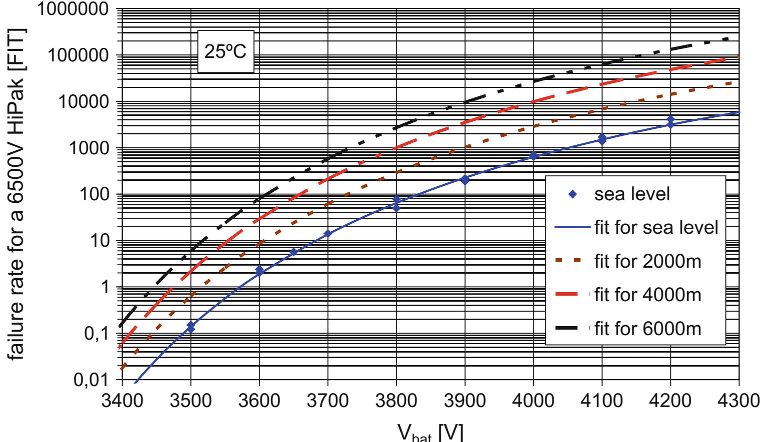


Fig. 12.53 Cosmic ray failure rate at $T = 25$ °C for the 6.5 kV IGBT module 5SNA0600G650100 from ABB. Figure from [Kam04]

decreasing again. Figure 12.53 shows an example for the dependency on height above sea level for a 6.5 kV IGBT module from [Kam04].

Following these rules and data, the probability of cosmic ray induced failures can be estimated for specific combination of application and device design and measures for a reduction of failure rates can deduced if necessary. Also a noticeable effect on the cosmic radiation induced failure rate from voltage peaks at switching is found [Hae12].

However, coming back to Fig. 12.52, the state of knowledge is unsatisfactory. Referring to the same area, the difference in failure rates resulting from different models amounts to almost two decades. Hence, the uncertainty is very large. Experimental data are necessary to reduce this uncertainty.

Furthermore, some of the models are for modules containing IGBTs and free-wheeling diodes, and these models do not distinguish between IGBT and diode. The device with higher failure rate will dominate the total failure rate, in one case it can be the IGBT, in another case it may be the freewheeling diode.

For high voltage devices, the dimensioning rules for cosmic ray stability are in contradiction with other rules for optimizing the device characteristics; e.g. for diodes it contradicts to the requirement for soft recovery behavior, which is hard to achieve if a strong PT dimensioning is used. A trade-off between different requirements must be made. To meet said demands, most of today's high-voltage devices use designs with a layer thickness w_B much higher than necessary for the required blocking capability. However, this leads to higher losses in the forward conduction mode and/or increased turn-off losses.

In a MOSFET, the failure mechanism is explained to be finally the activation of the parasitic bipolar transistor and second breakdown of the bipolar transistor [Was86]. In diodes, no parasitic component of any transistor type is present. Even a strong local avalanche breakdown should be stable, see Sect. 13.4. Field

redistribution effects in analogy to dynamic avalanche of the third degree are considered in deep submicron CMOS devices for improved ruggedness against radiation-caused single event pulses [DaG07].

In general no experimental data are publicly available for most power device manufacturers. Large customers received some data under the condition of non-disclosure agreements which required to keep the data confidential. Thus, no open scientific comparison was possible. However, new groups have started to work on cosmic ray stability. Si devices with a much higher onset voltage than predicted by the models discussed before were found. Meanwhile, some publications present not only experimental data but also extended explanations giving more insight.

12.8.6 Extended Model Considering the nn$^+$ Junction

Schulze and Lutz [Scu06] claimed the failure mechanism in diodes to be the occurrence of an Egawa-type electric field with 2 field peaks, similar to the third-degree dynamic avalanche. Egawa-type fields exhibit two field peaks, one at the pn-, one at the nn$^+$ junction. In Fig. 12.48c, the electric field in a streamer with high current density is shown. In this case, both pn- and nn$^+$ junction are abrupt. For this Egawa-type field, impact ionization occurs at both junctions. The figure has clear similarities to a dynamic avalanche of the third degree which will be discussed in detail in Sect. 13.4.2, compare Figs. 13.16 and 13.17 at the position of the filament. Figure 12.54a adapted from [Scu06] is very similar to Fig. 12.48c. Figure 12.54 is the key figure for understanding the extended model.

An electric field in the streamer as in Fig. 12.54b will be obtained, if the junction between low doped n$^-$ layer and n$^+$ layer shows a very slow increase of the doping, resp. if its gradient dN_D/dx is low. Since in a streamer an electron density of about 10^{17} cm^{-3} or more is expected, the dN_D/dx must be low especially in this range. Figure 12.55 shows a doping profile of the cathode layer which leads to an electric field in a streamer as in Fig. 12.54b.

The doping profile in Fig. 12.55 is taken from Fig. 7.3, it was used as collector layer of a bipolar transistor. The reason for this type of profile was to increase the stability of the bipolar transistor against second breakdown. The same type of profile hinders the formation of Egawa-type fields at dynamic avalanche, as will be discussed in Sect. 13.4. It also provides an increased cosmic ray stability.

According to Weiss [Wei15], the time in which the streamer formation happens is very short. The field peaks move with a velocity that is five times the saturation velocity. It takes only 200 ps, a time too short for a significant temperature increase that would lead to destruction. After arrival at the pn- and nn$^+$ junction, the streamer shortens both sides of the diode and is, at the first moment, combined with an Egawa-type avalanche at both sides. For an abrupt nn$^+$ junction, the cathode-side injection delivers the main part of the generated current. In the streamer, carriers diffuse laterally lowering the carrier density. It takes up to 20 ns until the channel,

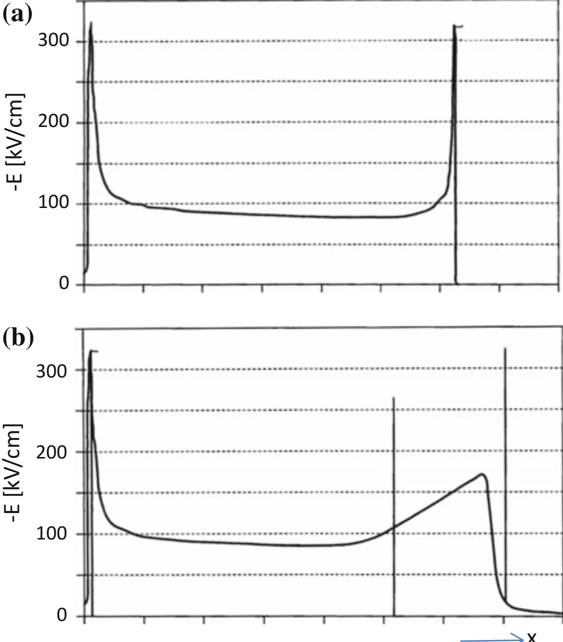


Fig. 12.54 Electric field at high current density in a streamer **a** abrupt pn$^-$ and nn$^+$ junction **b** abrupt pn junction, nn$^+$ junction with low gradient *dN*/*dx*. Fig. from [Scu06]

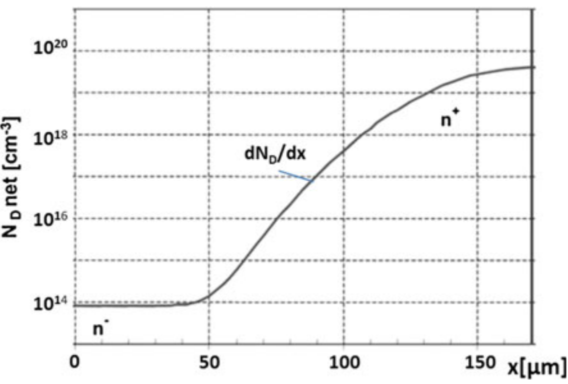


Fig. 12.55 Doping profile of the n$^+$ layer which leads to a field as in Fig. 12.54b in case of a streamer

which shortens anode and cathode, vanishes. Within this time, strong heating-up takes place, destroying the device. For the investigated diodes with an abrupt nn$^+$ junction, the highest temperature occurs at the nn$^+$ junction. However, for other device structures the temperature maximum can appear at the pn junction or inside the streamer close to the pn junction [Wei15].

In conclusion, the basic model considering the field strength at the neutron impact is not complete, it must be extended. Occurrence of impact ionization at the pn junction and formation of a streamer are necessary, but not sufficient conditions

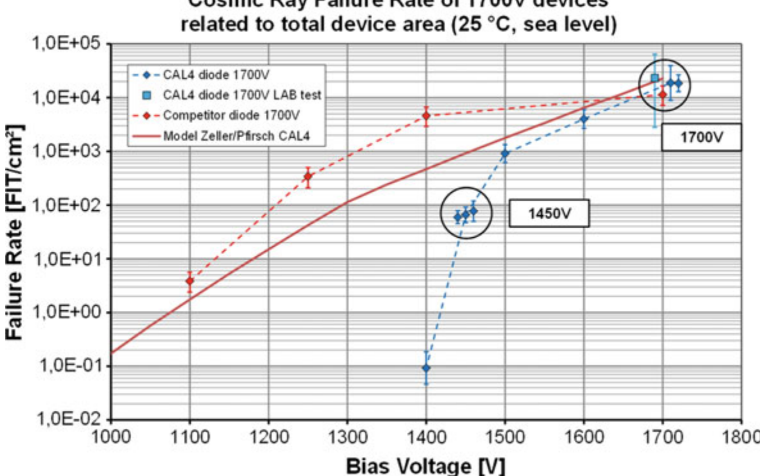


Fig. 12.56 Cosmic ray failure rate at $T = 25$ °C for a 1700 V CAL diode compared to model (Zeller with correction by Pfirsch) and to a competitor diode. Figure from [Scn15b]

for cosmic ray failures in diodes. The streamer is destructive if additional impact ionization at the nn^+ junction occurs, this is a sufficient condition that leads to failure. If a second avalanche in diodes or another amplifying mechanism in other devices can be avoided, a higher stability against cosmic ray failures is given.

Meanwhile, the extended model has been proven. For the 1700 V CAL diode [Lut94] it was shown that the threshold for cosmic ray destruction lies in the range of 1400 V [Scn15b]. Between 1000 and 1400 V, the failure rate is negligible or far below the prediction of the Pfirsch/Zeller model, and the detected threshold voltage is far above the threshold C1 of 983 V in Eq. (12.14) given for a 1700 V device in [Kam04]. The results for 1700 V are summarized in Fig. 12.56.

For the 1200 V CAL diode, a threshold in the range of 1200 V was found, and for the 600 V CAL diode no failures occurred even at 750 V.

Similar results have been published before by Shoji et al. [Sho13]. Failure rates of IGBTs and diodes are shown in Fig. 12.57a. The results indicate a threshold voltage for every device. Figure 12.57b shows the cosmic ray failure threshold voltage depending on the device base layer thickness for IGBTs with three different designs and one diode.

It is remarkable that a diode with 70 µm base width – which is typically used for a 600 V … 700 V diode – is better than an IGBT with approx. 160 µm wide n^- layer, which is typical for a 1700 V device.

Figure 12.58 shows the electric field for the diode from Fig. 12.57. For the onset of avalanche or moderate current density the field marked as line (1) is obtained. For an increased current density, the field has the shape of line (2). In a streamer, the device simulation shows line (3) for the electric field. The streamer front consists of electrons when it arrives at the cathode side. The higher the density of electrons, the

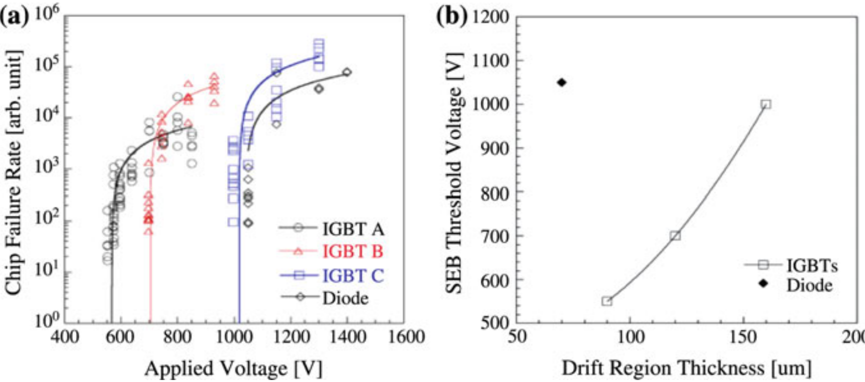


Fig. 12.57 Failure rate of IGBTs and diodes. **(a)** Failures, indicating a threshold voltage, **(b)** cosmic ray failure (SEB) threshold voltage depending on device base layer thickness, for IGBTs and one diode. Figure from [Sho13] © 2013 The Japan Society of Applied Physics

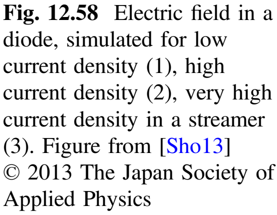
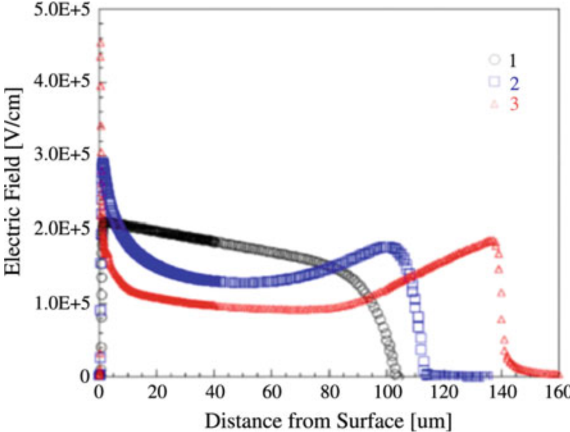


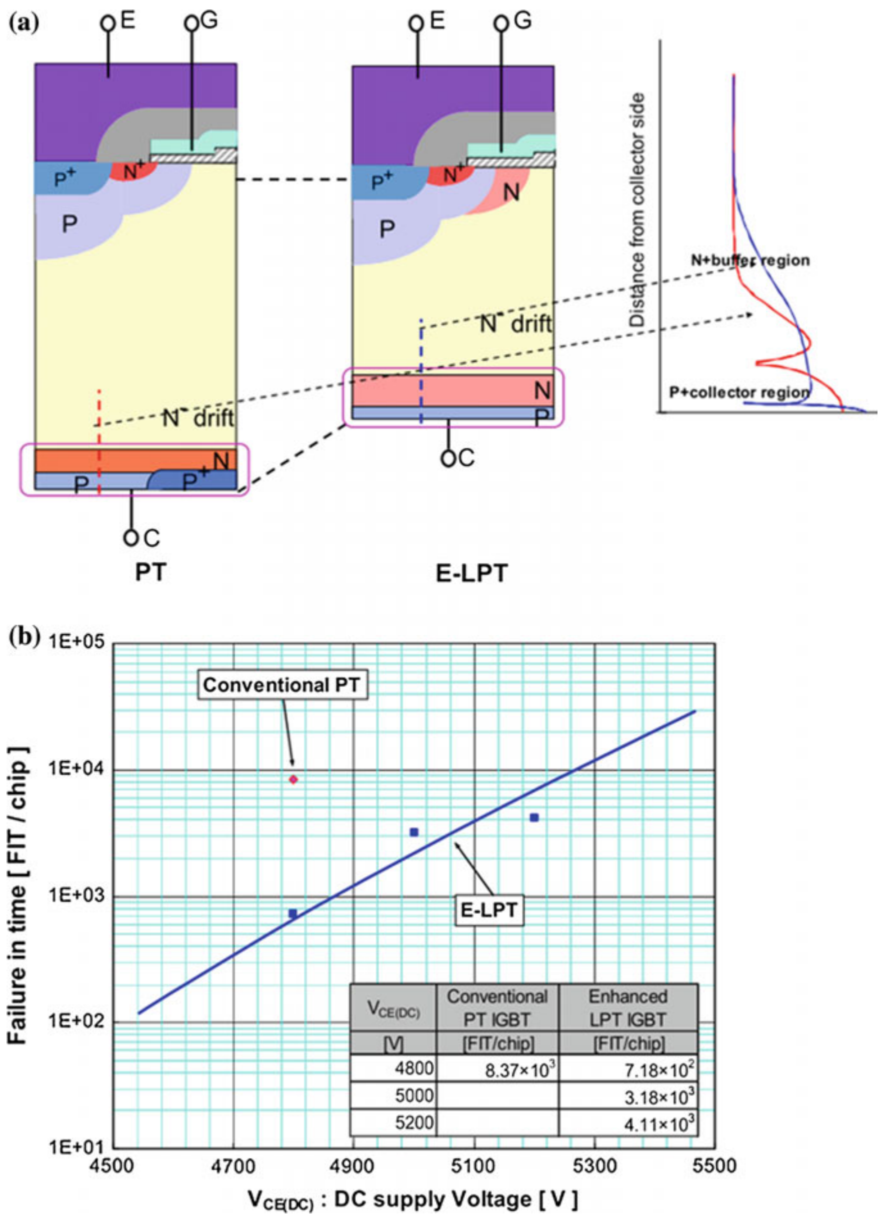
Fig. 12.58 Electric field in a diode, simulated for low current density (1), high current density (2), very high current density in a streamer (3). Figure from [Sho13] © 2013 The Japan Society of Applied Physics



more the doping of the n^+-layer is compensated, and the deeper the electric field penetrates into the n^+ layer. Figure 12.58 is very similar to Fig. 12.54b.

Both designs with very high ruggedness against cosmic ray, the CAL-diode [Scn15b] as well as the diode from [Sho13], contain similar n^-n^+ junctions with low doping gradients dN_D/dx. At high density of arriving electrons, they are partially compensated by the increasing density of positively charged donors, and the field can expand further into this compensation layer.

In summary, it can be stated that the failure mechanism in diodes with abrupt nn^+ junction is due to the occurrence of an Egawa-type electric field with two field peaks, similar to the dynamic avalanche of the third degree, as claimed in [Scu06]. This is experimentally confirmed by Scheuermann 2015 [Scn15b] and Shoi 2013 [Sho13] and supported with device simulation in [Wei15]. Designs with slowly



$V_{CE(DC)}$ [V]	Conventional PT IGBT [FIT/chip]	Enhanced LPT IGBT [FIT/chip]
4800	8.37×10^3	7.18×10^2
5000		3.18×10^3
5200		4.11×10^3

Fig. 12.59 Effect of the buffer doping on cosmic ray stability of a 6.5 kV IGBT. **a** Structure of the former (PT) and advanced generation (E-LPT), **b** failure rate for the former advanced generation. Figure from [Uem11] PCIM Europe 2011

increasing doping in the buffer can be very rugged. The threshold voltage can even be above the rated voltage.

The failure rate in modern IGBTs was found to be higher than that of cosmic ray stable diodes, yet much better than the competitor diode in Fig. 12.56 [Scn15b]. Figure 12.57 from [Sho13] also confirms a lower threshold voltage for IGBTs than for a well-designed diode. It is not possible to transfer the deep n^+ profile from Fig. 12.55 in the same form to an IGBT, since such a high buffer doping would lead to a very low current gain of the pnp-transistor. However, the buffer is of influence, results are shown in Fig. 12.59 [Uem11]. The recent design denoted as E-LPT has a more flat buffer layer with a lower gradient of doping dN_D/dx. Despite the new design having a lower base width, its failure rate is about one order of magnitude lower than for the conventional design denoted as PT.

The formation of an Egawa-field with a second field peak at the nn^+ junction to the buffer is supposed to be involved in the root cause for cosmic ray failures in IGBTs [Sho13]. The generated carriers activate the pnp-transistor. The onset of impact ionization at n^-n^+ interface can cause the parasitic transistor of IGBT to switch on; this phenomenon is known as destructive latching in IGBTs.

12.8.7 Further Design Aspects in Extended Models

Additionally to the impact of the nn^+ junction on the cosmic ray failure rate, the pn junction is of influence as well. Figure 12.60 compares two p layers [Wei15], where one layer is denoted as standard and the second is of lower depth and lower integral doping density G_n, compare Eq. (5.107). The shallow anode shows a reduced threshold for cosmic ray failures by approx. 100 V.

It has to be considered that in a streamer a carrier density in the range of 10^{17} cm^{-3} is to be expected. At the pn junction, holes of this density arrive and are capable to compensate the negative charged acceptors of the p layer. The electric field penetrates the p layer and may locally reach the semiconductor surface. Then a large number of carriers is injected by the metal layer. The effect is shown by device simulation in [Wei15].

It is reported there that the effect occurs suddenly for too shallow doping profiles, and above a certain limit other effects dominate the failure rate. There are clear similarities to the second-degree dynamic avalanche, where a too low doped anode layer can be compensated by the hole density in a current filament, compare Sect. 13.4.2. However, in a cosmic ray induced streamer even higher current densities can be expected than in case of dynamic avalanche.

With an additional very shallow and highly doped p^+ layer directly located at the metal-semiconductor interface (contact implantation), the electric field will no longer reach the semiconductor surface, and it is reported that diodes with shallow emitter no longer show a decreased threshold voltage for cosmic ray failures [Wei15]. A very similar result is published in [Mit15], where such an additional shallow p^+ layer was investigated and the higher doping demonstrated improved stability.

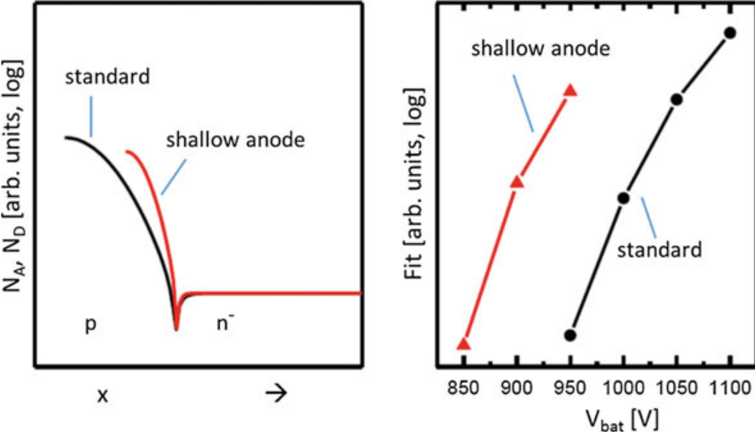


Fig. 12.60 Influence of anode design in diodes. Left: Comparison of standard and very shallow doped anode. Right: Failure rate measured with a proton beam. Figure adapted from [Wei15]

In a MOSFET, the failure mechanism is explained to be the activation of the parasitic bipolar transistor and second breakdown of the bipolar transistor [Was86]. For superjunction MOSFETs, it is reported that the distance between regions compensated with p columns and the n^+ substrate is of influence, compare Fig. 9.10. The higher this distance, the more cosmic ray stability was achieved [Wei15]. However, also the cell structure was found to be of influence. Regarding cosmic ray stability, superjunction devices are challenging.

A detailed final model of the failure mechanism related to ‘Single Event Burnout’ in IGBTs and MOSFETS is still a subject of research.

12.8.8 Cosmic Ray Stability of SiC Devices

SiC is sometimes considered as “radiation hard”, however, there was never any experimental evidence for a statement regarding higher stability against cosmic ray failure or ’Single Event Burnout’ (SEB). In fact, for the same voltage in SiC, the width of the space charge is 10 times lower, see Fig. 6.8. The volume where high field strength is given is also 10 times smaller for the same device area. On the other hand, the electric field strength at the same condition is 10 times higher. Which of these factors provide the higher influence?

First published experimental results were contradictory. For 600 V SiC Schottky diodes, the failure rate was found to be higher than for Si pin diodes. For 1200 V SiC Schottky diodes and JFETs, the failure rate was significantly below a Si pin diode. Therefore, early results were difficult to evaluate. Meanwhile, we know that there are pin diodes with failure rates varying over several orders of magnitude, see last paragraph. This might explain the conflicting results. For SiC, only since recently we

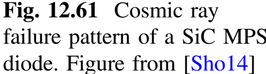
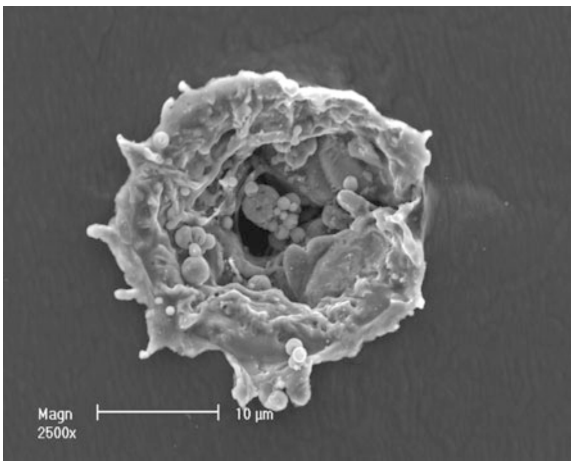


Fig. 12.61 Cosmic ray failure pattern of a SiC MPS diode. Figure from [Sho14]



are having results that allow a first evaluation. The defects in SiC resulting from cosmic ray failures look quite similar to that of Si, see Fig. 12.61 [Sho14].

Failure rates of different Si devices and one SiC MOSFET are compared in Fig. 12.62. The failure rate is normalized to the on-resistance, however, it is significant that there is a different threshold voltage for the different devices, especially some of the superjunction MOSFETS are sensitive. While for the 1200 V Si IGBT the threshold appears between 750 and 800 V, no failure is found for the SiC MOSFET up to 950 V.

So far, SiC seems to have an advantage. However, in an experimental investigation using the neutron beam in Uppsala with a spectrum comparable to the cosmic ray spectrum at sea level altitude, Consentino et al. [Con15] found the threshold voltage for a 1200 V SiC MOSFET in the range of 1020 V. In [Scn15b]

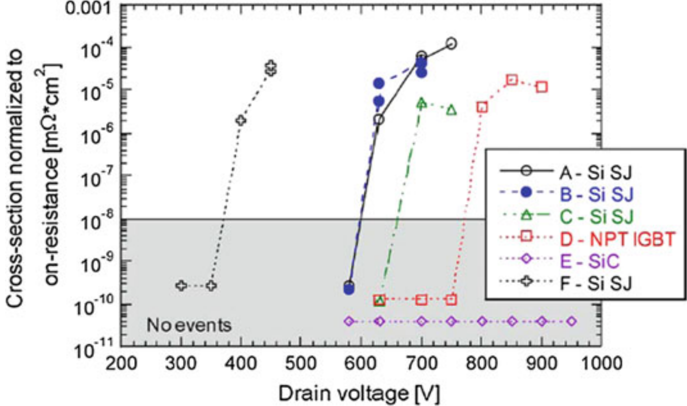


Fig. 12.62 Cosmic ray failure rates of four Si superjunction structures, Si IGBT and SiC MOSFET. The superjunction devices are rated 600 V (F), 900 V (C), 1000 V (B), 1100 V (A), Si IGBT and SiC MOSFET are rated 1200 V. Figure from [Gri12] © 2012 IEEE

the threshold voltage was found at 850 V for an Si IGBT of a more recent design (fieldstop) than the NPT IGBT in Fig. 12.60. So for a first quantitative comparison, the threshold voltage for cosmic ray failures can be assumed as 70% of V_{rated} for the Si IGBT and as 85% of V_{rated} for the SiC MOSFET (Fig. 12.63).

In simulations of cosmic ray failures in SiC MPS diodes by Shoji, similar effects like in Si diodes were found [Sho14]. A streamer shortens anode and cathode and the impact ionization generates higher charge at the cathode side. So we can conclude: The SEB current produced by impact ionization at the n^-n^+ interface is common to both Si and SiC power devices. SEB in power diodes thus corresponds to thermal destruction caused by an Egawa-type field (dynamic avalanche of the third degree, Sect. 13.4.2) In [Sho14], the same effect is denominated as “local second breakdown”.

For diodes, a comparison between Si pin-diodes and SiC diodes is given in [Feg16a]. The results are shown in Fig. 12.64. The Si-Diode-C has an abrupt nn^+ junction where the electric field in the streamer will be comparable to Fig. 12.54a, the Si-Diode-A (CAL) has a soft nn^+ junction with low gradient dN/dx and the electric field in the streamer is expected to be similar to Fig. 12.54b. The results for the two Si-diodes are very different, the SiC diode is more robust than the Si-Diode-C but less robust than the Si-Diode-A.

A detailed comparison for SiC MOSFETs and Si IGBTs is presented in [Feg16b]. The devices are compared not only for the rated voltage, but also for the measured breakdown voltage. The rated voltage compared to the measured breakdown voltage V_{BD} for the investigated 1200 V Si IGBTs was found to be 88–89%, for the 1200 V SiC MOSFET it is 73%. For 1700 V, the Si IGBT used 79% of the breakdown voltage as rated voltage, the SiC MOSFET used 64%. This shows that the high critical field strength of SiC is only partially used in the investigated SiC designs. The results analyzed in dependence of V_{DC}/V_{BD}, meaning normalized to applied DC voltage V_{DC} compared to the measured breakdown voltage V_{BD}, are as follows: For 1200 V devices there is a small advantage of the SiC-MOSFET (Fig. 12.65a), for 1700 V the failure rate becomes significant at similar V_{DC}/V_{BD} (Fig. 12.65b).

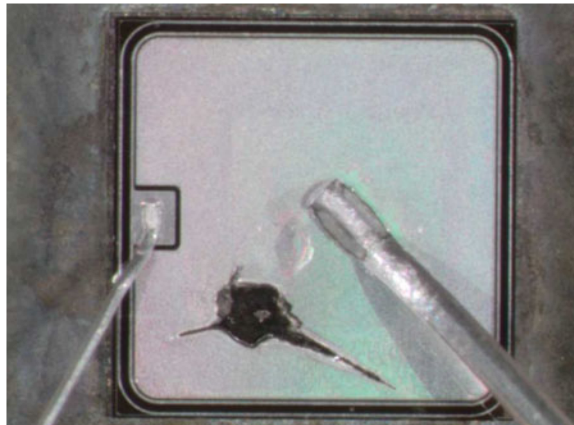


Fig. 12.63 Cosmic ray failure of a SiC MOSFET. Figure from [Con15] PCIM Europe 2015

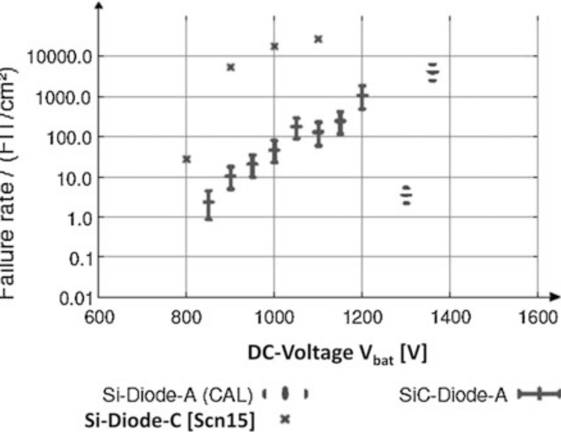


Fig. 12.64 Cosmic ray failure rate of two different Si diodes and one SiC diode. All diodes are rated 1200 V. Figure adapted from [Feg16a] PCIM Europe 2016

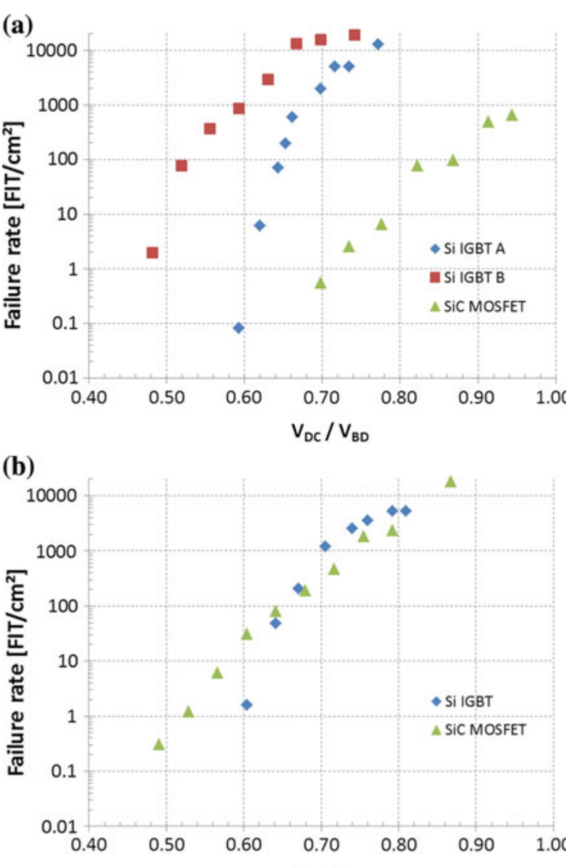


Fig. 12.65 Comparison of cosmic ray failure rates for Si IGBTs and SiC MOSFETs normalized to applied DC voltage V_{DC} compared to the measured breakdown voltage V_{BD}. **a** 1200 V rated devices, **b** 1700 V rated devices. Figures according to data from [Feg16b]

The physics of cosmic ray failures seems to be very similar for Si and SiC. However, with increased crystal quality in the future, for SiC the full potential of the material will also be exploited. More data for SiC and for Si are of high interest. For the same rated current there is a lower device area for SiC. Therefore, it still may remain a small advantage for SiC.

12.9 Statistical Evaluation of Reliability Test Results

The evaluation and interpretation of reliability test results can be considerably enhanced by statistical methods. Statistic software collections (e.g., Minitab® Statistical Software [MIN18]) provide powerful tools for the analysis of reliability test data or for the investigation of field failures. While statistical methods can be very helpful if applied correctly, they can also be misuse or misinterpreted by unexperienced engineers. It would exceed the frame of this book to provide fundamental knowledge on statistics, thus we will give some examples of statistical interpretations and we will also discuss some cases of incorrect application or misinterpretation of statistics, which are sometimes encountered in reliability investigations.

A good example for statistical analysis is the evaluation of end-of-life power cycling test results of power modules. If we only have a single result, we cannot say anything about the expected result of a second test performed under the same conditions. Often such a single test result is associated with a 50% failure probability. However, since we have no information on the statistical distribution of failures, this interpretation is not supported by statistics.

If several power cycling tests are conducted at the same test condition, much more information on the statistical character of failures can be extracted by statistical tools. The example in Table 12.6 shows the results of 6 power cycling tests which were conducted at identical test conditions. Power pulses with t_{on} = 1 s and t_{off} = 1 s were imposed by a DC load current of 244 A on the IGBTs of a 1200 V SKiM63 module bonded with an aspect ratio of ar = 0.31. The tests were conducted until the total failure of all wire bonds occurred and the number of cycles to reach this failure is indicated by N_{ex}.

Using a statistic software like Minitab allows to analyze the data obtained for N_{ex}. Minitab [MIN18] provides a distribution identification tool, which compares the probability that a given sample belongs to a specific statistical distribution by an Anderson-Darling-Test. For the values N_{ex}, this test confirms that distribution is most probably a Weibull distribution. The Weibull distribution is described by the probability density function

Table 12.6 Power cycling results for 6 IGBTs of a 1200 V SKiM63 module with temperature swing parameters, experimental results N_{ex} and corrected values N_{corr}

	T_{jmin} (°C)	T_{jmax} (°C)	ΔT_j (K)	T_{jm} (°C)	N_{ex}	N_{mod}^1	$f_{corr} = N_{ex}/N_{mod}$	$N_{corr} = N_{ref} \cdot f_{corr}$
1T	80	148	68	114	521,940	499,549	1.0448	459,829
1B	80	150	70	115	451,810	440,103	1.0266	451,810
2T	83	152	69	117.5	464,440	461,578	1.0062	442,832
2B	80	149	69	114.5	537,050	468,641	1.1460	504,346
3T	82	153	71	117.5	454,520	409,564	1.1098	488,411
3B	82	150	68	116	594,860	494,492	1.2030	529,432
Ref	80	150	70	115		440,103		

[1]SKiM63 lifetime model estimation with $ar = 0.31$ and $t_{on} = 1$ s

$$f(x, \alpha, \beta) = \frac{\alpha}{\beta^{\alpha}} \cdot x^{\alpha-1} \cdot \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right) \tag{12.16}$$

with the shape parameter α and the scale parameter β. The parameter x represents in this context the number of cycles to failure. The accumulated number of failures at a given number of cycles is obtained by the integration over the probability density function up to x which gives the cumulative distribution function of the Weibull distribution

$$F(x, \alpha, \beta) = 1 - \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right) \tag{12.17}$$

From the cumulative distribution function, the parameters of the Weibull distribution can be extracted as shown in Fig. 12.66. By plotting $\ln[-\ln(1 - F)]$ over $\ln(x/\beta)$ the Weibull distribution is represented by a straight line with the slope given by the shape parameter α. The two additional curves on both sides of estimated Weibull distribution indicate the confidence intervals for a 95% confidence level, i.e. the straight line for the real distribution can be located anywhere in the area between the curves with a confidence of 95%.

However, if we look at the temperature limits T_{jmin} and T_{jmax} of the swing in Table 12.6, we can see that the power cycling conditions are not exactly identical for all 6 tests. This observation, which is typical for power cycling tests, is caused by the active role of the device under test (DUT) in power cycling testing: The temperature swing is not only determined by external parameters like current and load pulse duration, but also by internal characteristics, i.e. the forward voltage drop and the thermal resistance of the DUT. These device parameters exhibit variations due to material tolerances and fluctuations in the production process, which results in small differences in the temperature swing. The parameters of the temperature swing must be determined in the stable phase of the test when all initial transient processes have saturated. These transient processes are the heating of the heat sinks and module parts, the redistribution of thermal interface materials between module and heat sink

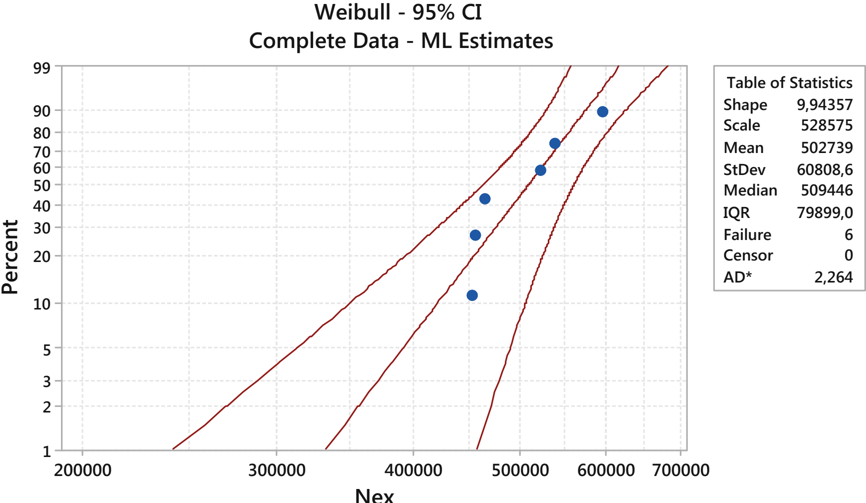


Fig. 12.66 Weibull distribution of N_{ex} in Table 12.6 with parameters and 95% confidence intervals extracted by maximum likelihood estimation with Minitab®

and the stabilization of the cooling liquid temperature controller. On the other hand, the initial temperature swing parameters must be determined before any degradation effects occur, so that no general rule can be applied. For the tests listed in Table 12.6, the initial values are determined by averaging T_{jmin} and T_{jmax} over several hundred cycle values after a few thousand cycles and rounding to integer values.

We can assume that the small variations of the temperature swing ΔT_j and medium temperature T_{jm} in Table 12.6 will have an impact on the test results. If a lifetime model is available that allows to calculate this impact on the number of cycles to failure, the test results can be corrected for this parameter variation. For the given example in Table 12.6, a lifetime model is available: The SKiM63 model introduced in Sect. 12.7.4. Applying the model parameters given in Table 12.4 to the temperature swing data, we can calculate the number of cycles to failure N_{mod} resulting from the lifetime model. The correction factor $f_{corr} = N_{ex}/N_{mod}$ describes the experimental results relative to the model prediction at the experimental temperature swing. If we would have tested at the reference temperature swing, we should have obtained for the cycles to failure $N_{corr} = N_{ref} \cdot f_{corr}$ if the lifetime model is correct. Thus, we can eliminate the variation in the temperature swing and can statistically analyze the corrected data as shown in Fig. 12.67.

To evaluate the correction process, we can compare the probability density functions of the two Weibull distributions in Fig. 12.68. The Weibull distribution for the uncorrected values exhibit a wider probability density function with a shape factor of 9.9 compared to 16.4 for the corrected values. This must be expected for a valid lifetime model, since the corrected values are more exact and the uncorrected values feature a broadening of the distribution by parameter variation.

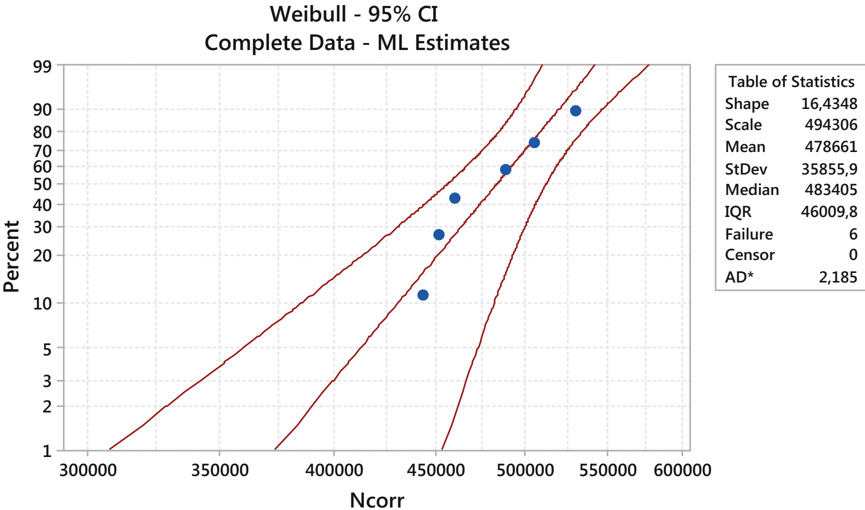


Fig. 12.67 Weibull distribution of corrected values N_{corr} in Table 12.6 extracted by maximum likelihood estimation with MiniTab®

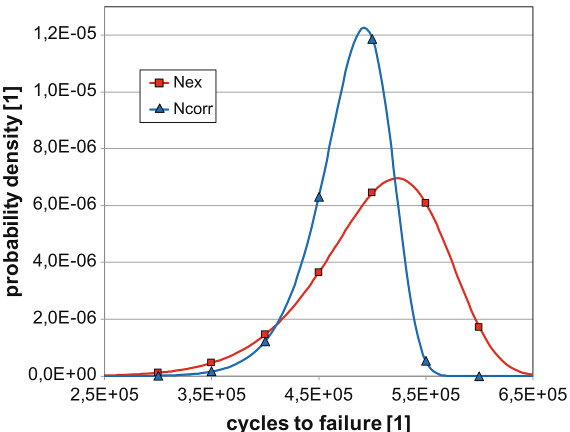


Fig. 12.68 Comparison of the probability density functions for the experimental result N_{ex} and the corrected results N_{corr} with the SKiM63 lifetime model

This correction process can be applied to perform a statistical evaluation of a complete lifetime model. Therefore, we could map all power cycling results with a wide variation of parameters to a single reference condition as we have done in Table 12.6. However, since the selected reference point is arbitrary, we can simply select the correction factor f_{corr} for the evaluation. This was done for the complete data set of more than 100 power cycling test results that was the basis for the parametrization of the SKiM63 lifetime model [Scn13]. Since the model coefficients (see Table 12.4) were determined by a least square fit, half of the data values

should be >1. Since the failure probability in Fig. 12.69 is shown as a function of the relative number of cycles to failure minus the threshold value (0.507) of a 3-parameter Weibull distribution, the 50% failure probability is reached at a value of $\sim$0.5 as expected (long arrow in Fig. 12.69). When the lifetime model predictions are multiplied with a margin factor of 0.8 (short arrow in Fig. 12.69), only 15% of parameter combinations would exhibit a failure before the estimated lifetime, while 85% would reveal a longer lifetime.

It should be emphasized, that the evaluation of a complete lifetime model has a higher significance than the evaluation of a single combination of test parameters in Figs. 12.66 and 12.67. The result for a single test condition provides no information on the statistical distribution at a different set of test parameters. The evaluation of a lifetime model gives a statistical information for the whole parameter space covered by the data set that was applied to determine the lifetime model coefficients.

The statistical analysis of test results delivers a correlation between the cycles to failure and the expected accumulated failure probability. However, an important question remains: Which failure probability should be defined for an estimated lifetime. Often, lifetime estimations for an accumulated failure probability of 1% or even lower are requested for a component by system designers. This can be understood as a quest for the lowest possible failure probability in application, however, this definition is problematic from a statistical point of view: The estimation of lifetime from a sample of tests run to end-of-life exhibits the highest accuracy for a 63% failure probability for a Weibull distribution, as can be seen in the probability plots shown before. At 63% probability, the difference between the

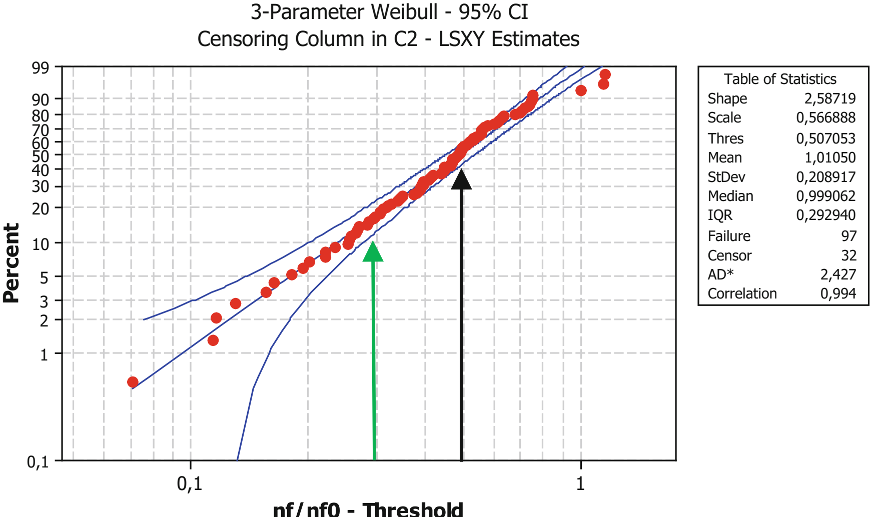


Fig. 12.69 Statistical evaluation of the SKiM63 lifetime model with a 3-parameter Weibull distribution. A margin factor of 0.8 (short arrow) indicates a failure probability of 15% [Scn13]

confidence intervals has its minimum. For higher or lower failure probabilities, the difference between the confidence intervals increases. Thus, the lower the defined failure probability, the higher the uncertainty of the medium distribution parameters indicated by the straight line in the probability plots. This results from the fact, that in a statistical sample a large fraction of elements will be in proximity of the maximum of the probability density function and only few elements will be found at the extremes. Therefore, failure probability of 10–15% seems to be a good compromise between low failure probability and uncertainty of prediction.

The statistical analysis is not restricted to power cycling tests; it can be applied to all failures observed in reliability testing. Figure 12.70 shows the probability plot of cosmic ray failures of 1700 V CAL diodes in an accelerated test with a mono-energetic 200 MeV proton beam of a flux density of 1.8×10^6 ps^{-1} cm^{-2} at a bias voltage of 1450 V. This test exhibits an acceleration factor of 6.48×10^8 related to the natural cosmic ray flux density at sea level (10 ph^{-1} cm^{-2}).

At the first glance, there seems to be no difference to the probability plots shown before. However, we can see in the data given in Fig. 12.70 that the shape factor is very close to 1.

To interpret the results, we have to look at the statistical definition of the failure rate. The reliability function or accumulated survival rate for a Weibull distribution is

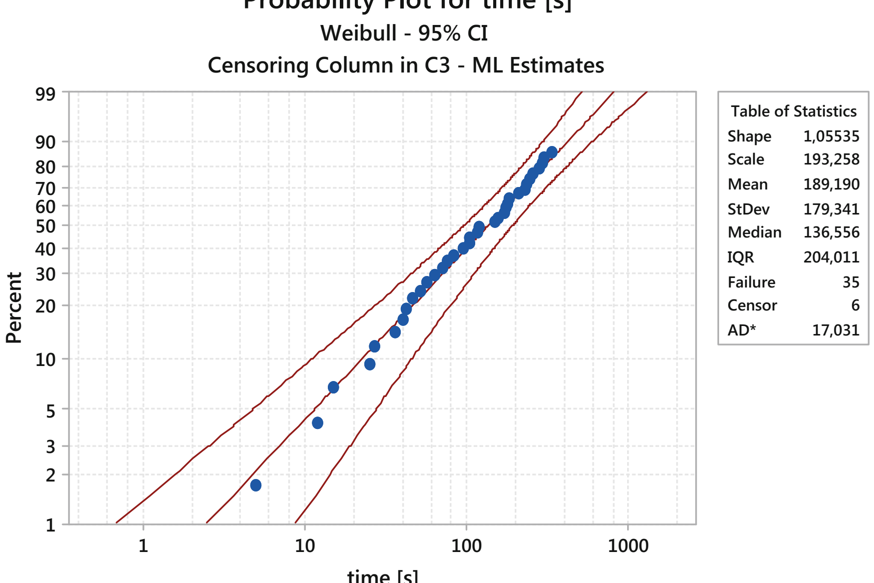


Fig. 12.70 Statistical analysis of cosmic ray failures of 1700 V CAL 4 diodes (area 44 mm^2) at 1450 V bias voltage in an accelerated test with a 200 MeV proton beam with a flux density of 1.8×10^6 $ps^{-1}cm^{-2}$ (middle value in Fig. 12.56) by MiniTab®

$$R(x, \alpha, \beta) = 1 - F(x, \alpha, \beta) = \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right) \tag{12.18}$$

The failure rate at any value x is given by the probability density function divided by the probability of survival at x, so that we obtain for a Weibull distribution

$$\lambda(x, \alpha, \beta) = \frac{f(x, \alpha, \beta)}{R(x, \alpha, \beta)} = \frac{\alpha}{\beta^{\alpha}} \cdot x^{\alpha-1} \tag{12.19}$$

From Eq. 12.19 we can see, that for a Weibull distribution with a shape factor $\alpha = 1$ the failure rate is a constant $\lambda = 1/\beta$ independent from the parameter x. Thus, we can determine from the scale factor $\beta = 193$ s the failure rate at sea level $\lambda = 29$ FIT for this 44 mm^2 diode. This is equivalent to a normalized failure rate of 65 FIT/cm^2 as shown in Fig. 12.56 for the middle value at 1450 V. The statistical analysis also allows to calculate defined confidence intervals for the measured data as described in standard textbooks on statistics, e.g. [Rau04].

Statistical methods also facilitate the calculation of the maximum resolution achievable with an accelerated test and the selection of sample size and test duration. An example is shown in Fig. 12.71 for accelerated cosmic ray tests. If at least 5 failures are required for statistical evaluation, the maximum acceleration factor given by the proton irradiation beam limits the detectable failure rate to $\sim$0.01 FIT for practical test conditions of 100 samples and 30 min beam time.

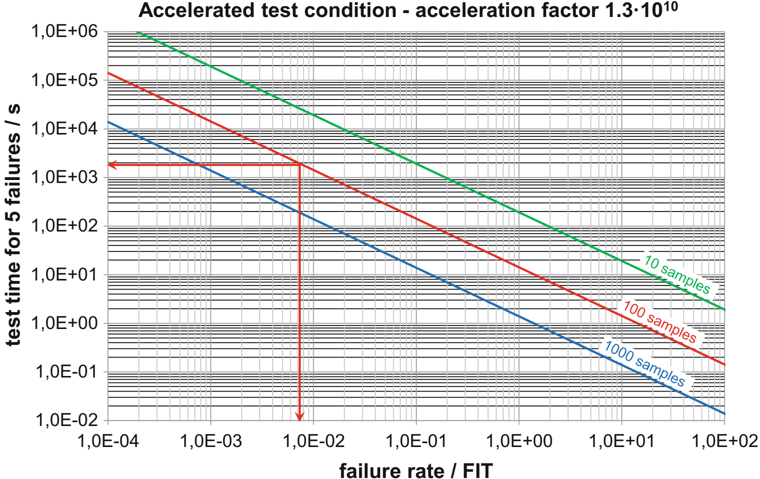


Fig. 12.71 Average test duration to obtain 5 failures at a specific failure rate for different sample sizes with an acceleration factor of 1.3×10^{10}. Arrows mark the practical detection limit for 100 samples and 30 min test time. Figure from [Scn16]

The analysis of the cosmic ray failures has shown, that they follow a Weibull distribution with a shape factor of 1, in contrast to the failures in power cycling tests which are characterized by $\lambda > 1$. The reason for this difference is, that they belong to different failure categories or failure types.

The stress during power cycling causes degradation and fatigue and is associated with accumulating damage in the system. Since it takes time to accumulate damage, a low failure rate at the beginning, which increases over time, can be expected. Therefore, power cycling failures belong to the failure type of 'end-of-life' failures in Fig. 12.72.

Cosmic ray failures are not increasing over time; the failure rates remain constant throughout the system life. The reason is, that an event originating from outside of the system is the root cause of failure. It does not matter, if a system has operated for 1 h or for 1 million hours; the probability of an impact of a cosmic ray particle remains constant. This failure type is referred to as random failure.

The third failure type, the 'early-life' failures, is characterized by a failure rate that decreases over time. This failure type is related to weak systems, which are selected from the population by sudden failures. When weak elements are removed, the remaining population becomes stronger and the failure rate decreases.

Since the bathtub curve in Fig. 12.72 is constructed by Weibull distributions, the reliability function or survival probability can be calculated from the distribution parameters and Eq. (12.18). The result is shown in Fig. 12.73. As discussed before, a survival probability of 99% is often requested for the limitation of useful life by degradation. This condition is reached in Fig. 12.73 after approx. 17.5 years of continuous operation. However, there could be considerably more failures after 17.5 years resulting from early-life and random failures. This must be kept in mind when targeting for so-called health monitoring concepts. The idea is to monitor degradation effects in power electronic components and issue a warning before the

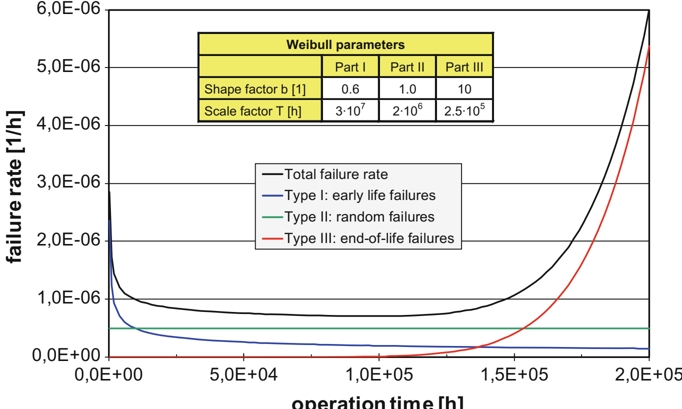


Fig. 12.72 Bathtub curve for a hypothetical system as a sum of failure rates for early-life-failures, random failures and end-of-life failures [Scn15a]

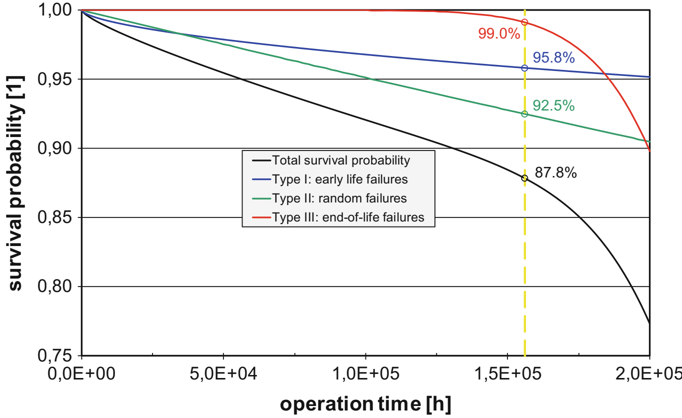


Fig. 12.73 Survival probability for the bathtub curve in Fig. 12.72 and contributions of the different failure types for a hypothetical system [Scn15a]

component fails. Since only end-of-life failures are associated with degradation, these concepts must be blind with respect to early-life and random failures. Therefore, health monitoring concepts will only make sense, when early-life and random failures are negligible.

In reliability assessments, MTTF (Mean Time To Failure) is often considered as an appropriate measure for the reliability of components or systems. However, MTTF is often misinterpreted. Let us assume a constant failure rate of 500 FIT or $0.5 \times 10^{-6}\ h^{-1}$. For a constant failure rate, the shape factor of the related Weibull distribution must be 1 and the MTTF is simply MTTF = β=1/λ = 2×10^6 h. Statements are sometimes heard like: “An MTTF of 2 million hours is a good value, because our inverter is designed for 2×10^5 h lifetime. Since the MTTF is 10 times higher than the lifetime, this inverter will not fail”. This interpretation assumes, that the MTTF-value can make predictions about failures of an individual inverter, which is not the case. MTTF is a characteristic of a statistical entity. An MTTF = 2×10^6 h predicts that for an entity comprising 1000 systems, one failure is expected every 2000 h of operation of all system.

MTTF is the expectancy value for a failure:

$$MTTF = \int_0^\infty R(x)dx \tag{12.20}$$

For a Weibull distribution, this expectancy value can be calculated with the gamma function $\Gamma(x)$:

$$MTTF = \beta \cdot \Gamma\left(1 + \frac{1}{\alpha}\right) \tag{12.21}$$

We can now calculate the MTTF values for the Weibull distributions in Fig. 12.72 with Eq. (12.21) and the MTTF value of the resulting bathtub curve using Eq. (12.20). We can also calculate the reliability function R(MTTF) using Eq. (12.18) for the 3 Weibull distributions. For the bathtub curve, the reliability function is determined by the product of the 3 contributing functions at the MTTF value of the bathtub curve. The results are collected in Table 12.7.

The results show, that at a time MTTF = 2×10^6 h, only ~37% of a population of inverters is expected to survive assuming a constant failure rate of 500 FIT. The survival probability at MTTF is with ~28% even worse for the early-life failures. This example demonstrates, that MTTF alone is not a good measure for the system reliability. Without additional information on the distribution of failures, the MTTF value gives no information about the survival probability of a statistical entity.

Finally, we will discuss the failure type of early-life failures. This failure category comprises failures with a decreasing failure rate. A way to reduce this failure rate is to establish appropriate tests that can screen out weak components and thus reduce the initial failure rate. However, it must be validated that the test conditions, which are often beyond the specification limits of the components, exhibit a decreasing failure rate as a function of test time.

An example of such a screening test is published in [Sie17]. SiC trench MOSFETs were subjected to a gate stress test at a temperature of 150 °C. The gate voltage applied was beyond the specification limit of 15 V. The observed failures as a function of test time are shown in Fig. 12.74. The results in this image show that the failure rate at each test level decreases over time.

Table 12.7 MTTF values and survival probabilities at MTTF for the Weibull distributions and the resulting bathtub curve in Fig. 12.72

Failure type	Shape α	Scale β [h]	MTTF [h]	R(MTTF) (%)
Early-life failures	0.6	3×10^7	4.5×10^7	27.9
Random failures	1	2×10^6	2×10^6	36.8
End-of-life failures	10	2.5×10^5	2.4×10^5	54.5
Bathtub curve	–	–	2.2×10^5	67.1

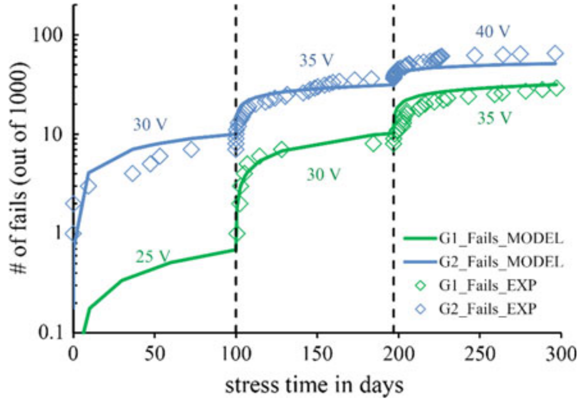


Fig. 12.74 Results of a long-time gate stress test at 150 °C on 1000 SiC trench MOSFETs with a rated maximum gate voltage of 15 V. Figure from [Sie17]

A statistical evaluation of the failures observed in group G1 at 30 V gate voltage (lower curve in Fig. 12.74) is shown in Fig. 12.75. In contrast to the previously discussed probability plots, the majority of the elements have not failed. Only 8 failed devices out of 1000 MOSFETs failed in 100 days test time at 30 V gate voltage. The MOSFETs that survived the test are taken into account as censored values. Elements without failure are marked as right-censored in the analysis. The shape factor of the Weibull distribution is 0.37, which verifies that this test produces early-life failures. The poor accuracy of the estimated shape factor is attributed to the small fraction of failed elements; it could be considerably improved by an increased test duration with more failures or by an increased stress, i.e. a higher gate voltage in this example

When the censored values are not taken into account, a totally different result is obtained, as Fig. 12.76 shows. This probability plot would be correct, if 7 MOSFETs were tested and all devices failed. However, in the test published in [Sie17] only a small fraction of elements had failed while more than 99% of devices survived the test level at 30 V gate voltage.

Statistical tools can also be applied to analyze field failures. However, if a short circuit of a device occurs in a wire bonded power module, the high energy involved in power electronic systems in general causes massive destruction and the identification of the root cause seldom possible. Therefore, only the accumulated bathtub curve in Fig. 12.72 can be determined. Furthermore, field failures are only found in a small fraction of total systems in operation. This is same situation as discussed in Figs. 12.75 and 12.76. The challenge in analyzing field failures is to collect the information on the operation time of all functional systems in the same application. If this information is

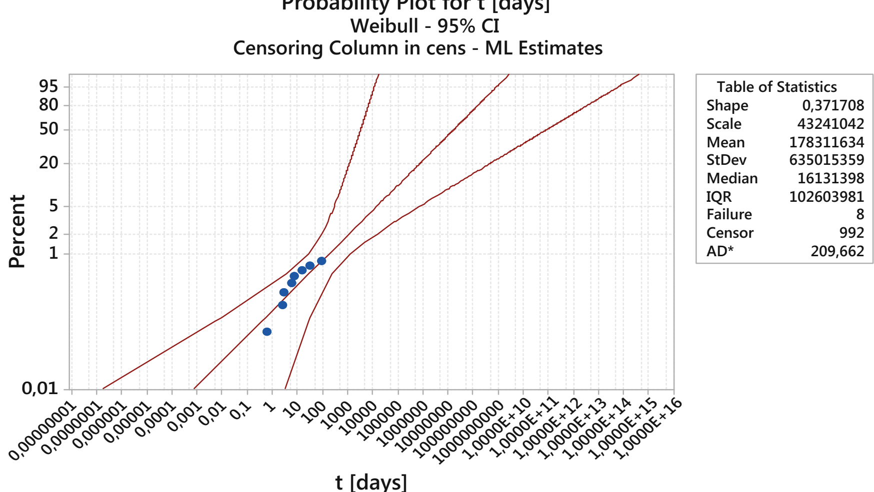


Fig. 12.75 Probability plot for the failures at 30 V gate voltage of SiC MOSFET of group G1 (lower curve) in Fig. 12.74 by MiniTab®

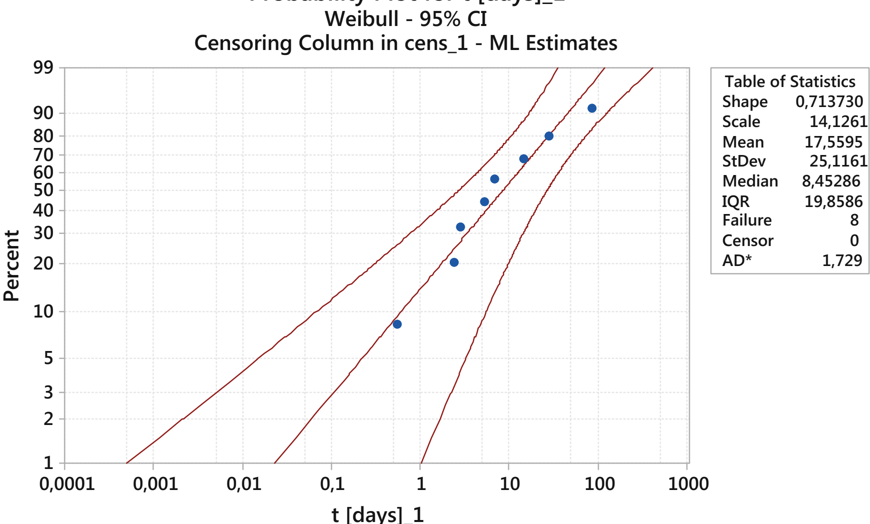


Fig. 12.76 Probability plot for the failures at 30 V gate voltage of SiC MOSFET of group G1 (lower curve) in Fig. 12.74 without censored values by MiniTab®

available or if it can be approximated by valid assumptions, statistical analysis can deliver estimations on the number of expected failures in the future.

12.10 Further Reliability Tests

A test sequence similar to the example given in Table 12.1 is mandatory for all series products of a power module manufacturer. However, additional test sequences can be negotiated for specific applications.

For applications in extreme environmental conditions, special tests under corrosive atmospheres are recommended. Corrosive gases can interfere seriously with the reliable function of power modules. The silicone soft mold represents almost no protection against corrosive gases. SO_2 interacts with all metal surfaces except noble metals, H_2S is highly corrosive for silver and silver alloys and unprotected Cu surfaces and Cl_2 together with high humidity will produce HCl, which corrodes non-noble metals, especially Al. The contribution of NO_x is not fully understood today, but its corrosion effect in connection with humidity is comparable to the impact of SO_2 and H_2S. The reliability of non-hermetically sealed modules must be verified in accelerated corrosive atmosphere tests in single or mixed corrosive gas environments for these applications.

Similar corrosion effects are connected with the impact of salt spray. This specific stress condition is found particularly in seaside or off-shore applications, which are typical environments for wind generator systems. The NaCl dissociates in

aqueous solution and produces HCl, which over time can penetrate the soft mold cover layer and dissolve the Al wire bonds. If non-hermetically sealed modules are used in these environments, an additional protection must be implemented in the system to prevent corrosive degradation of power modules and driver PCBs.

Additional tests can be conducted for special application conditions to verify the suitability of power electronics modules in extreme environments.

The standard test procedures discussed in the previous sections always use new power modules fresh from production for every reliability test. Some experts in the field of reliability testing propose the combination of tests to increase the reliability of power electronic components. Even though the consecutive or simultaneous application of different stress conditions could accelerate the time to failure of a power module, the number of possible sequences and combinations is quite high and there is no basis of experience for such test conditions.

However, the idea was adapted in the concept of HALT/HASS testing. The ‘Highly Accelerated Lifetime Test’ (HALT) is actually not a test, but much more a test philosophy, that accompanies the development process. During the HALT procedure, a single device parameter is selected and it is stepwise increased even far beyond the specification limit until the sample finally fails. Typical examples are the stepwise increase of the insulation test voltage until failure or the stepwise increase of the temperature swing during passive temperature cycling test. The goal of this procedure is to increase the stress to the destruction limit. From the analysis of the failure, small modifications can sometimes improve the robustness of a design considerably.

In a second step, a ‘Highly Accelerated Stress Screening’ (HASS) can be applied. Using the failure limits resulting form the HALT procedure, which typically are far beyond the specification limit, a HASS stress level can be defined, which is selected higher than the specification limit but a sufficient safety margin away from the destruction limit. For a fixed period of time 100% of a production is then subjected to an accelerated stress screening test, which allows to identify weak parts in a series production and improve the production process to enhance the product reliability. In this test sequence, combined stress levels are common. A typical example is the operation of a power module with high load currents under varying ambient temperature swings, sometimes even combined with the stress of mechanical vibration.

The general philosophy is developed further in the proposed concept of robustness validation [SAE08]. This concept combines experimentally determined failure limits and simulation results and compares these limits with the reliability requirements of applications, defined in mission profiles and specifications to increase the reliability of automotive electrical or electronic modules.

A general problem is the limited knowledge of the stress levels and the distribution of stress requirements in real applications. The implementation of the digital driver technology in power electronic systems gives the opportunity to monitor and record characteristic parameters during field operation and therefore has the potential to deliver more realistic data on the stress requirements of realistic field applications.

References

[Alb05] Albadri, A.M., Schrimpf, R.D., Walker, D.G., Mahajan, S.V.: Coupled electro-thermal simulations of single event burnout in power diodes. Trans. Nucl. Sci. **52**, pp. 2194–2199 (2005)

[All84] Allkofer, O.C., Grieder, P.K.F.: "Cosmic rays on earth", Physik Daten 25, Fachinformationszentrum Energie, Physik, Mathematik (1984)

[Amr04] Amro, R., Lutz, J., Lindemann, A.: Power cycling with high temperature swing of discrete components based on different technologies. In: Proceedings of PESC, pp. 2593–2598. Aachen (2004)

[Amr06] Amro, R., Lutz, J., Rudzki, J., Sittig, R., Thoben, M.: Power cycling at high temperature swings of modules with low temperature joining technique. In: Proceedings of ISPSD, pp. 1–4. Naples (2006)

[Ara08] Arab, M., Lefebvre, S., Khatir, Z., Bontemps, S.: Investigations on ageing of IGBT transistors under repetitive short-circuits operations In: Proceedings of PCIM Europe. Nuremberg (2008)

[Bay08] Bayerer, R., Licht, T., Herrmann, T., Lutz, J., Feller, M.: Model for power cycling lifetime of IGBT modules – various factors influencing lifetime. In: Proceedings of CIPS, pp. 37–42. Nuremberg (2008)

[Bay16] Bayerer, R., Lassmann, M., Kremp, S.: Transient hygrothermal-response of power modules in inverters – the basis for mission profiling under climate and power loading. IEEE Trans. Power Electron. **31**, pp. 613–620 (2016)

[Bei16] Beier-Möbius, M., Lutz, J.: Breakdown of gate oxide of 1.2 kV SiC-MOSFETs under high temperature and high gate voltage. In: Proceedings of PCIM Europe, pp. 172–179. Nuremberg (2016)

[Bei17] Beier-Möbius, M., Lutz, J.: Breakdown of gate oxide of SiC-MOSFETs and Si-IGBTs under high temperature and high gate voltage. In: Proceedings of PCIM Europe, pp. 365–372. Nuremberg (2017)

[Beu89] Beuhler, A.J., Burgess, M.J., Fjare, D.E., Gaudette, J.M., Roginski, R.T.: Moisture and purity in polyimide coatings. In: Material Research Society Symposium Proceedings, vol 154, pp. 73–90, doi:https://doi.org/10.1557/PROC-154-73 (1989)

[Bla75] Blackburn, D.L., Oettinger, F.F.: Transient thermal response measurements of power transistors. IEEE Trans. Ind. Electr. Control Instrum., IECI-22, pp. 134–141 (1975)

[Blu10] Blümer, J.: Partikel in der Pampa. Physik J **9**, 31–36 (2010)

[Cia96] Ciappa, M., Malberti, P.: Plastic-strain of aluminium interconnections during pulsed operation of IGBT multichip modules. Qual. Reliab. Eng. Int. **12**, pp. 297–303 (1996)

[Cia01] Ciappa, M.: Some reliability aspects of IGBT modules for high-power applications. Dissertation, ETH Zürich, (2001)

[Cia02] Ciappa, M.: Selected failure mechanisms of modern power modules. Microelectron. Reliab. **42**, pp. 653–667 (2002)

[Cia08] Ciappa, M.: Lifetime modeling and prediction of power devices. In: Proceedings of CIPS, pp. 27–35. Nuremberg (2008)

[Cle97] Clech, J.: Solder reliability solutions: a PC-based design-for-reliability tool. Soldering Surface Mount Technol. **9**, pp. 45–54 (1997)

[Cle02] Clech, J.: Review and analysis of lead-free solder material properties. Report to NIST, on: http://www.metallurgy.nist.gov/solder/clech/Intro-duction.htm (2002)

[Con15] Consentino, G., Laudani, M., Privitera, G., Pace, C., Giordano, C., Hernandez, J.: Are SiC HV power MOSFETs more robust of standard silicon devices when subjected to terrestrial neutrons? In: Proceedings of PCIM Europe, pp. 512–517. Nuremberg (2015)

[Coo97] Cooper, Jr, J.A.: Oxides on SiC. In: IEEE/Cornell Conference on Advanced Concepts in High Speed Semiconductor Devices and Circuits, pp. 236–243. Ithaca (1997)

[DaG07] DasGupta, S., Witulski, A.F., Bhuva, B.L., Alles, M.L., Reed, R.A., Amusan, O.A., Ahlbin, J.R., Schrimpf, R.D., Massengill, L.W.: Effect of well and substrate potential modulation on single event pulse shape in deep submicron CMOS. IEEE Trans. Nucl. Sci. **54**, pp. 2407–2412 (2007)

[Dar02] Darveaux, R.: Effect of simulation methodology on solder joint crack growth correlation and fatigue life prediction. J. Electron. Packag. **124**(3), pp. 147–154 (2002)

[Das17] Dashdondog, E., Harada, S., Shiba, Y., Sudo, M., Omura, I.: The failure rate calculation method for high power devices in low earth orbit. In: International Symposium Space Technology and Science, pp. 1–5. Matsuyama City (2017)

[Dep06] Déplanque, S., Nuchter, W., Wunderle, B., Schacht, R., Michel, B.: Lifetime prediction of SnPb and SnAgCu solder joints of chips on copper substrate based on crack propagation FE-analysis. In: Proceedings of EuroSime, pp. 1–8. Como (2006)

[Dep07] Déplanque, S.: Lifetime prediction for solder die-attach in power applications by means of primary and secondary creep. Ph.D. thesis, (2007)

[Dow82] Downing, S.D., Socie, D.F.: Simple rainflow counting algorithms. Int. J. Fatigue **4**, pp. 31–40 (1982)

[Feg16a] Felgemacher, C., Vasconcelos, S.A., Nöding, C., Zacharias, P.: Benefits of increased cosmic radiation robustness of SiC semiconductors in large power-converters. In: Proceedings of PCIM Europe, pp. 573–780. Nuremberg (2016)

[Feg16b] Felgemacher, C., Araújo, S.V., Zacharias, P., Nesemann, K., Gruber, A.: Cosmic radiation ruggedness of Si and SiC power semiconductors. In: Proceedings of ISPSD, pp. 235–238. Prague (2016)

[Fer08] Feller, M., Lutz, J., Bayerer, R.: Power cycling of IGBT- modules with superimposed thermal cycles. In: Proceedings of PCIM Europe. Nuremberg (2008)

[Gai13] Gaisser, T.K., Stanev, T., Tilav, S.: Cosmic ray energy spectrum from measurements of air showers. Front. Phys. **8**, pp. 748–758 (2013)

[Gri12] Griffoni, A., van Duivenbode, J., Linten, D., Simoen, E., Rech, P., Dilillo, L., Wrobel, F., Verbist, P., Groeseneken, G.: Neutron-induced failure in silicon IGBTs, silicon super-junction and SiC MOSFETs. Trans. Nucl. Sci. **59**, pp. 866–871 (2012)

[Hae12] Haertl, A., Soelkner, G., Pfirsch, F., Brekel, W., Duetemeyer, T.: Influence of dynamic switching on the robustness of power devices against cosmic radiation. In: Proceedings of ISPSD, pp. 353–356. Bruges (2012)

[Ham01] Hamidi, A., Kaufmann, S., Herr, E.: Increased lifetime of wire bond connections for IGBT power modules. In: Proceedings of IEEE APEC, pp. 1040–1044. Anaheim (2001)

[Hed14] Herold, C., Poller, T., Lutz, J., Schäfer, M., Sauerland, F., Schilling, O.: Power cycling capability of modules with SiC-Diodes. In: Proceedings of CIPS, pp. 36–41 (2014)

[Hed16] Herold, C., Franke, J., Bhojani, R., Schleicher, A., Lutz, J.: Requirements in power cycling for precise lifetime estimation. Microelectron. Reliab. **58**, pp. 82–89 (2016)

[Hed17] Herold, C., Sun, J., Seidel, P., Tinschert, L., Lutz, J.: Power cycling methods for SiC MOSFETs. In: Proceedings of ISPSD, pp. 367–370. Sapporo (2017)

[Hel97] Held, M., Jacob, P., Nicoletti, G., Sacco, P., Poech, M.H.: Fast power cycling test for IGBT modules in traction application. In: Proceedings of Power Electronics and Drive Systems, pp. 425–430 (1997)

[Her07] Herrmann, T., Feller, M., Lutz, J., Bayerer, R., Licht, T.: Power cycling induced failure mechanisms in solder layers. In: Proceedings of EPE. Aalborg (2007)

[Heu14] Heuck, N., Guth, K., Ciliox, A., Thoben, M., Oeschler, N., Krasel, S., Speckels, R., Böwer, L.: Aging of new interconnect-technologies of power modules during power cycling. In: Proceedings of CIPS, pp. 69–74. Nuremberg (2014)

[IEC10] Process management for avionics – Atmospheric radiation effects – part 1: accommodation of atmospheric radiation effects via single event effects within avionics electronic equipment, E DIN EN 62396-1:2010-11 (IEC 107/129/DTS:2010)

[IEC13] Process management for avionics – Atmospheric radiation effects – part 4: design of high voltage aircraft electronics managing potential single event effects, IEC 62396-4, Edition 1.0, 2013-09

[Ibr16] Ibrahim, A., Ousten, J., Lallemand, R., Khatir, Z.: Power cycling issues and challenges of SiC-MOSFET power modules in high temperature conditions. Microelectron. Reliab. **58**, pp. 204–210 (2016)

[Joo06] Jooss, C., Lutz, J.: The evolution of the universe in the light of modern microscopic and high-energy physics. In: Proceedings of AIP Conference, vol. 822, pp. 200–205 (2006) https://cds.cern.ch/record/945132

[Jun15] Junghaenel, M., Schmidt, R., Strobel, J., Scheuermann, U.: Investigation on isolated failure mechanisms in active power cycle testing. In: Proceedings of PCIM Europe, pp. 251–258. Nuremberg (2015)

[Jun17] Junghaenel, M., Scheuermann, U.: Impact of load pulse duration on power cycling lifetime of chip interconnection solder joints, Micorelctron. Reliab. **76–77**, pp. 480–484 (2017)

[Kab94] Kabza, H., Schulze, H.-J., Gerstenmaier, Y., Voss, P., Wilhelmi, J., Schmid, W., Pfirsch, F., Platzöder, K.: Cosmic radiation as a possible cause for power device failure and possible countermeasures. In: Proceedings of ISPSD, pp. 9–12. Davos (1994)

[Kai04] Kaindl, W., Soelkner, G., Becker, H.W., Meijer, J., Schulze, H.J., Wachutka, G.: Physically based simulation of strong charge multiplication events in power devices triggered by incident ions. In: Proceedings of ISPSD, pp. 257–260. Kitakyushu (2004)

[Kai05] Kaindl, W.: Modellierung höhenstrahlungsinduzierter Ausfälle in Halbleiterleistungsbauelementen. Dissertation, Munich (2005)

[Kam04] Kaminski, N.: Failure rates of HiPak modules due to cosmic rays. ABB Application Note 5SYA 2042–02 (2004)

[Kim97] Kimura, M.: Oxide breakdown mechanism and quantum physical chemistry for time-dependent dielectric breakdown. In: Proceedings of IEEE International 35th Annual Reliability Physics Symposium, pp. 190–200. Denver (1997)

[Kov15] Kovacevic-Badstuebner, I., Schilling, U., Kolar, J.W.: Modelling for the lifetime predication of power semiconductor modules. Reliab. Power Electron. Converter Syst., pp. 103–140 (2015)

[Lee88] Lee, J.C., Chen, I.-C., Chenming, H.: Modeling and Characterization of Gate Oxide Reliability. IEEE Trans. Elect. Dev. **35**(12), pp. 2268–2278 (1988)

[Lei09] Lei, T.G., Calata, J.N., Lu, G.-Q.: Effects of large temperature cycling range on direct bond aluminum substrate. Trans. Device Mater. Reliab. **9**, pp. 563–568 (2009)

[Lip99] Lipkin, L.A., Palmour, J.W.: Insulator investigation on SiC for improved reliability. IEEE Trans. Elect. Dev. **46**, pp. 525–532 (1999)

[Lut94] Lutz, J., Scheuermann, U.: Advantages of the new controlled axial life-time diode. In: Proceedings of PCIM, pp. 163–169. Nuremberg (1994)

[Lut08] Lutz, J., Herrmann, T., Feller, M., Bayerer, R., Licht, T., Amro, R.: Power cycling induced failure mechanisms in the viewpoint of rough temperature environment. In: Proceedings of CIPS, pp. 55–58. Nuremberg (2008)

[Lut17] Lutz, J., Aichinger, T., Rupp, R.: Reliability evaluation. In: Suganuma, K.(eds.) Wide Bandgap Power Semiconductor Packaging: Materials, Components, and Reliability, to be published. Elsevier (2017)

[LV324] Qualification of Power Electronics Modules for Use in Motor Vehicle Components, General Requirements, Test Conditions and Tests, supplier portal of BMW:GS 95035, VW 82324 Group Standard, Daimler, (2014)

[Mat94] Matsuda, H., Fujiwara, T., Hiyoshi, M., Nishitani, K., Kuwako, A., Ikehara, T.: Analysis of GTO failure mode during DC voltage blocking. In: Proceedings of ISPSD, pp. 221–225. Davos (1994)

[McP85] McPherson, J.W., Baglee, D.A.: Acceleration factors for thin gate oxide stressing. In: 23rd Annual Reliability Physics Symposium, pp. 1–5 (1985)
[Mik01] Mikkelsen, J.J.: Failure analysis on direct bonded copper substrates after thermal cycle in different mounting conditions. In: Proceedings of PCIM, pp. 467–471. Nuremberg (2001)
[MIN18] Minitab® Statistical Software, Version 18, www.minitab.com
[Min45] Miner, M.A.: Cumulative damage in fatigue. J. App. Mech. **12**, pp. A152–A164 (1945)
[Mit15] Mitsuzuka, K., Yamada, S., Takenoiri, S., Otsu, M., Nakagawa, A.: Investigation of anode-side temperature effect in 1200 V FWD cosmic ray failure. In: Proceedings of ISPSD, pp. 117–120, Hong Kong (2015)
[Moz01] Morozumi, A., Yamada, K., Miyasaka, T.: Reliability design technology for power semiconductor modules. Fuji Electric. Rev. **47**, pp. 54–58 (2001)
[Nor96] Normand, E.: Correlation of in-flight neutron dosimeter and SEU measurements with atmospheric neutron model. Trans. Nucl. Sci. **48**, pp. 1996–2003 (2001)
[Oet73] Oettinger, F.F., Gladhill, R.L.: Thermal response measurements for semiconductor device die attachment evaluation. Int. Electron Dev. Meet. **19**, pp. 47–50 (1973)
[Pad68] Paddock, A., Black, J.R.: Hillock formation on aluminum thin films presented at the Electrochemical Society Meeting, Boston, May 5–9 (1968)
[Par63] Paris, P., Erdogan, F.: A critical analysis of crack propagation laws. ASME J. Basic Eng. **85**, pp. 528–533 (1963)
[Pfi10] Pfirsch, F., Soelkner, G.: Simulation of cosmic ray failures rates using semiempirical models. In: Proceedings of ISPSD, pp. 125–128, Hiroshima (2010)
[Pfu76] Pfüller, S.: Halbleiter Messtechnik. VEB Verlag Technik, Berlin, pp. 89ff, (1976)
[Phi71] Philofsky, E., Ravi, K., Hall, E., Black, J.: Surface reconstruction of aluminum metallization – a new potential wearout mechanism. In: Proceedings of Reliability Physics Symposium, vol. 9, pp. 120–128. Las Vegas (1971)
[Pol10] Poller, T., Lutz, J.: Comparison of the mechanical load in solder joints using SiC and Si chips. In: Proceedings of ISPS. Prague (2010)
[Ram00] Ramminger, S., Seliger, N., Wachutka, G.: Reliability model for Al wire bonds subjected to heel crack failures. Microelectron. Reliab. **40**, pp. 1521–1525 (2000)
[Rau04] Rausand, M., Hoyland, A.: System reliability theory – models, statistical methods, and applications, 2nd edn. Wiley, Hoboken (2004)
[Rup14] Rupp, R., Gerlach, R., Kabakow, A., Schörner, R., Hecht, C.h., Elpelt, R., Draghici, M.: Avalanche behaviour and its temperature dependence of commercial SiC MPS diodes: influence of design and voltage class. In: Proceedings of ISPSD, pp. 67–70 (2014)
[Sad16] Sadik, D.-P., Nee, H.-P., Giezendanner, F., Ranstad, P.: Humidity Testing of SiC Power MOSFETs, pp. 3131–3136. IPEMC-ECCE, Asia (2016)
[SAE08] SAE/ZVEI.: Handbook for Robustness Validation of Automotive Electrical/Electronic Modules. www.zvei.org/ecs, (2008)
[San69] Santoro, C.J.: Thermal cycling and surface reconstruction in aluminum thin films. J. Electrochem. Soc. **116**, pp. 361–364 (1969)
[Scn99] Scheuermann, U.: Power module design for HV-IGBTs with extended reliability. In: Proceedings of PCIM, pp. 49–54. Nuremberg (1999)
[Scn02b] Scheuermann, U., Hecht, U.: Power cycling lifetime of advanced power modules for different temperature swings. In: Proceedings of PCIM, pp. 59–64. Nuremberg (2002)
[Scn08] Scheuermann, U., Beckedahl, P.: The road to the next generation power module – 100% solder free design. In: Proceedings of CIPS, pp. 111–120. Nuremberg (2008)
[Scn09] Scheuermann, U., Schmidt, R.: Investigations on the $V_{CE}(T)$ method to determine the junction temperature by using the chip itself as sensor. In: Proceedings of PCIM Europe, pp. 802–807. Nuremberg (2009)

[Scn11] Scheuermann, U., Schmidt, R.: Impact of solder fatigue on module lifetime in power cycling tests. In: Proceedings of EPE. Birmingham (2011)

[Scn13] Scheuermann, U., Schmidt, R.: A new lifetime model for advanced power modules with sintered chips and optimized al wire bonds. In: Proceedings of PCIM Europe, pp. 810–817. Nuremberg (2013)

[Scn15a] Scheuermann, U.: Packaging and reliability of power modules – principles, achievements and future challenges. In: Proceedings of PCIM Europe, pp. 35–50. Nuremberg (2015)

[Scn15b] Scheuermann, U., Schilling, U.: Cosmic ray failures of power modules – the diode makes the difference. In: Proceedings of PCIM Europe, pp. 494–501. Nuremberg (2015)

[Scn16] Scheuermann, U., Schilling, U.: Impact of device technology on cosmic ray failures in power modules. IET Power Electron. **9**, pp. 2027–2035 (2016)

[Scr10] Schuler, S., Scheuermann, U.: Impact of test control strategy on power cycling lifetime. In: Proceedings of PCIM Europe, pp. 355–360. Nuremberg (2010)

[Sct12a] Schmidt, R., Scheuermann, U.: Separating failure modes in power cycling tests. In: Proceedings of CIPS, pp. 97–102. Nuremberg (2012)

[Sct12b] Schmidt, R., Koenig, C., Prenosil, P.: Novel wire bond material for advanced power module packages. Microelectron. Reliab. **52**, pp. 2283–2288 (2012)

[Sct13] Schmidt, R., Zeyss, F., Scheuermann, U.: Impact of absolute junction temperature on power cycling lifetime. In: Proceedings of EPE, pp. 1–10. Lille (2013)

[Sct17] Schmidt, R., Werner, R., Casady, J., Hull, B., Barkley, A.: Power cycle testing of sintered SiC-MOSFETs. In: Proceedings of PCIM Europe, pp. 694–701. Nuremberg (2017)

[Scu06] Schulze, H.J., Lutz, J.: Patent application DE 102006046845A1, 2.10.2006

[Sho11] Shoji, T., Nishida, S., Ohnishi, T., Fujikawa, T., Nose, N., Hamada, K., Ishiko, M.: Reliability design for neutron induced single-event burnout of IGBT. Trans. Ind. Appl. **131**, pp. 992–999 (2011)

[Sho13] Shoji, T., Nishida, S., Hamada, K.: Triggering mechanism for neutron induced single-event burnout in power devices. Jpn. J. Appl. Phys. 52, pp. 04CP06-1–04CP06-7 (2013)

[Sho14] Shoji, T., Nishidas, S., Hamada, K., Tadano, H.: Cosmic ray induced single-event burnout in power devices. In: Proceedings of ISPS, pp. 5–14. Prague (2014)

[Sie17] Siemieniec, R., Peters, D., Esteve, R., Bergner, W., Kück, D., Aichinger, T., Basler, T., Zippelius, B.: A SiC trench MOSFET concept offering improved channel mobility and high reliability. In: Proceedings of EPE ECCE Europe. Warzaw (2017)

[Soe00] Soelkner, G., Voss, P., Kaindl, W., Wachutka, G., Maier, K.H., Becker, H.W.: Charge carrier avalanche multiplication in high-voltage diodes triggered by ionizing radiation. Trans. Nucl. Sci. **47**, pp. 2365–2372 (2000)

[Sti09] Stiasny, T: Cosmic Rays Failure in Power Devices Part 1. ISPSD Short Course Lecture Notes, pp. 5–16, Barcelona (2009)

[Tit98] Titus, J.L., Wheatley, C.F.: Proton-induced dielectric breakdown of power MOSFETs. Trans. Nucl. Sci. **45**, pp. 2891–2897 (1998)

[Uem11] Uemura, H., Iura, S., Nakamura, K., Kim, M., Stumpf, E.: Optimized design against cosmic ray failure for HVIGBT modules. In: Proceedings of PCIM Europe, pp. 10–15. Nuremberg (2011)

[Was86] Waskiewicz, A.E., Groninger, J.W., Strahan, V.H., Long, D.M.: Burnout of power MOS transistors with heavy ions of 252-Cf. Trans. Nucl. Sci. NS-33, pp. 1710–1713 (1986)

[Wei15] Weiß, C.: Höhenstrahlungsresistenz von Silizium-Hochleistungs-bauelementen. Ph.D. Thesis, Munich 2015

[Wil12] Wilkinson, D.: Space environment overview. https://commons.wikimedia.org/wiki/File:SpaceEnvironmentOverview_From_19830101.jpg, SpaceEnvironmentOverview From 19830101, Excerpt by J.Lutz,https://creativecommons.org/licenses/by-sa/3.0/legalcode

[Yag13] Yang, L., Agyakwa, P.A., Johnson, C.M.: Physics-of-failure lifetime prediction models for wire bond interconnects in power electronic modules. IEEE Trans. Dev. Mater. Reliab., pp. 9–17 (2013)

[Yan00] Yano, H., Kimoto, T., Matsunami, H., Bassler, M., Pensl, G.: MOSFET performance of 4H-, 6H-, and 15R-SiC processed by dry and wet oxidation. Mater. Sci. Forum **338–342**, pp. 1109–1112 (2000)

[Zel94] Zeller, H.R.: Cosmic ray induced breakdown in high voltage semiconductor devices, microscopic model and phenomenological lifetime prediction. In: Proceedings of ISPSD, pp. 339–340. Davos (1994)

[Zel95] Zeller, H.R.: Cosmic ray induced failures in high power semiconductors. Solid State Electr. **38**, pp. 2041–2046 (1995)

[Zom90] Zombeck, M.V.: Handbook of Space Astronomy and Astrophysics, 2 edn. Cambridge University Press, Cambridge (1990)

[Zor14] Zorn, C., Kaminski, N.: Temperature humidity bias (THB) testing on IGBT modules at high bias levels. In: Proceedings of CIPS, pp. 101–107. Nuremberg (2014)

[Zor15] Zorn, C., Kaminski, N.: Acceleration of temperature humidity bias (THB) testing on IGBT modules by high bias levels. In: Proceedings of ISPSD, pp. 385–388. Hong Kong (2015)

Chapter 13
Destructive Mechanisms in Power Devices

This chapter will deal with some destructive mechanisms in power devices, and typical failure pictures for them will be shown. Failure analysis requires a lot of experience, especially regarding the conditions in the power circuit at failure, which must be carefully considered. Although some of the failure pictures appear to be similar, it is difficult to draw conclusions only from pictures. However in practice the engineer often has the problem to find the reason for failures, and the following sections might be helpful.

13.1 Thermal Breakdown—Failures by Excess-Temperature

In Chap. 2 the intrinsic carrier density n_i was explained, it strongly depends on the temperature as described in Eq. (2.6). In silicon, n_i amounts approx. to 10^{10} cm^{-3} at room temperature and it is negligible compared to the background doping. But n_i increases rapidly with increasing temperature. Therefore, at very high temperatures, the thermal generation becomes the dominant mechanism for the creation of carriers.

With the introduction of an intrinsic temperature T_{int} similar to [Gha77], one can estimate, when the rise of some critical mechanisms in a device with increasing temperature can be expected. T_{int} is that temperature, at which the density of carriers n_i generated by thermal generation is equal to the background doping N_D. It is drawn in Fig. 13.1 as a function of N_D. Below T_{int} the carrier density is only weakly dependent on temperature. Above T_{int} the carrier density increases exponentially with temperature according to Eq. (2.6). From Fig. 13.1 we can see that for a high voltage device, which requires a N_D in the range of 10^{13} cm^{-3}, T_{int} will be reached at a much lower temperature compared to a device with a lower voltage rating, at which e.g. a range of 10^{14} cm^{-3} is used for N_D.

© Springer International Publishing AG 2018

J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8_13

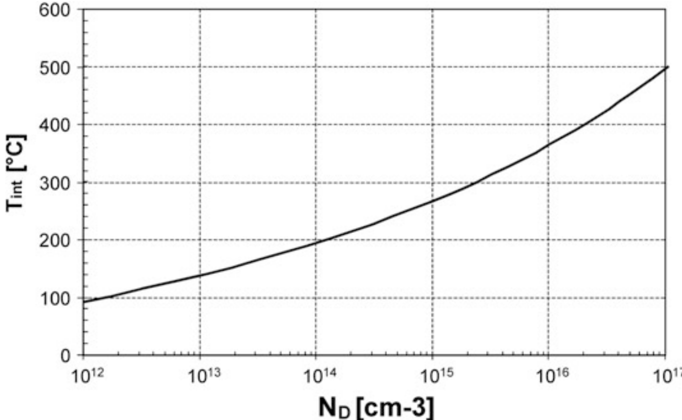


Fig. 13.1 Intrinsic temperature in silicon as a function of background doping

However, this point of view is much too simplified. The intrinsic carrier density is defined for thermal equilibrium, while a device in application is usually never operating in thermal equilibrium. Therefore, one must also consider in which operation mode and by which effect an increased or locally increased temperature is reached.

For low voltage MOSFETs operated in a short time interval in avalanche breakdown, it was found that destruction occurs if the temperature increases just below the temperature T_{int}, which was in the range 320 °C for the used 60 V MOSFET [Ron97]. This agrees with Fig. 13.1, if the doping N_D of 4×10^{15} cm^{-3} is assumed which is reasonable for this voltage range.

If a bipolar device is in the forward conduction mode, e.g. in a surge current event, it is flooded with free carriers in a density in the range above 10^{16}, even slightly above 10^{17} cm^{-3} can also occur. When the thermal generation amounts to a carrier density in this range, it becomes the dominating mechanism. Therefore, in such short time events, very high temperatures may occur without a failure in the device. In the surge current mode, a T_{int} up to the range of 500 °C is to be expected.

In the blocking mode, a space-charge region builds up and carriers are removed from the depleted zone. Their density is given by the leakage current, which is low for most modern power devices (except for gold-diffused devices). Thermal stability is now determined by the height of the leakage current.

Therefore the associated mechanism which leads to a high temperature must be taken into account. For the investigation of stability, one has to consider the temperature dependency of the electrical mechanism which leads to high losses and therewith high temperature.

If the heating is caused by a high leakage current, the leakage current will further increase in regions of high temperature. These regions will get hotter, which again leads to more increased leakage current. Such behavior shall be termed here as a

positive feedback. If the high power loss density can not be extracted by the cooling of the device, the device will be destroyed inevitably.

If the high losses are generated by a high voltage above the breakdown voltage V_{BD}, which drives the device in the avalanche breakdown, losses are created and the temperature increases. However, with increasing temperature, V_{BD} increases. The region, where avalanche breakdown occurs, will move to the regions of the device where the temperature is lower. Even if such electrical mechanisms lead to local filaments, the increase of temperature releases the local stress, leading to a *negative feedback effect*.

However if the temperature reaches T_{int}, the thermal generation will become the dominating effect, and then the temperature increase acts as a positive feedback. Inhomogeneities in current density, even if they are small, will be amplified rapidly. If the temperature reaches T_{int}, one has to expect current tubes or filaments, considering a device with an area above some mm^2.

If P_{gen} is the generated power density and P_{out} is the power density that can be maximally be drawn out via the package and heatsink, one can express a condition for thermal runaway [Lin08]:

$$\frac{\partial P_{gen}}{\partial T} > \frac{\partial P_{out}}{\partial T} \tag{13.1}$$

If this condition is fulfilled for a stationary operation point, a fast exponential temperature increase will occur. Equation (13.1) is a general form of an equation that was used for thermal stability of bipolar transistors [How74]:

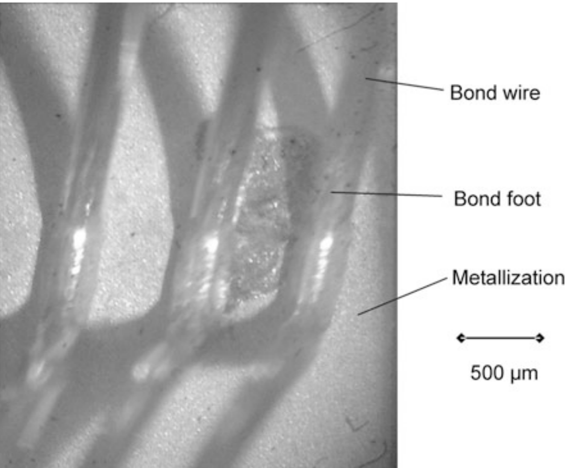
$$S = R_{th} \cdot V_C \frac{\partial I_C}{\partial T} \tag{13.2}$$

If the stability factor S increases > 1, any temperature disturbance will grow and thermal runaway will occur.

Finally, the destruction of the device is always due to high temperature. Failed devices show local regions of molten semiconductor material. If the temperature increase occurs locally, in a very small point-like area of the device, one can find even cracks in the crystal lattice. However it must be distinguished by which effect the temperature increase was generated. Some of these effects will be described below.

As a simple example, Fig. 13.2 shows an IGBT device, which has failed due to very high power losses. The IGBT was only stressed with forward current, and the failure was caused by a too low gate voltage V_G. One can recognize a comparatively large molten area (> 1 mm^2) at the emitter side. It is located typically close to the centre of the device and close to the bond wires.

If over-temperature occurs in an application in which the device is switched between the blocking mode and the conducting mode at a high frequency, the destruction picture may be different. With increased temperature, the blocking



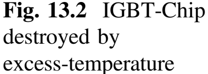


Fig. 13.2 IGBT-Chip destroyed by excess-temperature

capability is lost first. In almost all devices with planar junction termination the breakdown will occur at the edge. Therefore the point of destruction will be at the edge of the device, or at least a small part of the edge should be included.

13.2 Surge Current

In the application of a diode or a thyristor in a rectifier, momentary high over-current pulses can occur. Therefore the possible surge current is determined and given in the datasheets for rectifier diodes, fast diodes and thyristors. During the qualification of a diode, a single sinus half-wave of the grid current is imposed in the forward direction of the diode. Figure 13.3 shows the surge current measurement of a fast 1200 V diode with an area of 7 × 7 mm^2, the waveforms of current, voltage and additionally the power $p = v \cdot i$ are pictured as a function of time.

Due to the junction voltage of the measured diode and the junction voltage of a thyristor in the measurement equipment, the current pulse duration is not 10 ms as it should be for a grid frequency of 50 Hz, but it is 7.5 ms. In Fig. 13.4 the I–V characteristic from the measurement in Fig. 13.3 is shown. Since a high temperature is reached in the device, the characteristics split into an ascending and descending branch. In the descending branch the voltage drop is significantly lower.

For the dissipated power in Fig. 13.3 the temperature in the semiconductor is estimated in Fig. 13.5 with a thermal simulation using the Simulator SIMPLORER. The diode is packaged as shown in Fig. 11.13, the thickness of the Al_2O_3 ceramic is 0.63 mm. The heat flux is fed in the volume of the low-doped n^--layer of the device in the form of a sinus-square function within a time of 7.5 ms and amplitude of 3060 W, according to the measurement in Fig. 13.3. The temperature dependency of thermal conductivity in silicon is considered according Eq. (11.5), since this will

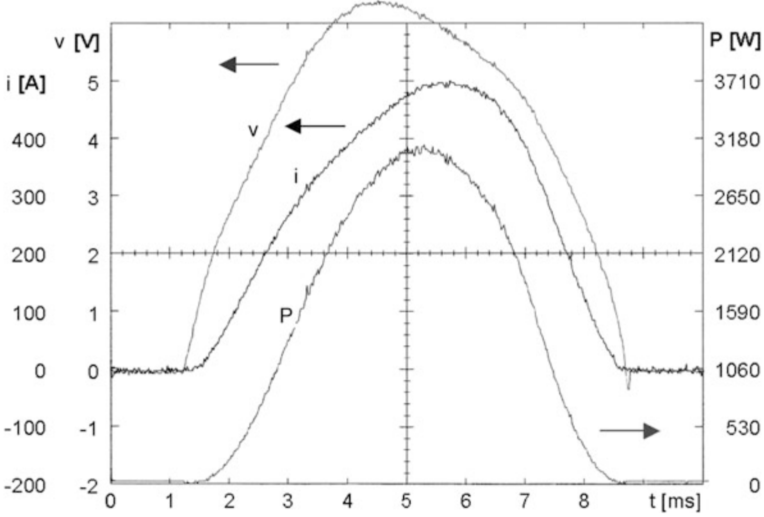


Fig. 13.3 Surge current load of a fast diode. Voltage, current and power depending on time

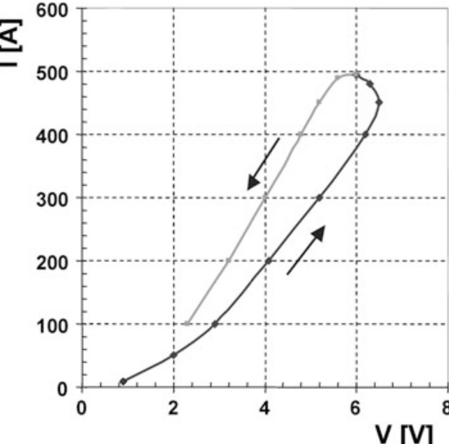


Fig. 13.4 I–V characteristic of the surge current load in Fig. 13.3

be significant for the result at the expected high temperature increase. In the estimation in Fig. 13.5, the temperature in the n^--layer (Si active) of the device increases up to 382 °C.

This high temperature may explain the strong difference of the forward voltage in Fig. 13.4 in the descending branch compared to the ascending branch. The voltage drop V_F consists of $V_F = V_j + V_{drift}$ · V_j decreases at high temperature because n_i is strongly increasing with temperature. For V_{drift}, the temperature dependency can be discussed using Eq. (5.47)

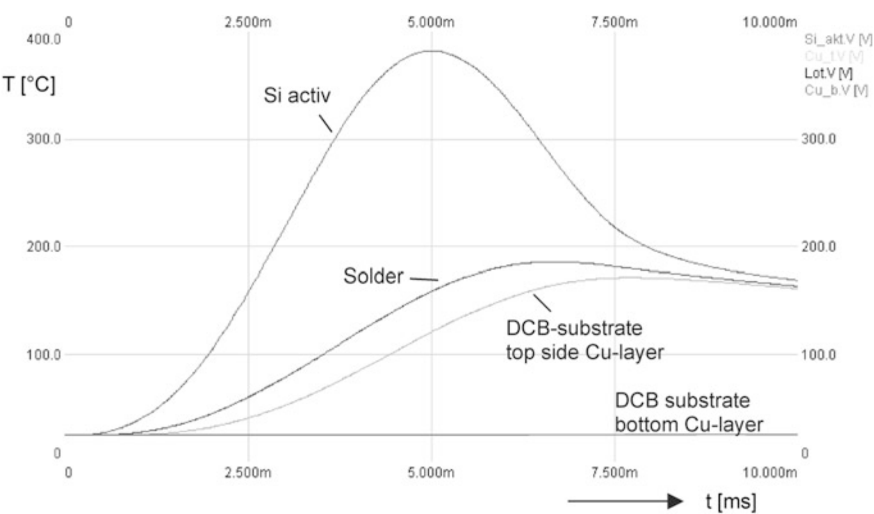


Fig. 13.5 Simulation of temperatures for the surge current event in Fig. 13.3

$$V_{drift} = \frac{w_B^2}{(\mu_n + \mu_p) \cdot \tau_{eff}} \tag{13.2}$$

where τ_{eff} contains the carrier lifetime τ_p as well as the emitter influence, see Eq. (5.52).

To this temperature dependency contribute the following effects:

- The carrier lifetime. It increases with temperature. This leads to a decrease of the forward voltage drop with temperature.
- The emitter-recombination. For this, Auger-Recombination is important, and at a high current density τ_{eff} will become very small. Additionally, in many modern devices the emitter depth is smaller than the diffusion length, and then the emitter depth must be used in the emitter-parameter. However, no strong temperature dependency of emitter recombination is to be expected.
- The mobilities. They decrease strongly with temperature. This effect leads to an increase of the forward voltage.
- The temperature dependency of the specific resistance of metallization and bond wires. This resistance increases with temperature.
- Finally the density of thermal generated carriers n_i depends strongly on temperature. If n_i significantly contributes to the density of free carries – this is in the range of 10^{17} cm^{-3} for surge current events – then a significant decrease of the forward voltage is to be expected. Now Eq. (13.2) is no longer valid. However, Eq. (5.34) which was used to derive the basics of forward conduction, can be written as

$$V_{drift} = \frac{j}{q} \int_0^{w_B} \frac{1}{\left(\mu_n(x) + \mu_p(x)\right) p(x)} dx \tag{13.3}$$

The density of free carriers $p(x) \approx n(x)$ is strongly increasing and V_{drift} is decreasing.

The form of the I–V characteristic will therefore be very different for different diode fabrication technologies. The behavior like Fig. 13.4 is normally observed in some special fast diodes. Finally, the failure can be caused by the following mechanisms:

(a) Melting of the top-side metallization. This occurs especially in bonded diodes in power modules.
(b) Mechanical destruction, cracks in the device, caused by very high temperature and mechanical stress resulting from thermal expansion.
(c) If finally n_i is dominating, the characteristic behavior of a resistor with a negative temperature coefficient occurs [Sil73]. The forward voltage decreases strongly. A positive feedback occurs. A filamentation of the current into tubes with very high current density is to be expected.

In [Sil73] it is estimated that the negative temperature coefficient of the resistance according mechanism (c) occurs after the temperature has increased to an amount, at which n_i approximates to 0.3 $\bar{n}$. There are some hints, that there is a dependency on the device area. Small diodes can bear a higher current density since there are fewer possibilities for the formation of filaments. If the failure occurs according (c), then it will be typically close to the edge of the active area, since these are the locations with the highest current densities.

For wire bonded diodes in IGBT modules, as shown in Fig. 11.13, the anode layer with a typical low junction depth is on the top. The heat dissipating volume is close to the metallisation layer and bond wires. For such diodes, failure according mechanism (a) is expected. In a more detailed simulation with the Sentaurus$_{\mathrm{TCAD}}$ device simulator [SYN07], it was found that the metallisation layer and the bond foot arrangement are of high influence to the occurring temperatures and the surge current capability [Hei08b]. The thicker the metallisation the better is the capability to absorb the heat especially in short times, since it acts as additional thermal capacitance. Another important factor on the surge current capability of diodes is the location and size of the contact area of the bond wires. A high ratio of bond foot area to the diodes anode will increase the surge current capability.

A surge current stress below the destruction limit will not lead to irreversible modifications in the semiconductor itself. If the surge current is increased above the value in Fig. 13.4, the split of the characteristics will increase and higher temperature will occur in the device. Finally destruction of the device occurs.

However, already in the simplified temperature estimation in Fig. 13.5 the temperature in the chip solder layer grows up to 186 °C. Such a temperature is already close to thermal softening of solder layers. Therefore irreversible

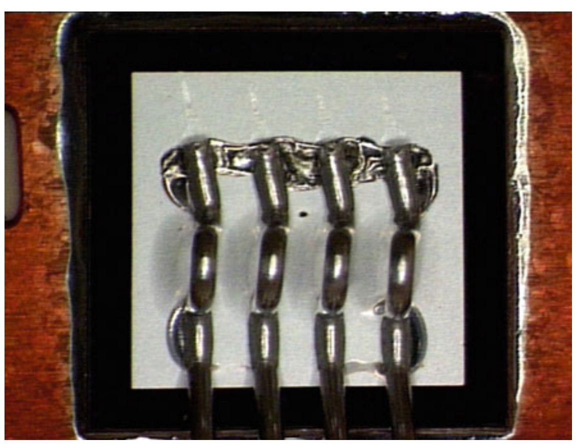


Fig. 13.6 Wire-bonded diode, destroyed by surge current

modifications in solder layers, and also in metallization and bond wires may occur. The surge current capability therefore is intended for singular overload events, and it is not intended for regular operation of a power semiconductor.

The temperature in the bottom Cu-layer of the DCB-substrate has only grown up to 27 °C, this is negligible. Therefore the influence of further components of the package can be neglected at surge current conditions, hence, all effects happen in the semiconductor and in the immediately adjacent layers.

The surge current capability of a fast diode is typically at 10 - 12 times of the rated current. The surge current capability of a diode for grid frequency operation or of a thyristor is typically in the range of 20 times of the rated current, since these devices are manufactured with high carrier lifetime, and the forward voltage drop is lower.

A diode which failed due to surge current is shown in Fig. 13.6. The molten regions close to the bond feet are typical. The hottest surface-spot at surge current is beside the bond feet [Hei08]. The failure mechanism here is according to (a). The molten area for this case is always in the active area of the device. Such pictures allow a clear identification as a surge current fault during analysis in the quality department of the device manufacturer.

The failure in an application must not be during a sinus-shaped pulse. The current pulse might be of a different shape. A general description is done by the i^2t value in data sheets, which holds for arbitrary current pulses. The failures are due to exorbitant heating of the device by too high currents and the occurrence of one of the described mechanisms.

13.3 Overvoltage – Voltage Above Blocking Capability

The blocking capability of power devices is limited by avalanche breakdown. Avalanche breakdown occurs above the rated voltage of the device. Most power devices can sustain some current in the avalanche breakdown mode. However the

data sheet of the manufacturer excludes operation of the device in the avalanche mode, if the device is not avalanche rated.

Several MOSFETs and diodes in the range up to 1000 V are avalanche rated; this allows short-time operation in the avalanche mode. The maximum dissipated energy in the avalanche E_{av} is specified in the data sheets of the manufacturers, in general form it is given by

$$E_{av} = \int_{t_{av}} V_{BD} \cdot i(t) dt \tag{13.4}$$

where V_{BD} is the device breakdown voltage, $i(t)$ the current pulse in avalanche and t_{av} is the time of the current pulse. Usually avalanche occurs at "unclamped inductive switching", where the MOSFET is turned-off in a circuit with an inductor L in series. The voltage now rises up to the breakdown-voltage, and the current decreases with

$$\frac{di}{dt} = \frac{V_{BD} - V_{bat}}{L} \tag{13.5}$$

If $i(t)$ decays linearly from $I_{av(peak)}$ to zero during t_{av}, as it is the case when the energy of an inductor is dissipated, (13.4) can be expressed as

$$E_{av} = \frac{1}{2} \cdot V_{BD} \cdot I_{av(peak)} \cdot t_{av} \tag{13.6}$$

and the dissipated energy in this case is the stored energy in the inductor $0.5 \cdot L \cdot I_{av(peak)}^2$ and additionally the energy $0.5 \cdot V_{bat} \cdot I_{av(peak)} \cdot t_{av}$ which is delivered by the voltage source V_{bat} during t_{av}. Avalanche rating of low-voltage MOSFETs can be up to E_{av} = 1 J, such energy can only be dissipated in single events and never in continuous mode at a high operation frequency.

The design of the MOSFETs for avalanche capability is in a way, that the breakdown occurs in the volume of the device, e.g. in a planar MOSFET at the p^+-layer in the center of the cell (Fig. 9.4) where the n^--base is narrowest. Avalanche capability is also possible for trench MOSFETs (Fig. 9.6), for these structures an additional design effort is necessary to avoid the location of avalanche at the trench corners [Kin05].

The occurrence of breakdown in the volume of the device is also possible for diodes with a beveled junction termination of positive angle (Fig. 4.22). The breakdown will occur at the edge for diodes with planar junction terminations with floating potential rings (Fig. 4.24). Well designed potential rings can also bear current in the avalanche mode, even if it flows mainly at the edge. However, above 1200 V, one can only rarely find avalanche rating, and if, it is with more restrictive conditions.

For avalanche capability, as second condition is that branches with negative differential resistance (NDR) must be avoided. This will be explained more in detail in the following.

For a higher rated blocking voltage the device must have a lower doping, see Fig. 3.17. If avalanche occurs, the shape of the electric field is triangular or trapezoidal, and electron hole pairs are generated in the region with high electric field. The generation depends exponentially on the electric field, and the main part is generated close to the pn-junction.

Holes are flowing to the anode and electrons to the cathode. Within the space-charge, we can include the generated carries p_{av} and n_{av} in the basic Eq. (2.107) according to their polarity, in one-dimensional expression as

$$\frac{dE}{dx} = \frac{q}{\varepsilon_0 \cdot \varepsilon_r} \left(N_{D^+} + p_{av} - n_{av} \right) \tag{13.7}$$

At the anode side boarder of the space-charge, no electrons arrive, and the reverse current is carried only by holes. We neglect the small contributions of diffusion current and recombination-center induced leakage current. Then at this position holds

$$p_{av} = \frac{j_R}{q \cdot v_{sat}(p)} \tag{13.8}$$

where j_R denotes the reverse current and $v_{sat}(p)$ denotes the saturation velocity of holes since at the given high electric field. At the cathode side boarder the arriving reverse current is pure electron current. The density of generated electrons is

$$n_{av} = \frac{j_R}{q \cdot v_{sat}(n)} \tag{13.9}$$

with $v_{sat}(n)$ of 1.05×10^7 cm/s at $T = 300$ K.

The generated free carriers influence the effective doping N_{eff} and thereby the shape of the electric field. The feedback is shown with device simulation in Fig. 13.7. At the left hand side, and at the pn-junction, it holds $N_{eff} = N_D + p_{av}$. With increasing j_R the gradient of the electric field becomes more steep according to Eqs. (13.7) and (13.8), and at the pn-junction occurs an increased field peak.

At the nn$^+$-side, N_{eff} is lowered since $N_{eff} = N_D - n_{av}$. If n_{av} is in the range of N_D, a part of the positively charged donors will be compensated. If we assume a high voltage diode with N_D of 1.1×10^{13} cm^{-3} as in Fig. 13.7a, only a current density $j_R = 19$ A/cm^2 is necessary to compensate the background doping for this design (a) completely. The gradient $d\text{E}/dx$ becomes flat. In Fig. 13.7b the same conditions are taken, but the doping N_D is increased to 1.7×10^{13} cm^{-3} for this design (b).

The voltage corresponds to the area below $\text{E}(x)$. For design (a) the negative voltage-area is always larger than the positive voltage area that occurs at the nn$^+$-junction. For design (b), up to a current density of 40 A/cm^2 there is a larger

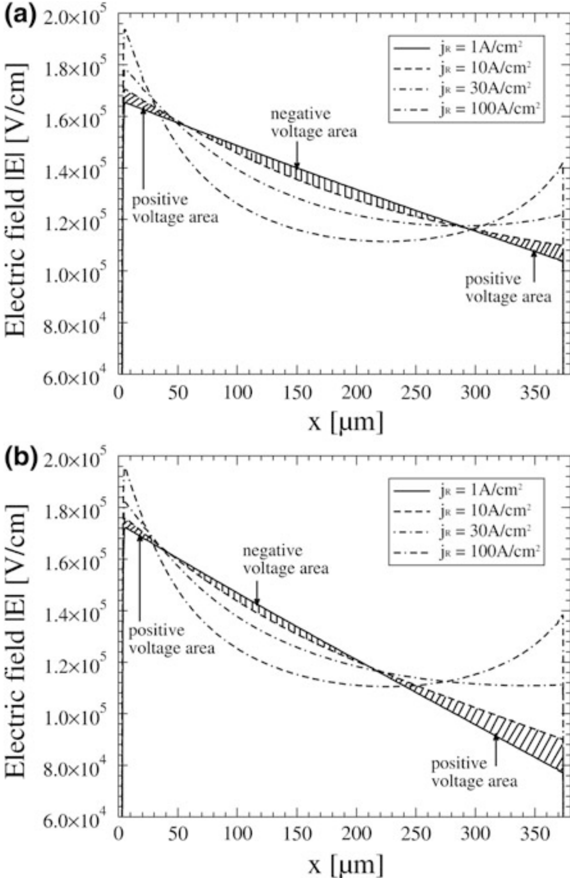


Fig. 13.7 Electric field |E| at low and at increased avalanche current. **(a)** Design with $N_D = 1.1 \times 10^{13}$ cm^{-3}, negative differential resistance, **(b)** design with $N_D = 1.7 \times 10^{13}$ cm^{-3}, positive differential resistance up to some 10 A/cm^2. Figure from [Lut09] © 2009 IEEE

positive voltage-area at the nn$^+$-junction, resulting in a positive branch of the post-avalanche characteristics.

In the simulated I–V-characteristic branches with positive and with negative differential resistance (NDR) occur at increased carrier density in avalanche, depending on the respective design [Hei05]. Figure 13.8 shows the post-avalanche behavior. Diode design (a) with very low background doping $N_D = 1.1 \times 10^{13}$cm^{-3} shows NDR already at 0.1 A/cm^2. Diode design (b) with increased doping shows as consequence a lower breakdown voltage, but then follows a branch with positive differential resistance.

For a further increased current density, a second voltage peak builds up at the nn$^+$-junction. Finally, at a very high avalanche current density, design (b) runs at the same line as (a), since at this condition the blocking capability is no longer determined by the doping, but only by the free carriers and the device thickness. Design (c) in Fig. 13.8 with thicker base has an extended branch of positive differential resistance.

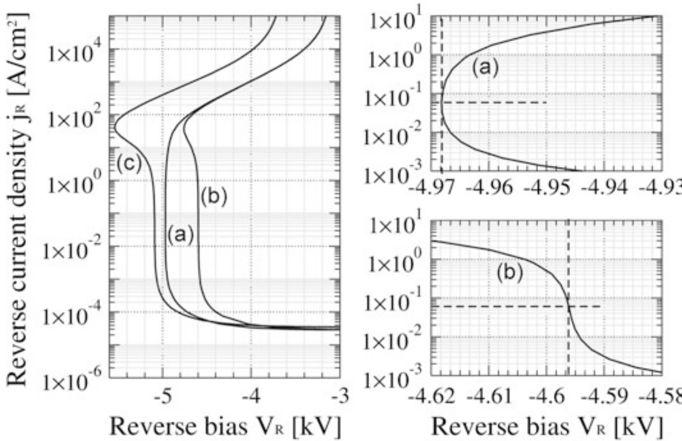


Fig. 13.8 Influence of base doping and base width to the static avalanche characteristics. **a** and **b** with w_B = 375 µm, and different doping of 1.1 × 10^{13} cm^{-3} (**a**) and 1.7 × 10^{13} cm^{-3} (**b**), **c** with 1.7 × 10^{13} cm^{-3} and wider w_B of 450 µm. Figure from [Lut09] © 2009 IEEE

The occurrence of branches with the NDR was first explained by Egawa [Ega66]. They are in conjunction with a hammock-like field shape, as to be seen in Fig. 13.7a, b for high current density. This shall be called Egawa-type field. Such fields and branches with NDR have the precondition that the electron current n_{av}, calculated with Eq. (13.9), is higher than N_D. This condition is fulfilled for low voltage devices only at very high current densities, but it may be reached at a moderate current density in high voltage devices with low background dopings. This is one of the reasons why high voltage devices are usually not avalanche-rated.

At the nn$^+$-junction, avalanche is triggered by electrons and we have to use the multiplication factors for electrons, which are much higher than that for holes, see Fig. 3.15. Therefore impact ionization will occur at the nn$^+$-junction already at a lower electric field. Impact ionization at the nn$^+$-junction will create electron-hole-pairs, the holes are flowing to the pn-junction and will increase the avalanche at the pn-junction. A positive feedback between impact ionization at the left and the right side will occur. Avalanche at an nn$^+$-junction has been described as a failure mechanism of devices in [Ega66, How70].

In Fig. 13.7, the difference between (a) and (b) is only the increased doping in (b). The range of positive differential resistance can be extended by a thicker base w_B. Most effective are buffer layers, in which a region of increased doping is arranged in front of the nn$^+$-layer. With special buffers, fields at the nn$^+$-junction can be limited and branches of negative differential resistance can be widely avoided. Details are given in [Fel06].

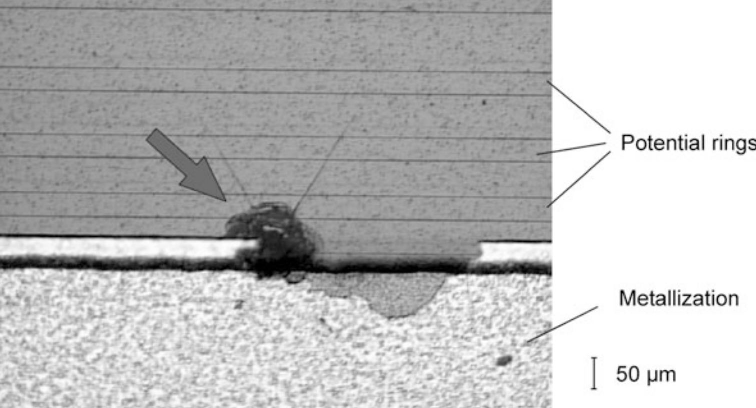


Fig. 13.9 1700-V diode destroyed by voltage

In devices with planar junction termination, the edge usually limits the maximal possible blocking capability.[1] Avalanche breakdown first occurs at the edge. Therefore current densities high enough to lead to Egawa-type fields can occur locally at the edge. For a failure by overvoltage it is typical that the edge of the device is included in the destroyed area. Figure 13.9 shows the failure picture of a 1700 V diode. The device has a planar junction termination with potential rings similar to Fig. 4.24. Three potential rings can be seen in the upper part of the figure. The point of destruction is located between the p-anode layer and the first potential ring. This is one of the positions of the highest electric field, as it is marked in Fig. 4.24.

The occurrence of such a failure position indicates that the failure was caused due to voltage. However Fig. 13.9 does not allow a clear decision, whether an overvoltage above the rated voltage of the device was applied, or the particular device had a weak point induced in the manufacturing process. A failure picture as in Fig. 13.9 is only then occurring, if no high current was flowing across the point of destruction.

A picture of a destroyed semiconductor, in which after failure a high current was flowing, is shown in Fig. 13.10. Part of the edge and a big part of the active area are evaporated. If such a picture occurs, one can presume that the destruction occurred at first at the edge, and then it propagated towards the bond-wires.

However the cracks in the crystal lattice are not typical. These cracks indicate a local hot spot in a small, point-like position. Such failure pictures can also be found at strong dynamic avalanche (dynamic avalanche of the third degree, see Sect. 13.4.2). Therefore this failure picture is not unequivocal.

[1]Exception devices with cells (MOSFETS, IGBTs, MPS-diodes) in which the cell geometry is adjusted that the avalanche breakdown occurs first below the cells and not at the edge, see above.

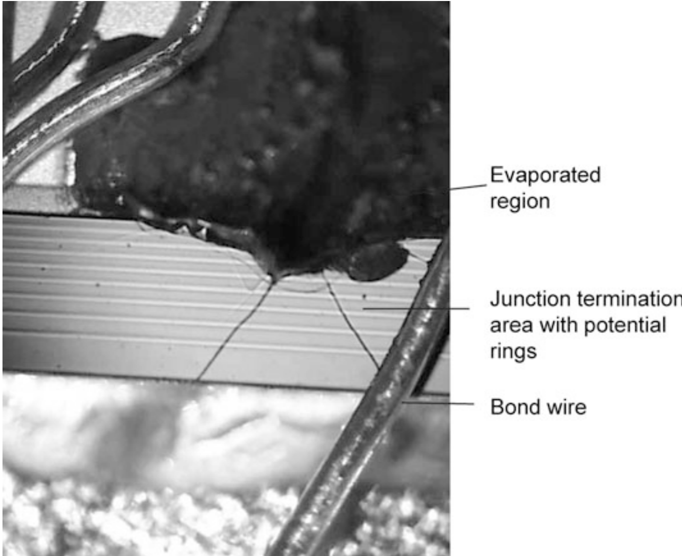


Fig. 13.10 3.3-kV diode possibly destroyed by overvoltage

13.4 Dynamic Avalanche

13.4.1 Dynamic Avalanche in Bipolar Devices

During switching of all bipolar devices the increase of the voltage occurs at an instant, at which a large part of the stored carriers, which have conducted the forward current before, is still present in the device. This stored charge is partially removed during the voltage increase, and it flows as hole current through the space-charge region.

Figure 13.11 shows the process in a simplified way. The pn-junction at the position $x = 0$ represents the blocking pn-junction of a bipolar device. Between the junction and w_{SC} the space-charge region has extended, for supporting the applied voltage. Between w_{SC} and the end of the lowly doped layer exists a plasma zone, in which $n \approx p$ holds. The effects at the right side shall be neglected in this first approximation. At this position will be either an nn^+-junction of a diode, or the collector layer of an IGBT, or the anode layer of a GTO thyristor etc. As long as not very hard switching conditions are applied, no space-charge is build up at this position.

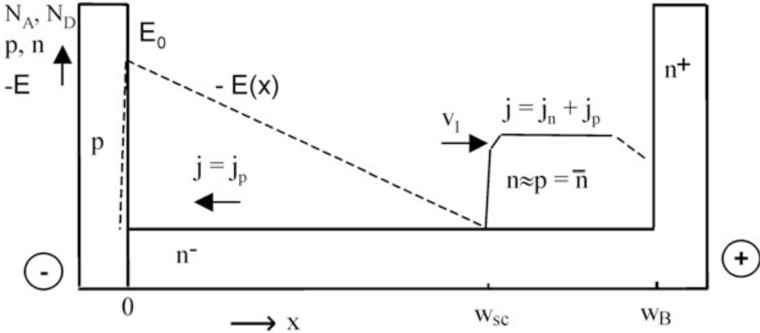


Fig. 13.11 Bipolar semiconductor device during the turn-off process

Through the space-charge the current flows as hole current, $j = j_p$. The density of holes p can be calculated from the current density at this instant:

$$p = \frac{j}{q \cdot v_{sat(p)}} \tag{13.10}$$

In this equation $v_{sat(p)}$ is the saturation drift velocity of holes under the condition of high fields, it amounts in silicon to approximately 1×10^7 cm/s and is close to the saturation drift velocity of electrons $v_{sat(n)}$. A current density j of 100 A/cm^2 leads to $p = 8.2 \times 10^{13}$ cm^{-3}, which is already in the order of the background doping of a bipolar 1200 V device. The hole density p can no longer be neglected.

Holes have the same polarity as the positively charged ionized donors, hence their density now adds to the background doping to an effective doping N_{eff}:

$$N_{eff} = N_D + p \tag{13.11}$$

With the Poisson-equation, N_{eff} determines the gradient of the electric field

$$\frac{dE}{dx} = \frac{q}{\varepsilon}(N_D + p) \tag{13.12}$$

E/dx is increased. With this, the field shape is steeper, E_0 is increased and the voltage, which drops across the space-charge of width w_{SC}, is increased in the first instance. However E_0 can reach to a maximum the value of the avalanche field strength $E_C \cdot E_C$ will now be reached at an applied voltage far below the specified rated blocking voltage of the device and avalanche will set in. This process, which is now dominated by free carriers, is called dynamic avalanche.

This process occurs during turn-off of diodes, GTOs and IGBTs, for further details however the specific peculiarities of the respective device and their physics must be considered.

13.4.2 Dynamic Avalanche in Fast Diodes

In the state which is drawn in Fig. 13.11, the shape of the electric field is approximately triangular, since no space-charge can penetrate into the plasma layer. For a triangular field we can use the relation (3.84) between the avalanche breakdown voltage V_{BD} and doping

$$V_{BD} = \frac{1}{2} \cdot \left(\frac{8}{B}\right)^{\frac{1}{4}} \cdot \left(\frac{q \cdot N_{eff}}{\varepsilon}\right)^{-\frac{3}{4}} \tag{13.13}$$

where for the ionization rates the proposal of Shields [Shi59] and Fulop [Ful67] is used with $B = 2.1 \times 10^{-35}$ cm^{6}V^{-7} and $n = 7$ at room temperature. The derivation of this equation can be found in Chap. 3, Eqs. (3.75)–(3.84).

If now (13.10) and (13.11) are inserted in (13.13), we obtain a relation between the onset of dynamic avalanche and the current density in the space-charge region, as shown in Fig. 13.12. For the background doping N_D the typical values for a 1700 V device and for a 3300 V device are used, $N_D = 4.3 \times 10^{13}$ cm^{-3} and 1.7×10^{13} cm^{-3}, respectively.

The derivation of Eq. (13.13), however, assumed that at the boarders of the space charge the minority carrier currents enter from adjacent neutral regions, as it is the case for static avalanche breakdown, see Sect. 3.3. In fact, the space charge now is in touch with a plasma zone, see Fig. 13.11. For the share of electron and hole current in the plasma holds Eq. (5.85), the hole current j_p given in this equation, which is significantly higher, penetrates the space charge. As breakdown condition the ionization integral $\int_0^w \alpha(E(x))\,dx$ in Eq. (3.71) must not be unity, it is already sufficient if it is reaching $\int_0^w \alpha(E(x))\,dx = 1 - \mu_p/(\mu_n + \mu_p)$ [Dom03]. The further derivation is analogously to Eqs. (3.75)–(3.79), it results [Bab11]

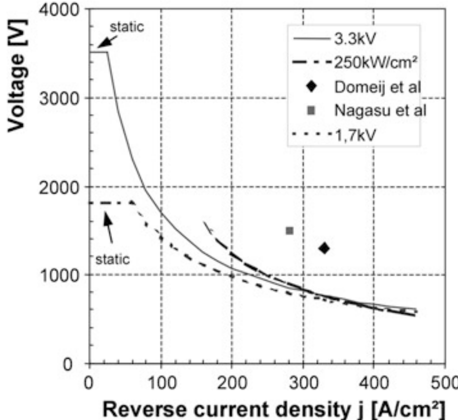


Fig. 13.12 Avalanche onset voltage depending on current density in reverse direction according Eq. (13.13), former assumed limit of 250 kW/cm^2, reported operation points of 3.3-kV power diodes at max. power density of former work Domeij et al. [Dom99] and Nagasu et al. [Nag98]. Reprinted from [Lut03] with permission from Elsevier

$$V_{av,dyn} = \frac{1}{2} \cdot \left(\frac{8}{B}\right)^{\frac{1}{4}} \cdot \left(\frac{q \cdot N_{eff}}{\varepsilon}\right)^{-\frac{3}{4}} \cdot \left(\frac{\mu_n}{\mu_n + \mu_p}\right)^{\frac{1}{4}} \tag{13.13a}$$

using for the ionization rates the proposal of Shields and Fulop [Shi59], as done in Eq. (13.13). The movement of the plasma front is neglected. The comparison with Eq. (13.13) using the in Si approximatively given relation $\mu_n \approx 3 \cdot \mu_p$ leads to $V_{avdyn} = 0.93 \cdot V_{BD}$. The onset of dynamic avalanche is shifted to a value somewhat lower compared to Fig. 13.12.

From Fig. 13.12 one can recognize that for a device with high static blocking capability the avalanche onset voltage decreases strongly. In a 3.3 kV diode, dynamic avalanche sets in already at a reverse current density of 30 A/cm^2. For a reverse current density of 200 A/cm^2 the limit for the onset of dynamic avalanche has decreased down to 1050 V, and it is not much above the limit of a 1700 V device. There are only small differences between the two diodes voltage ratings, because the second term in (13.11) representing the free charge carriers is dominating.

13.4.2.1 Dynamic Avalanche of the First Degree

Close to the onset of dynamic avalanche, failures have been detected and assumed to be an unavoidable phenomenon of semiconductor physics [Por94]. Even a “silicon limit” at a power density $V \cdot I/A = 250$ kW/cm^2 was sometimes used [Sit02]. The line for 250 kW/cm^2 shown in Fig. 13.12 is close to the calculations with Eq. (13.13), the onset of dynamic avalanche at the pn-junction.

However Schlangenotto [Sco89b] published another point of view. Moderate dynamic avalanche, dynamic avalanche of the first degree, should not be critical. Dynamic avalanche generates electrons in the space-charge which counteract the positive charge of holes, and then holds

$$N_{eff} = N_D + p - n_{av} \tag{13.14}$$

The increased hole density is partially compensated, and this mechanism should be self-stabilizing. This was confirmed by diode designs bearing a significant amount of dynamic avalanche current [Lut97]. For 3.3 kV diodes Nagasu et al. [Nag98] and Domeij et al. [Dom99] reported measurements with 1.7 times higher voltages than the alleged limit. These results are also shown in Fig. 13.12.

The power loss density gives an estimation of the avalanche intensity, but it gives no physical explanation of a failure limit, since the high power density occurs only during some 10 ns, while the dissipated energy is low and the temperature increase, if it is a one-time effect, is negligible.

However, for this to hold true it is a precondition that the device has no weak points in its design. In [Nag98, Tom96] it was shown that the edge of the active area is of great importance. In diodes with the common planar field limiting structures,

the anode area is smaller than the cathode area. At the edge of the anode additional current contributions from the n^+-region are occurring during forward conduction, as shown schematically in Fig. 13.13. The current density at the edge of the anode is increasing, and in device simulation one can observe a current filament already at the beginning of turn-off period. The implementation of a resistive zone at the edge, as shown in Fig. 13.14, is reducing this weak point. The p-layer is extended by a length R below the passivation layer. The anode-side p-layers are only moderately doped in soft recovery diodes, therefore the region R acts as a pre-resistor for a current at the edge of the p-layer, and the current density at the edge will be reduced. Such a structure at the edge was called "HiRC"-structure (High Reverse-recovery Capability) in [Mor00].

13.4.2.2 Dynamic Avalanche of the Second Degree

With an increasing current density, dynamic avalanche leads to filamentation of the current due to space-charge effects as reported by Oetjen et al. [Oet00]. Figure 13.15 shows a static simulation of the electric field distribution and I–V characteristic of a p^+np^+ structure. This figure represents the process of Fig. 13.11, however for an increased current density. The plasma layer of Fig. 13.11 is replaced by a p^+-layer which injects holes. Figure 13.15a displays the electric field at the current densities 500 and 1500 A/cm^2. The field has reached the avalanche field strength level at the pn-junction, in this figure at x = 8 μm. The field shape for j = 500 A/cm^2 is approximately triangular. Avalanche generates electron-hole-pairs in the region with high field strength. However the generation is not local, since it

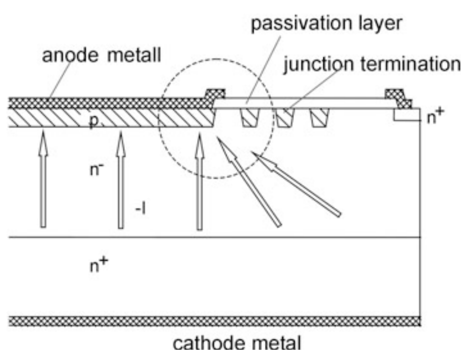


Fig. 13.13 Planar junction termination of a diode with increased current density at the edge of the active area. Reprinted from [Lut03] with permission from Elsevier

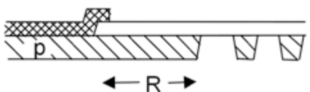


Fig. 13.14 Resistive layer for reduction of the current density at the edge of the anode. Reprinted from [Lut03] with permission from Elsevier

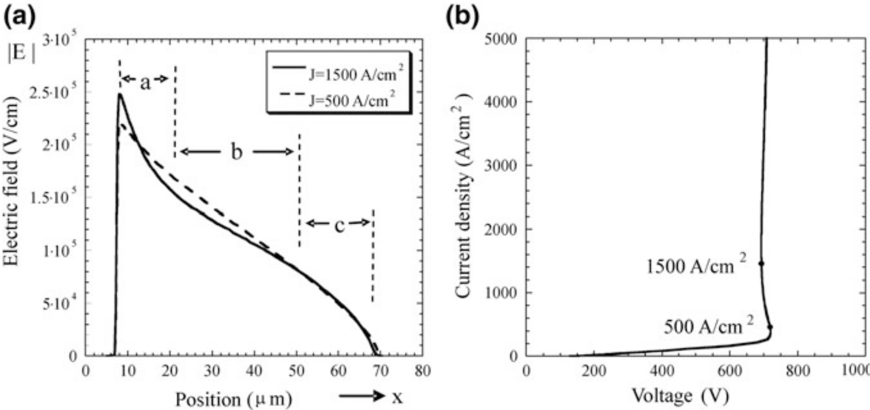


Fig. 13.15 Illustration for dynamic avalanche at an increased current density. Field shape (left hand side), I–V characteristic (right hand side). Reprinted from [Lut03] with permission from Elsevier

happens within some length in the x-direction which is necessary for carrier acceleration. The holes are flowing to the left hand side, the electrons are flowing to the right hand side. Close to the pn-junction - region a - the hole density is increased by the holes generated by avalanche. At the pn-junction holds

$$\frac{dE}{dx} = \frac{q}{\varepsilon}(N_D + p + p_{av}) \tag{13.15}$$

With this, dE/dx at the pn-junction has become very high. This is drawn for j = 1500 A/cm^2. Further away from the pn-junction, in region b, the electrons n_{av} generated by dynamic avalanche are flowing to the right hand side. There are also some generated holes p_{av}, but their density is decreasing with distance to the pn-junction. It can be summarized for region b

$$\frac{dE}{dx} = \frac{q}{\varepsilon}(N_D + p + p_{av} - n_{av}) \tag{13.16}$$

In region b a partial compensation of generated electrons and the flowing holes occurs, p_{av} is decreasing to the right hand side, while n_{av} is increasing. Hence, the gradient of the electric field $d\mathsf{E}/dx$ becomes flat. At the boarder to the plasma layer $\mathsf{E} = 0$ must hold. Therefore in region c the gradient $d\mathsf{E}/dx$ must increase again. The field in regions a to c will form a typical bowed shape.

The voltage correlates to the area under $\mathsf{E}(x)$ which is slightly lower for the higher current density (see Fig. 13.15b). The I–V characteristic has a region of weakly negative differential resistance. Such a characteristic splits a homogenously distributed current to areas with lower current density and in filaments with high

current density. The type of characteristic in Fig. 13.15b is described as S-shaped in [Wak95], which can lead to the formation of stable current filaments.

Device simulation of such events shows a current density of 1000 – 2000 A/cm^2 in the filaments. Nevertheless, there are counteracting mechanisms and destruction should be avoidable for the following reasons:

– The differential resistance is weakly negative, so that a further increase of current increases the voltage again. Therefore, the current density in the filaments is limited.
– The temperature inside the filament will increase leading to a reduction of impact ionization in the area of the filament, this counteracts the filamentation.
– The high local current density in a filament quickly removes the stored carriers locally from the plasma layer, counteracting the driving force for dynamic avalanche.

Therefore a state of rising, expiring, jumping or moving filaments is to be expected, but a diode should still be capable to withstand this effect. It has to be taken care that the edge of the active area does not lead to a locally fixed filament, which will be the case for the structure in Fig. 13.13. Moving and jumping filaments are to be seen in 2D-device simulations [Hei06, Nie05]. It must be mentioned that 2D-simulations are limited, because filaments are a 3D-effect. However such simulations are enormously time consuming and have not been possible up to now. In real devices, the edge of the active area is a strong inhomogeneity, and it triggers the first filaments [Hei07]. For such a geometry including the edge of the active area, a 2D simulation is capable of showing the main effects.

Most simulations in this field have been done under isothermal conditions that means without consideration of any temperature increase in filaments. However, the avalanche coefficients are strongly temperature dependent. The first temperature dependent simulations of such problems showed an influence of temperature, filaments building up and decreasing at different speeds, compared to isothermal simulations. However the qualitative behavior was similar [Hei07].

13.4.2.3 Dynamic Avalanche of the Third Degree

Further increase of the stress by dynamic avalanche leads to the situation that a field at the nn$^+$-junction occurs while there is still dynamic avalanche at the pn-junction. The effects are shown in Fig. 13.16 in a simplified way. There is still remaining plasma, and the electric field between the plasma and the pn-junction is bowed as in dynamic avalanche of the second degree. Between the plasma and the nn$^+$-junction a second electric field builds up, whose peak is at the nn$^+$-junction and whose gradient is inverted compared to the left space-charge zone.

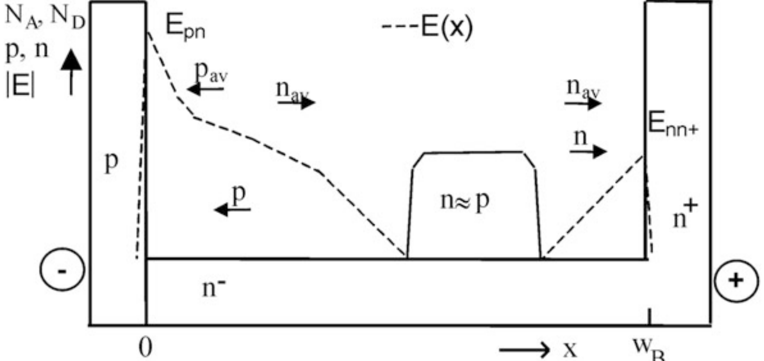


Fig. 13.16 Effects at strong dynamic avalanche. Figure adapted from [Lut09] © 2009 IEEE

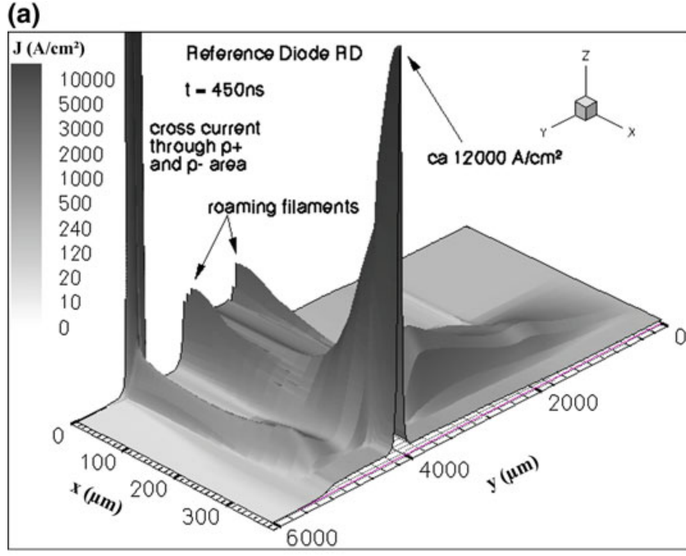
This inverted gradient of the electric field can only occur, if the density of free electrons is so high that it overcompensates the positively charged donors of the background doping:

$$\left|\frac{dE}{dx}\right| = \frac{q}{\varepsilon}(n + n_{av} - N_D) \tag{13.17}$$

If n_{av} is increasing, the absolute value of $d\mathsf{E}/dx$ will increase, and finally impact ionization sets on at the nn$^+$ junction also. This is usually initiated locally in a region where a high current density filament at the pn-junction has formed. A double-sided dynamic avalanche is given at this position.

Figure 13.17 shows results of a simulation of a diode under such conditions [Hei08]. In Fig. 13.17a the current density is shown, between y = 3600 and 4400 µm a wide filament at the pn-junction has formed, at the nn$^+$-junction occurs a narrow high-current filament at y = 4000 µm. In Fig. 13.17b the electric field is shown for the same instant. At the nn$^+$-junction a second electric field appears with a high field peak at y = 4000 µm.

This hammock-shaped electric field has already occurred as an Egawa-type field [Ega66] in Sect. 13.3. There are also similarities to the field at the nn$^+$-junction at second breakdown in bipolar transistors. Applying Wachutka's model for the destruction limit of GTO-thyristors [Wac91] to fast recovery diodes using the boundary conditions in [Ben67] leads to the conclusion, that dynamic impact ionization at the pn-junction should be stable, whereas impact ionization at the nn$^+$-junction is highly unstable. It is sufficient, that the ionization integral $\int_0^w \alpha(E(x))\,dx$ (see Eq. 3.67) reaches a value of 0.3 to get to an instability mode. The current density in a filament will increase with a time constant in the range of some ns [Dom03]. Impact ionization at the nn$^+$-junction is triggered by electrons and will occur at lower electric field strength because of the higher ionization rates of electrons. The rapid increase of the local current density is assumed here to be the cause of destruction.



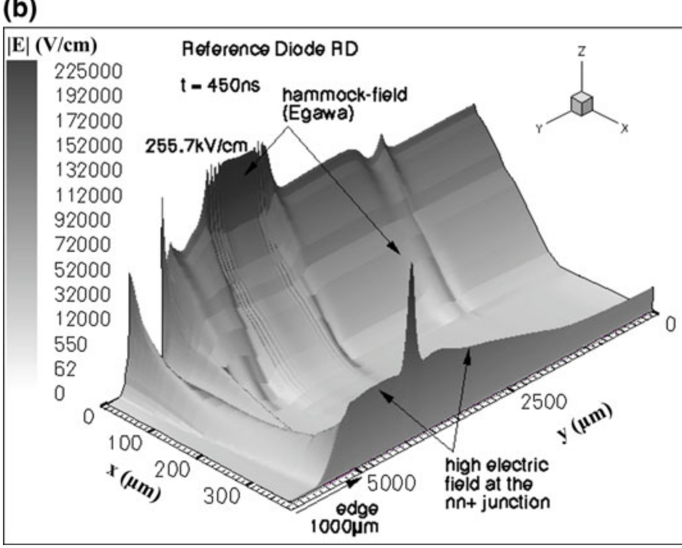


Fig. 13.17 Effects during strong dynamic avalanche: Current filaments and electric field in simulation of the turn-off of a 3.3 kV diode. Conditions V_{bat} = 1800 V, di/dt = 1600 A/µs, L_{par} = 1.125 µH. **a** Current density distribution at strong dynamic avalanche, current tube at the pn-junction, rising of a high current filament at the nn$^+$-junction at x = 4000 µm. **b** Electric field distribution, rising of an Egawa-type field at x = 4000 µm. Figure from [Lut09] © 2009 IEEE

In recent work, it has been shown by numerical simulation that the plasma layer expands to the nn$^+$-junction in the vicinity of a cathode-side filament. This behaviour differs from anode-side filaments and results from the velocity saturation of

the electrons and holes in the high-field region [Bab09]. The reduced depletion layer in the vicinity of the filament inhibits the lateral movement of the filament and, therefore, leads to strong local heating. This generates a destructive thermal filament.

The assumption that an Egawa-type field above a certain limit of 100 kV/cm at the nn^+-junction (Fig. 13.16 right hand side) leads to the destruction of the diode, could explain an experimentally found destruction limit. Figure 13.18 shows the failure of a 3.3 kV rated diode under the condition of very fast commutation *di*/*dt*. The reverse-recovery current peak I_{RRM} of 360 A represents a reverse current density of 400 A/cm^2. The voltage across the diode is rising very fast, since in this experimental setup an additional capacitor of 22 nF was built-in between the gate and emitter of every switching IGBT. Already 200 ns after I_{RRM} the voltage has climbed-up to 2000 V. The diode is destroyed shortly after the voltage peak.

If the employed IGBT withstands the short-circuit stress, which occurs after the destruction of the diode, and if it turns off the current successfully, one can find a characteristic failure picture as shown in Fig. 13.19. At one point of the active area a small molten channel is found. In the case of dynamic avalanche of the third degree, cracks in angles of 60° were observed. The failure picture corresponds to the destruction of a 111-oriented silicon wafer by a point-shaped stress. This is a hint for a current filament of very high current density and very high temperature in a small area.

Such a failure limit was reproduced for devices of the same type of different production lots. Figure 13.20 shows the measurement results (black dots). The *x*-axis in Fig. 13.20 corresponds to the current density at a reverse-recovery current peak I_{RRM}, which can be adjusted by variation of *di*/*dt*. The *y*-axis shows the value of the voltage peak V_{pk} (see Fig. 13.18) which occurred before failure. It can be

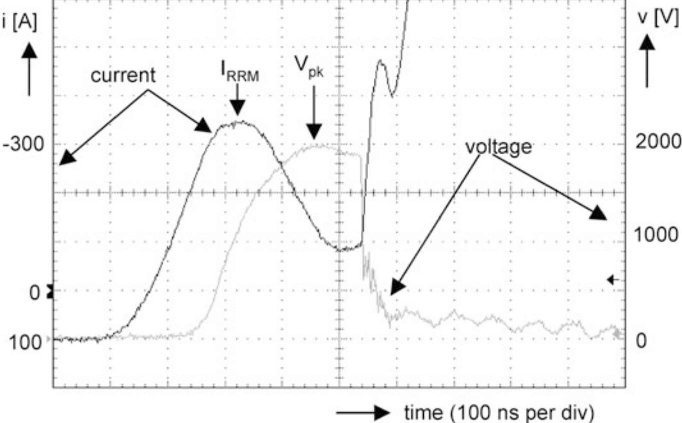


Fig. 13.18 Failure of a 3.3 kV diode under extremely strong dynamic avalanche conditions. Reprinted from [Lut03] with permission from Elsevier

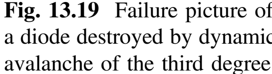


Fig. 13.19 Failure picture of a diode destroyed by dynamic avalanche of the third degree

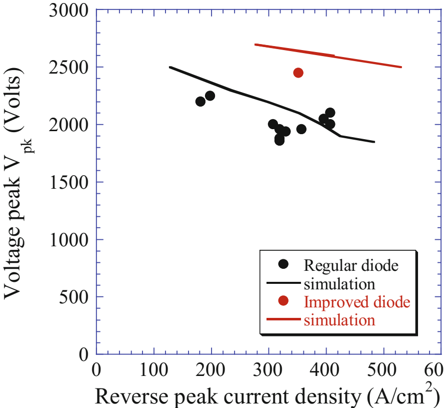


Fig. 13.20 Experimental failure limit and simulated failure limit. Reprinted from [Lut03] with permission from Elsevier

recognized that there is a limit for the voltage, and this limit depends only weakly on the current density in the measured interval.

A device simulation with AVANT Medici resulted in the strait line shown in Fig. 13.20. The condition for failure was taken that a second electric field builds up at the nn$^+$-junction, and that a failure occurs if this field grows up to 100 kV/cm.

The simulation predicts that an Egawa-type field occurs at higher voltages if the width of the n$^-$-layer of the diode is increased (dotted line in Fig. 13.20). This was confirmed by an experiment for a diode with a 50 µm thicker middle layer. However, the experimental found failure is at lower voltage than the simulated limit. Nevertheless Fig. 13.20 shows that with the selection of the necessary thickness, the ruggedness can be improved.

More recent results showed that freewheeling diodes even up to the range of 3300 V could achieve a very high capability to withstand dynamic avalanche [Rah04]. Figure 13.21 shows a measurement with very high stress on the diode by the high battery voltage $V_{bat} = V_{DC}$, high di/dt and especially a high inductance of 2.4 µH. Already shortly after the voltage increase, strong dynamic avalanche

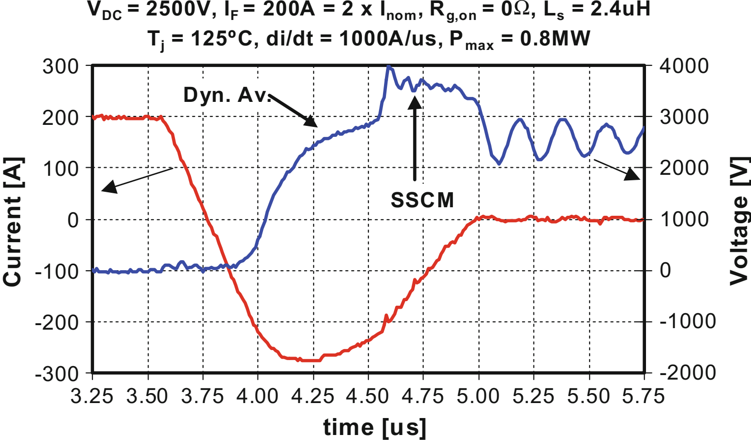


Fig. 13.21 3.3 kV diode with strong dynamic avalanche and SSCM-Mode [Rah04] © 2004 IEEE

occurs, and above a voltage of 1500 V at $t = 4.1$ μs a flattening in the increasing voltage is observed, the reverse-recovery current peak is widened and the voltage increases very slowly. After strong dynamic avalanche, at $t = 4.55$ μs the internal plasma is suddenly exhausted. The hole current, which before this instant is fed by the internal plasma, suddenly disappears, thus, the reason for dynamic avalanche is no longer there. Now the voltage climbs up steeply, but it is limited by the diode itself. This limit at approx. 3700 V is close to the static avalanche breakdown voltage of the diode.

However in the current waveform in Fig. 13.21, at this point in time occurs no snap-off, but only a small dip. Thereafter follows a current generated in static avalanche. This mode was called "switching self-clamping mode" (SSCM) [Rah04], where the diode clamps the voltage peak.

In the SSCM mode holds [Hei05]:

$$\frac{di}{dt} = \frac{V_{SSCM} - V_{bat}}{L_{par}} \tag{13.18}$$

where V_{SSCM} is a voltage close below the static breakdown voltage V_{BD}. The capability of the diode to transit from the mode of dynamic avalanche into static avalanche is a high progress in ruggedness. According to [Rah04] the design of this diode observed the necessary measures to achieve a high ruggedness: The p^+ anode layer was sufficiently high doped. The edge of the active area is designed in a way that local current crowding is avoided. Especially, the nn^+-junction of this diode has a very shallow gradient. If at the nn^+-junction the doping increases slightly, N_D is increased, according Eq. (13.17) more electrons would be necessary for inversion of the gradient of the electric field, and the danger that an Egawa-type field can arise is lessened.

13.4.3 Diode Structures with High Dynamic Avalanche Capability

Increased doping at the n^+-side by a buffer reduces Egawa-type fields, since the increased density of positively charged donor ions compensates part of the arriving electrons. However most efficient are structures which inject holes for compensation of avalanche-generated electrons.

One of these structures is the “Field Charge Extraction” (FCE) structure [Kop05]. A part of the cathode area contains a p^+-layer, as shown in Fig. 13.22 at the cathode side. At the anode side, a He^{++}-implantation was applied to reduce locally the lifetime and to adjust for soft recovery behavior, see Chap. 5. Soft recovery can be achieved with different methods. The effect of the FCE structure is at the cathode side: If a space-charge builds up at the cathode side, the p^+-layer will inject holes. The injected holes compensate the electrons generated by dynamic avalanche.

The disadvantage of this structure is that part of the cathode area is lost for injection of carriers into the internal plasma. Therefore, the forward voltage drop of the diode will be increased.

The “Controlled Injection of Backside Holes” (CIBH) structure avoids this disadvantage. It contains floating p-layers in front of the cathode n^+-layer [Chm06]. The structure is shown in Fig. 13.23.

A continuous p-layer at the backside would realize a four-layer diode, this structure was discussed in [Mou88]. Such a four-layer diode acts in a similar way as a thyristor. It conducts in the forward direction after the cathode-side pn-junction is overcome by breakover-triggering. This leads to an additional voltage peak at turn-on of the diode and makes the structure unusable. Therefore in the CIBH diode this p-layer is interrupted by areas of the distance c, which form a resistor parallel to the additional junctions J2 and J3. The thickness of the p layer between J2 and J3 is very low. Both sides of the junction J3 are highly doped zones, which results in the onset of the avalanche breakdown at a small reverse bias like in an avalanche diode. Correspondingly the CIBH diode can be analyzed as a pin diode with an integrated avalanche diode, which is accompanied with a parallel resistor. The distance c must be wide enough to avoid a deterioration of the turn-on behavior, and it must be small enough to avoid high fields at the cathode side.

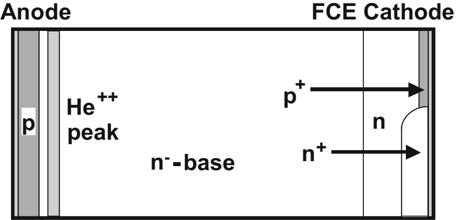


Fig. 13.22 Field Charge Extraction (FCE) diode, drawn after [Kop05] © 2005 IEEE

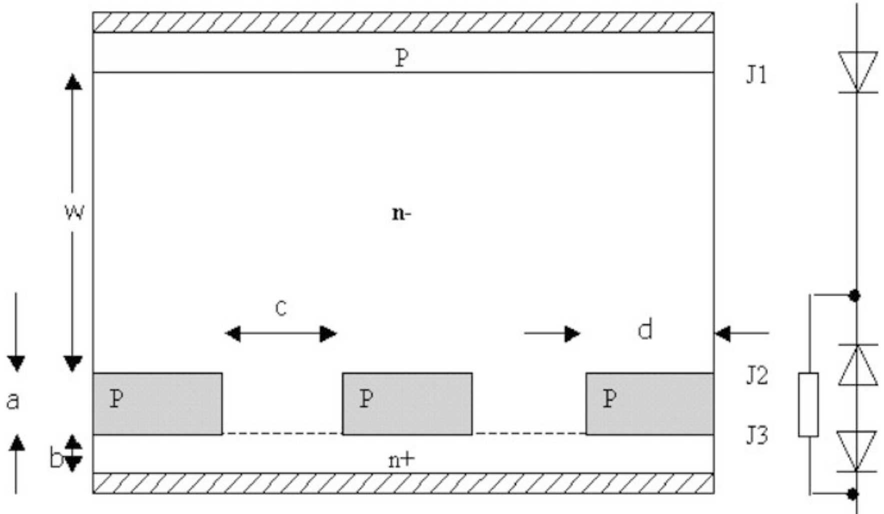


Fig. 13.23 "Controlled Injection of Backside Holes" (CIBH) diode [Chm06] © 2006 IEEE

The effect of the suppression of electric fields is shown in Fig. 13.24. For comparison, Fig. 13.24a shows the electric field distribution during reverse-recovery of a reference diode. The formation of the electric field peak at the nn^+ junction starts at t = 400 ns and leads to a critical Egawa-type field distribution after t = 400 ns, as described before.

Under the same switching condition the CIBH diode presents a completely different transient electric field distribution in the diode when compared to the reference diode, as illustrated in Fig. 13.24b. The situation at the backside is obviously improved. After the onset of the controlled avalanche at junction J3, the voltage drop at the backside is clamped at the breakdown voltage of the junction J3. Avalanche at J3 injects as much holes as necessary to compensate the electrons. The evolution of a second peak of electric field strength, which is correlated to a space-charge region at the cathode side, is successfully suppressed.

Since Egawa-type fields can be effectively avoided, the CIBH diode has an extremely high ruggedness in dynamic avalanche. Figure 13.25 shows the turn-off of two 3.3 kV CIBH diodes under extreme stress. The maximal power density at turn-off is 2.5 MW/cm^2 – a factor of 10 above formerly assumed limits!

The CIBH diode has additional advantages in the turn-off behavior: During plasma removal, the plasma does not detach the cathode zone, but remains connected to the cathode, see Fig. 13.24b. In contrast to the usual process at reverse-recovery as described in Fig. 5.26, the plasma in the CIBH diode is removed only from the anode side.

This effect occurs, since early during reverse-recovery, junction J3 goes into the avalanche mode and injects a current $j_{p,ava}$. Then for the movement of the cathode-side plasma layer front holds [Bab08]

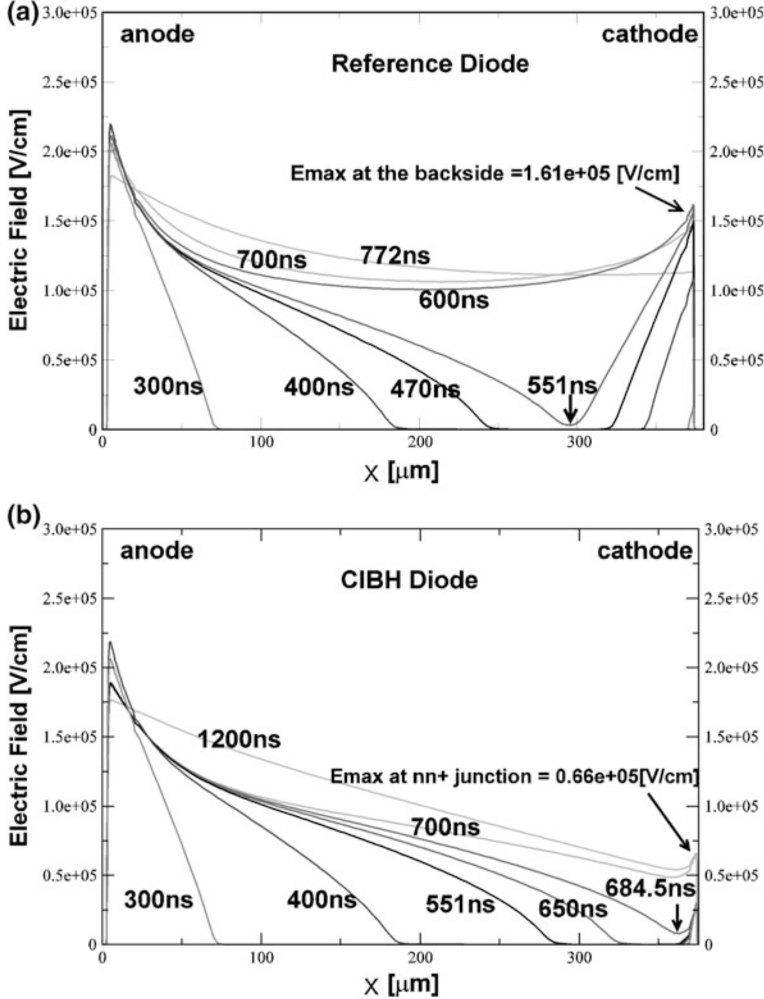


Fig. 13.24 Simulated electric field distribution of a reference diode (**a**) and of the CIBH diode (**b**) at very high stress in dynamic avalanche $T = 300$ K, $J_F = 100$ A/cm^2, $V_{bat} = 2500$ V, $di/dt = 2000$ A/μs cm^2, $L_{par} = 1.25$ μH. Figure from [Chm06] © 2006 IEEE

$$v_R = \frac{j_p - j_{p,ava}}{q \cdot \bar{n}} \tag{13.19}$$

The absolute value of v_R is reduced, and if $j_{p,ava} = j_p$, v_R becomes 0. The plasma stays connected to the nn$^+$-junction. In Eq. (5.94) v_R is equal to 0 and we obtain $w_x = w_B$. Therefore the whole base width w_B is supporting the voltage V_{sn}, and in Eq. (5.95) $w_x = w_B$ holds.

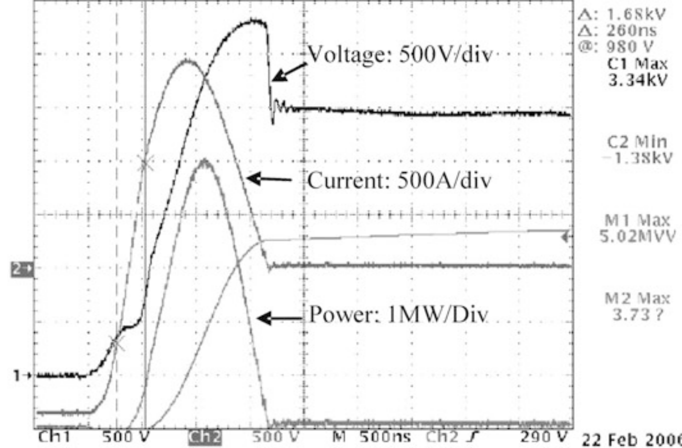


Fig. 13.25 Experimental ruggedness measurements of two CIBH diodes in a module at 5.5× nominal current and $T = 400$ K, $V_{bat} = 2500$ V, $di/dt = 6500$ A/µs, $L_{par} = 0.75$ µH, voltage (CH1), current (CH2), dissipated power (Math1). Figure from [Chm06] © 2006 IEEE

If now the space-charge region reaches w_B – this is the case at turn off under conditions of low forward current, high parasitic inductance and high voltage, where usually freewheeling diodes show snappy recovery – the p-layers inject additional holes. If the applied voltage is further increased, the reverse-recovery is further improved. This was introduced as DSDM (Dynamic Self Damping Mode) [Fel08].

The forward voltage drop of the CIBH diode is not distinctly increased compared to another diode of comparable thickness and stored charge, since during forward conduction we have a parallel connection of a triggered thyristor and a diode. A possible drawback could be a loss in the static breakdown voltage V_{BD} because of the hole injection into the electric field. By adjusting the process parameters even at high p-doses and high p area ratios the loss in the static breakdown voltage can be avoided while the softness is improved. The loss of breakdown voltage is in maximum 7% for diodes with a high implantation dose [Fel08].

The CIBH diode was introduced in the market in a 3.3 kV Infineon module [Bie08]. The CIBH diode shows that power devices can be designed in a way that they can withstand high stress in dynamic avalanche. A review on dynamic avalanche in high-voltage devices is given in [Lut09]. However manufacturers still restrict the allowed reverse current I_{RRM} with safe-operation area (SOA) diagrams. This SOA diagrams are usual for devices with blocking voltages of 3.3 kV and above, and have a shape similar to Fig. 13.12. Such diagrams forbid the application of devices in dynamic avalanche.

We also conclude from Fig. 13.12: The higher the rated voltage of the device, the more the allowed reverse current density must be restricted if one wants to avoid

dynamic avalanche. On the other hand, the stored charge Q_{RR} in a device increases with the base width of a device at a power of two — see Chap. 5, Eq. (5.64). With increased Q_{RR} also I_{RRM} is increasing. To keep I_{RRM} low, the current slope di/dt must be kept small. However with a small *di/dt* during turn-on of a transistor, the turn-on losses in the transistor increase. Therefore, the required restriction of I_{RRM} restricts the possibility to reduce the switching losses in the application.

It was found that the higher the rated voltage, the lower the current densities at which formation of filaments and current tubefilaments occur [Nie04]. Therefore, in spite of the progress, further research is necessary to understand the behavior in power devices at the border of the safe operation area, especially for high voltage devices.

Similar effects of negative differential resistance and moving filaments have been found in low-voltage ESD protection devices at very high current densities [Pog03]. The occurrence of Egawa-type electric fields with tow field peaks is also found in 60-V CMOS devices at high current densities [Bab10]. Finally, in the Section on cosmic ray failures Egawa-type fields have been found as final destruction mechanism, see Sect. 12.8.6.

13.4.4 Turn-off of Over-Current and Dynamic Avalanche in IGBTs

The turn-off of over-current is a very critical operation point of an IGBT. In difference to short circuit, the turn-off process is executed while the device is highly flooded with free carriers. The first IGBT generations for 3.3 kV restricted the turn-off capability of over-current to two-time the rated current [Hie97]. A higher current was not allowed to be turned off. If a higher current occurs in these IGBTs, the driver has to wait until the IGBT is in the mode of saturation current, while the voltage increases and the IGBT then is turned off in the short circuit mode. Note that in the short circuit mode 5–6 times of the rated current are switched off without problems.

During turn-off of an over-current, the channel is first turned-off. Hence, the electron current flowing via the channel becomes extinct. The total current must flow for a short instant completely as hole current. Figure 13.26 shows the process for example of an NPT IGBT.

The hole current, fed by the remaining plasma, flows across the n^--layer, in which an electric field has built up. The density of free holes adds to the background doping. The gradient of the electric field gets steeper, as already treated with Eqs. (13.10)–(13.12). The blocking capability is therefore reduced as given in Eq. (13.13). Dynamic avalanche now generates electron-hole pairs in the region close to the blocking pn-junction. The holes are flowing to the left side in Fig. 13.26, while the electrons flow to the right side. The hole current, increased by dynamic avalanche, must flow through the p-well and via the resistor R_S (Fig. 10.2b).

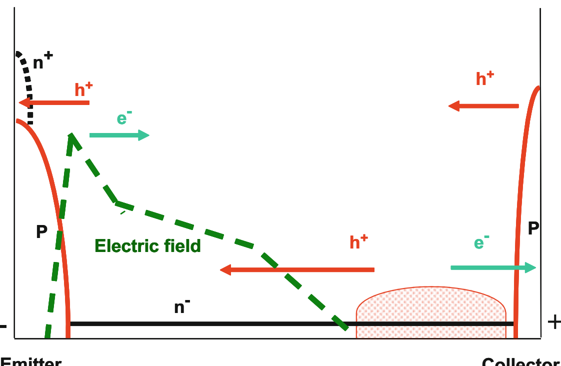


Fig. 13.26 NPT-IGBT at turn-off of an over-current and occurrence of dynamic avalanche

In this operation mode the hole density is highest and the danger of turn-on of the parasitic npn-transistor and latching of the IGBT is greatest. If R_S is low enough, the IGBT will successfully withstand such conditions.

The electrons generated close to the pn-junction are flowing to the right hand side, and they compensate the hole current. An electric field will rise as drawn in Fig. 13.26 for a strong dynamic avalanche mode. This field has close similarity to the S-shaped field in dynamic avalanche of the second degree, and the arguments given in context of Eqs. (13.15) and (13.16) hold respectively.

An electric field shaped as in Fig. 13.26 leads to a weakly negative differential resistance in the I–V characteristics, and an investigation of the effect using device simulation [Ros02] showed the occurrence of current tubes, which arise in certain regions and which jump to neighboring cells. This process also is similar to dynamic avalanche of the second degree.

However a fundamental difference to a diode is that on the right side of the plasma a p-layer exists, and the given collector-side pn-junction is forward biased. This layer injects holes, which compensate the electrons arriving from exhaustion of the remaining plasma and from dynamic avalanche. This opposes to a removal of the remaining plasma from the collector side. Due to the same polarity of negatively charged acceptor ions in the p-collector and the negative charge of electrons arriving at the collector side, no space-charge region can build up at this point even with a high electron density. Therefore, dynamic avalanche of the third degree is not possible in an NPT-IGBT.

Another situation is given in a modern IGBT with a field stop layer or an n-buffer layer in front of the collector layer. This is shown in Fig. 13.27. The electron current flows to the right hand side. If the hole current coming from the collector is not high enough to compensate the electron current, then at the n^--n-junction a space-charge can build up between the negative charged free electrons and the positively charged donor ions in the n-buffer, and an electric field can rise.

In [Rah05] it is explained that this process is dangerous especially at the final plasma removal and the transition of the IGBT into the “Switching Self Camping Mode” (SSCM). SSCM occurs in an IGBT during turn-off of an over-current under

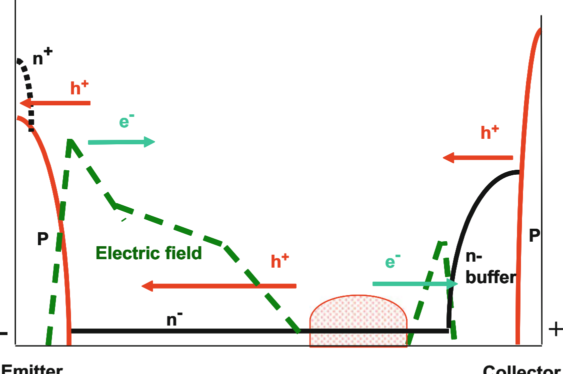


Fig. 13.27 IGBT with an n-buffer at turn-off of an over-current with dynamic avalanche. Under the worse conditions a second electric field can rise in front of the n-buffer

the condition of a high parasitic circuit inductance. In SSCM a field distribution with a second field peak at the nn$^+$-junction can occur, similar to second breakdown in a bipolar transistor. This effect is unstable. However if a sufficiently high hole current is delivered from the p-collector layer and it compensates the electron current, then the SSCM event is stabilized in the IGBT. To achieve this, the emitter efficiency of the p-collector and the corresponding current gain α_{pnp} must not be too small.

Figure 13.28 shows the turn-off of a 3.3 kV-IGBT at a high over-current level. After the voltage has climbed up to approx. 2000 V, a flattening of the voltage course is to be seen. This is a sign of strong dynamic avalanche, where the holes flowing through the space-charge limit the blocking capability in this time interval.

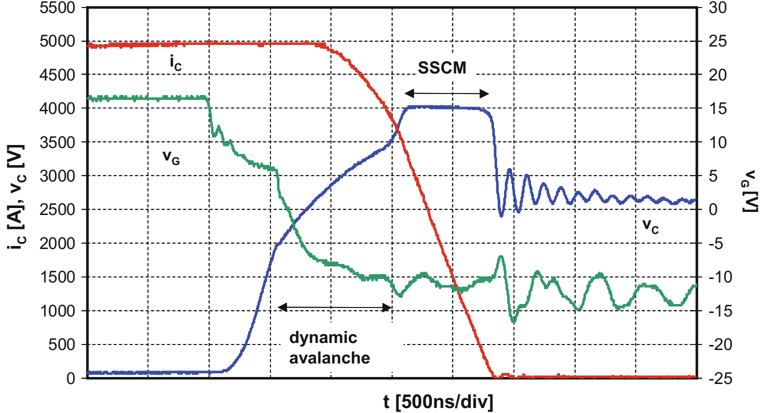


Fig. 13.28 Turn-off of a 3.3 kV 1200 A IGBT module at 4-times the rated current against a batters voltage V_{bat} = 2600 V. After an interval of strong dynamic avalanche, the SSCM event follows. Figure from [Rah05] © 2005 IEEE

After the voltage has reached the applied DC link voltage of 2600 V, the current starts to decay, and still strong dynamic avalanche exists in the device. At a voltage of 3500 V the device transits into the SSCM. The voltage ramps up to 4000 V, which is close to the static avalanche breakdown voltage of the device. The course of the current and voltage waveforms seems to be more stable compared to the SSCM process in a diode (see Fig. 13.21). The hole current, which comes from the p-collector, adds a stabilizing effect during the SSCM [Rah05]. Figure 13.28 shows that very high stress in dynamic avalanche is also possible for IGBTs.

Dynamic avalanche in 6.5 kV IGBTs was investigated in [Mue15]. At turn-off with low R_G, lower than recommended by the manufacturer, and subsequent high dv_C/dt, the IGBT survived millions of avalanche events. However, a small increase of the turn-off delay time and an increase of the di/dt at turn-on was found. The avalanche has injected charge carriers into the gate oxide. With this, feed-back capacity from the collector to gate, the miller capacitance, is modified [Mue15]. Strong dynamic avalanche is not recommended for repetitive operation of high-voltage IGBTs.

13.5 Exceeding the Maximum Turn-off Current of GTOs

During turn-off of a GTO thyristor, the current below the emitter finger is extracted from the edge towards the middle. This was shown in the paragraph on GTO thyristors, see Fig. 8.16. Finally a narrow current-conducting area remains in the middle of the emitter finger, representing a current filament before the anode current decays. Even if high accuracy of the fabrication process technology is given, not all emitter fingers of the GTO thyristor are ideally identical, and even in a single emitter finger, there will be a position which will be the last for current conduction. If the maximum turn-off current is exceeded, a molten zone is found at this point. After chemical removal of the metallization and etching in concentrated potash KOH, which solves polysilicon faster than mono-crystalline silicon, a narrow hole can be found as shown in Fig. 13.32.

The dark area in Fig. 13.29 shows the position of the final filament or current tube, while the molten channel reaches down to the backside of the device.

A failure picture as in Fig. 13.29 can be caused by exceeding the maximum turn-off current capability of the device. However it can also occur because of a failure of an element in the RCD-snubber (see Fig. 8.18), since during turn-off of a GTO thyristor the voltage slope dv/dt must be limited.

Fig. 13.29 GTO-Thyristor failed by exceeding the maximum turn-off current. Molten channel, created by a current filament in the middle of the cathode finger

13.6 Short-Circuit in IGBTs

13.6.1 Short Circuit Types I, II and III

Three types of short circuit have to be distinguished for the IGBT [Eck94, Eck95, Let95].

Short circuit I is a direct turn on of the IGBT to a short circuit. The course of the voltage V_C, the collector current I_C and the gate voltage V_G is shown in Fig. 13.30. Before turn-on, the voltage V_C is high, while at the gate a negative voltage is applied. After turn-on into the short circuit, the current increases to more than 6 kA, this is the value of the saturation current. This value is specified in some data sheets as I_{SC} and correspondents to the saturation current at $V_G = 15$ V. The IGBT is able to withstand the simultaneous load of a high current and a high voltage during the short circuit pulse for some time. It must be turned off within a defined time, typically specified to 10 µs or lower, to ensure safe operation and to avoid failures by overheating.

In Fig. 13.30 I_{SC} decreases with time due to self-heating of the device. During turn-off of the short circuit current, an inductive voltage peak is generated. In the figure this voltage peak amounts to approx. 2000 V. The condition for surviving of a short circuit pulse is that the peak voltage must remain lower than the rated voltage within the specified safe operation area. To ensure this, short circuit must be turned-off with a limited *di*/*dt*. Usually, for turn-off of the short circuit the driver uses a higher gate resistor to limit the *di*/*dt*. This soft turn-off is visible in Fig. 13.30.

Short circuit II is the occurrence of a short circuit during the conducting mode of the IGBT [Eck94, Eck95, Let95]. The occurrence is shown in Fig. 13.31 for the same IGBT as used in Fig. 13.30. The IGBT is carrying the load current I_{Load} and the voltage drop is V_{CEsat}. As soon as the short circuit has occurred, the collector current will increase very steeply. The *di*/*dt* is determined by the DC-link voltage V_{bat} and the inductance of the short-circuit loop. During the time interval I, the IGBT is desaturated. The consequently high dv_C/dt of the collector emitter voltage will induce a displacement current through the gate-collector capacitance, which

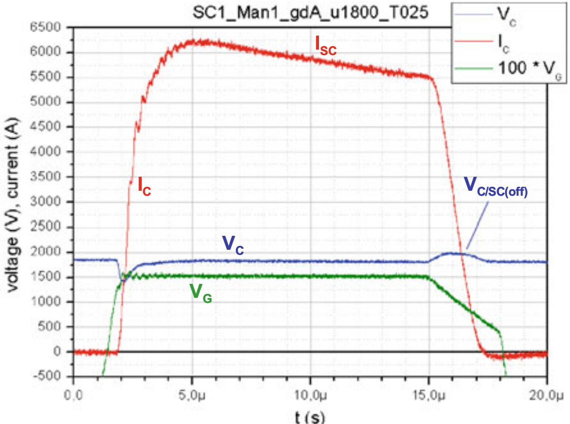


Fig. 13.30 Short circuit I. Figure from [Lut09b] © 2009 EPE

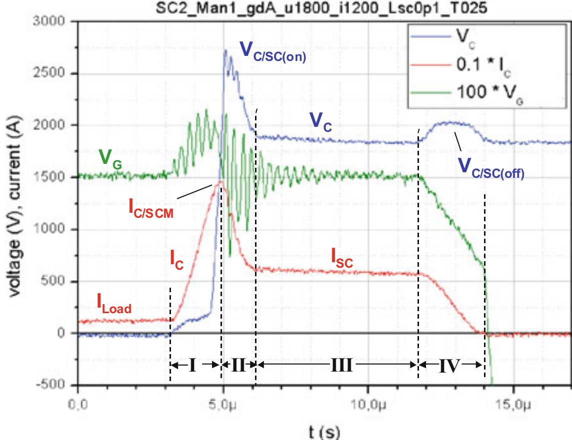


Fig. 13.31 Short circuit II. Figure from [Lut09b] © 2009 EPE

increases the gate-emitter voltage [Nic00]. As can be seen in Fig. 13.31, V_G is dynamically increased up to 20 V due to this effect. This increase of V_G in turn causes a high short-circuit current peak $I_{C/SC(on)}$, which in this case reaches a value of 14 kA.

Since the gate collector capacitance is high at low V_C, the effect of the displacement current will be most significant at low voltage. The increased gate voltage during desaturation leads to a negative gate current back to the driver. The onset of a negative gate current together with the interconnection parasitics and IGBT input capacitances are supposed to cause the oscillations in V_G showed in Figs. 13.31 and 13.33. After having completed the desaturation phase, the short-circuit current will drop to its static value I_{SC} (time interval II). Due to the current fall with negative di/dt, a voltage $V_{C/sc(on)}$ will be induced over the parasitic inductances, which becomes visible as an voltage peak on the IGBT:

$$V_{C/SC(on)} = V_{bat} + L_{SC} \cdot \left| \frac{di_C}{dt} \right|_{\max} \tag{13.20}$$

The stationary short-circuit phase (time interval III) is followed by the turn-off of the short circuit current. Due to the negative *di*/*dt*, the commutation circuit inductance will again induce an overvoltage $V_{C/SC(off)}$ across the IGBT (time interval IV). Once again it must be ensured that this voltage peak stays below the rated voltage respectively within the specified safe operation area. To limit $I_{C/SCM}$ and to keep the gate-emitter voltage within the permissible limits, V_G has to be clamped.

It should be noted that short circuit II is a harder condition than short circuit I. While $V_{C/SC(off)}$ can be limited by a soft turn-off function of the driver, this does not hold for $V_{C/SC(on)}$, which can easily overshoot the rated voltage of the device.

Additionally to the well understood dv_C/dt effect, an effect of accumulation of holes under the gate oxide is found for high-voltage IGBTs in [Mue16]. This so-called “self turn-on” is caused by a self-charging displacement current and increases the gate-voltage. It is found for some IGBTs even at low-inductive short circuit I, where V_C remains nearly constant.

Short circuit III is the occurrence of a short circuit across the load during the conducting mode of the freewheeling diode. SC III can occur in all typical IGBT applications, since there is always an interval where the freewheeling diode is conducting. In motor drive applications, the duty cycle for the IGBT is higher than for the freewheeling diode, and SC II is more probable. However, if e.g. a train drives down from a mountain, the motor is used as generator and energy is transferred back from the train to the grid. The duty cycle for the freewheeling diode is higher compared to the IGBT in this operation mode, and therefore SC III is more probable to occur than SC II.

The occurrence of a SC III event is explained using Fig. 13.32, in which a single phase inverter is shown.

Inverters with pulse width modulation usually are driven with complementary signals. As starting point we assume that IGBT 1 and IGBT 4 are on and the current flows through IGBT 1, the load and IGBT 4. As next, the IGBTs 1 and 4 are turned-off, the gate signal of IGBTs 2 and 3 is set to “on” after a short dead time. Because the inductance of the load determines the current direction, the current will in this case flow through diodes 2 and 3 back to the voltage source. If now a short

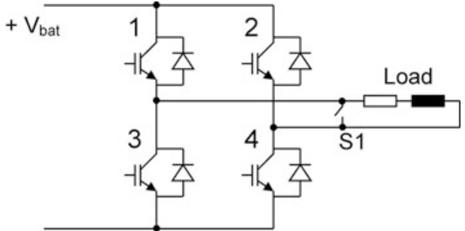


Fig. 13.32 Explanation of an SC III on example of a single phase inverter

circuit occurs across the load – symbolised by closing the switch S1 – diode 2 and 3 will be commutated with a high di/dt and IGBTs 2 and 3, which at this time interval have a gate-signal which is already set to "on", will be exposed to a short circuit. The current is thereby rapidly commutated from the diode to the IGBT, and the IGBT is turned-on passively into the short circuit. During the short circuit event, a reverse recovery process occurs at the freewheeling diode; however the voltage course is determined by the IGBT and the occurring high *dv*/*dt* while the IGBT transits to its desaturation mode.

Before the short circuit, a positive V_G = 15 V is already applied at the IGBT, however its current is zero because the inverse diode is conducting at this stage in the PWM pattern. When short-circuit occurs, the current changes its sign and the IGBT is turned-on in a passive, diode-like manner. At usual turn-on, the voltage across the IGBT is high before turn-on. In the now given case, the voltage is low. Therefore a forward recovery peak V_{FRM} at the IGBT occurs. This is related to the forward recovery of a diode. It was observed in [Pen98] in a specially designed zero-voltage switching circuit with relatively low voltage peaks. This forward recovery peak V_{FRM} can get very high, depending on *di*/*dt*. Several hundred volts can occur with wide base high voltage IGBTs. Forward recovery voltage peaks at IGBTs may be higher than at diodes [Bab09]. However, the measurement in Fig. 13.33 is at the outer terminals, and the parasitic module inductance contributes significantly. Additionally, the diode parallel to the IGBT is in the reverse recovery process, the measured current at this instant contributes of diode and IGBT current and cannot be resolved in the used setup.

The dynamic short circuit peak current $I_{C/SCM}$ amounts to 14 kA, this is almost the same value as in SC II. $I_{C/SCM}$ in SC III was found to be similar to $I_{C/SCM}$ in SC II for IGBTs from different manufacturers [Lut09b].

The occurrence of V_{FRM} is not supposed to be an extraordinary stress for the IGBT, since it occurs in the IGBT forward direction, where a high blocking capability is specified. The main additional stress in SC III for the module is the

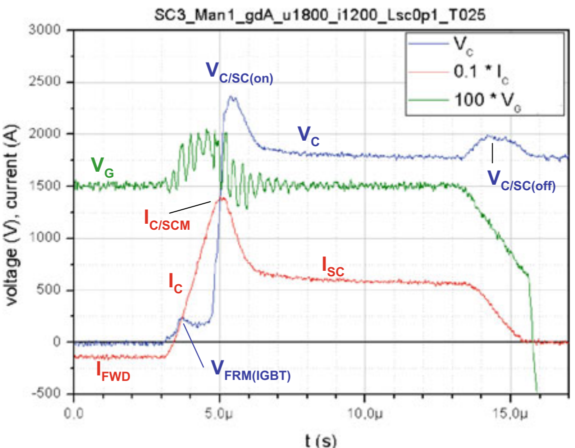


Fig. 13.33 Short circuit III. Figure from [Lut09b] © 2009 EPE

reverse recovery of the freewheeling diode during short circuit. The two-step voltage slope at reverse recovery of the diode, in which the second step may occur with an extraordinary high *dv/dt*, is a special stress event for the FWD. Usually, IGBTs fail in short circuit. However, in one of the SC III experiments, a diode failure was observed while the IGBT withstood [Lut09b].

While in Fig. 13.33 only the module current could be measured, a measurement of IGBT- and FWD current at SC III was executed in [Fuh15]. As to be seen, the reverse-recovery of the diode occurs first at low voltage between 0.6 and 1 µs, then at t = 1 µs a voltage with high *dv/dt* is exposed to the diode. A second reverse recovery peak occurs, the diode is exposed to strong dynamic avalanche (Fig. 13.34).

A typical picture of a 3.3-kV IGBT, destroyed by short circuit is shown in Fig. 13.35. The large-area burned emitter regions are typical. A picture of a destroyed IGBT similar to Fig. 13.35 gives the specialist a hint that short circuit is probably the failure reason. The burned-off emitter regions are typically found for SC I as well as for SC II and SC III failures.

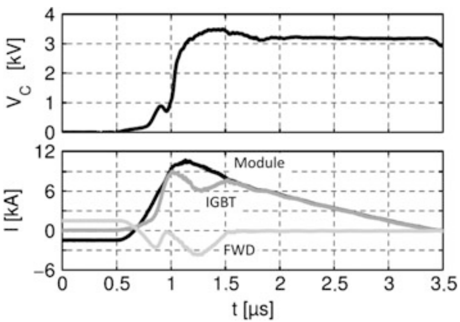


Fig. 13.34 Measurement of short-circuit type III with a 3.3-kV IGBT at V_{bat} = 2.9 kV and I_C = −1.5 kA. Figure adapted from [Fuh15]

Fig. 13.35 IGBT die (3.3 kV) destroyed by short circuit

13.6.2 Thermal and Electrical Stress in Short Circuit

From its basic function, the IGBT has a short circuit capability that limits the current. I_{SC} is the current in the active region of the IGBT forward I–V characteristic and can be approximated similar to Eq. (10.3)

$$I_{SC} = \frac{1}{1-\alpha_{pnp}} \cdot \frac{\kappa}{2} \cdot (V_G - V_T)^2|_{V=V_{bat}} \tag{13.21}$$

An electric field exists in the device (SC I), or is built up at desaturation (SC II, SC III). In short circuit II, there might be remaining plasma from the saturation mode before short circuit. The electric field is built up quickly, and the remaining plasma is extracted in a short time. Figure 13.36 shows the process in an NPT-IGBT in the short circuit mode. An electric field has built up at the blocking pn-junction, which takes over the applied voltage V_{bat}. On the emitter side, the left hand side in Fig. 13.36, the n-channel is conducting. Electrons are flowing into the space-charge region, holes are injected from the p-emitter. In the electric field, the carriers flow with their drift velocities given in Chap. 2 with Eq. (2.39). The current density j_{SC} is composed of $j_n + j_p$, where

$$\begin{aligned} j_n &= q \cdot n \cdot v_n \\ j_p &= q \cdot p \cdot v_p \end{aligned} \tag{13.22}$$

Since v_n, v_p are now much higher than in a plasma layer, the total amount of electrons and holes is much lower than at forward conduction where n, p are in the range of 10^{16} cm^{-3}. Sometimes $v_{sat(n,p)}$ is used for calculation of the carrier densities, however especially for holes the drift velocity is often not saturated at the given fields. n, p are typically in the range of several 10^{14} cm^{-3} and clearly above

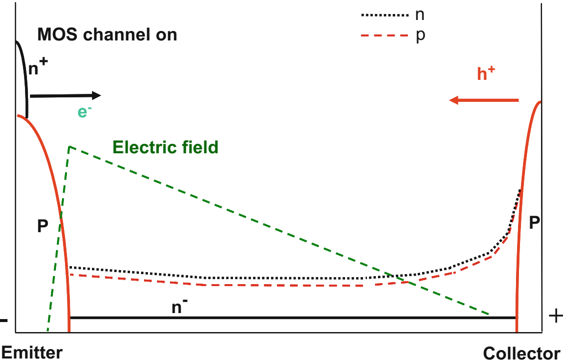


Fig. 13.36 Process in an NPT-IGBT during short circuit

the background doping N_D. The mobile carriers therefore lead to a feedback to the electric field, and it holds

$$\frac{dE}{dx} = \frac{q}{\varepsilon}\left(N_D - \frac{j_n}{q \cdot v_n} + \frac{j_p}{q \cdot v_p}\right) \tag{13.23}$$

where the terms in the bracket form an effective doping $N_{eff} = N_D - n + p$. The detailed feedback depends on the used IGBT technology and the conditions. In PT-IGBTs with the applied high p-emitter efficiency, typically the term caused by the hole current is dominating, and N_{eff} will be increased. In NPT-IGBTs, the electron density n is above the hole density, and the reduced dE/dx leads to a wider extension of the space charge into the n^--layer, reducing the electric field at the pn-junction [Las92]. In IGBTs with buffer layer, even $n > N_D + p$ may occur, the gradient of the field will be inverted and the field peak shifts to the collector side.

The active region is a stable condition for a transistor; the deposited energy in the case of a SC I is

$$E = V_{bat} \cdot I_{SC} \cdot t_{SC} \tag{13.24}$$

and due to the simultaneous high V_{bat} and I_{SC}, the temperature increases fast. If the short circuit is turned off within the requested time interval – 10 μs for former IGBT generations, 7 μs for new generations – the IGBT can in most cases withstand the thermal stress occurring during the short circuit mode.

The holes are flowing through the p-well close to the emitter layer; this was shown in Fig. 10.14. The voltage drop in this path must be well below the built-in voltage V_{bi} of the junction [Las03]. The density of carriers in short circuit is much lower than at rated current, where the IGBT is flooded with plasma. On the other hand, during short circuit very high temperatures occur. A high temperature decreases V_{bi}, therefore the danger of latch-up increases with increasing temperature. Nevertheless, there was much progress in manufacturing highly doped p^+-wells, therefore latch-up is usually no more a limit for the short circuit capability of latest IGBT generations.

Equation (13.21) has given the parameters that determine I_{SC}. α_{pnp} is adjusted to be low in modern IGBTs by a low emitter efficiency at the collector side. It is typically in the range of 0.33–0.4. The other decisive factor in (13.21) is the channel parameter κ which gives the conductivity of the MOS-channel. Since a comparatively high cell distance is of advantage in modern IGBTs – see Sect. 10.6 – the channel parameter κ is kept moderate. Therefore, in modern IGBTs a low I_{SC} can be achieved despite a high plasma density at the emitter side during usual forward conduction. As an example, Fig. 13.37 shows a comparison of a modern and conventional IGBT [Mor07]. For the modern IGBT (HiGT), the saturation current is not increased, despite the lower V_C at the typical operation condition (denoted $V_{CE(sat)}$ in Fig. 13.37). The I_{SC} of the 50-A rated IGBT is approximately 175 A. Since the thermal short circuit capability depends on the deposited energy in short

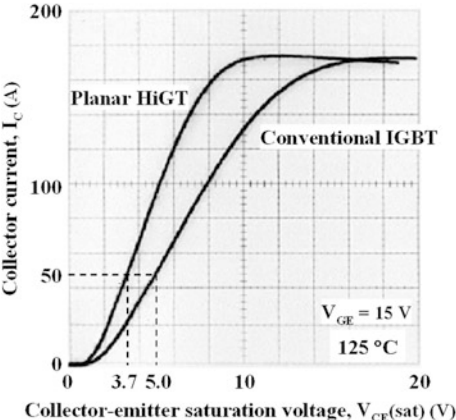


Fig. 13.37 Measured forward characteristic of a 3.3-kV conventional IGBT and planar HiGT (modern type). Figure from [Mor07] © 2007 IEEE

circuit, which is given by $I_{SC} \cdot V_{bat} \cdot t_{sc}$, the modern IGBT has the comparable short-circuit capability as the conventional one.

In several data sheets of 1200-V IGBTs, I_{SC} is even lower for the new IGBT generations. I_{SC} is specified to approx. 4 times the rated current, compared to 5–6 times the rated current for the older generations. In combination with a very high doping of the p-well (Fig. 10.14), leading to a very low resistance R_S (see Fig. 10.1 b), very high short circuit ruggedness is achieved in modern IGBTs. However in future IGBT-generations with further reduced device area and device volume, the allowed time t_{sc} before turn-off of the short circuit pulse will be limited from 10 µs to a lower value of 5–7 µs. This is possible in the application with meanwhile available fast-reacting improved gate drive units.

13.6.2.1 Thermal Limits for Medium-Voltage IGBTs

In medium-voltage IGBTs, the short circuit capability is limited by temperature. The destruction of a 600-V IGBT by short circuit I is shown in Fig. 13.38. The applied stress in this case is beyond the short circuit capability of the device. In Fig. 13.38a, the short circuit occurs at a battery voltage of 540 V during a time of approx. 60 µs, until the device is destroyed. In Fig. 13.38b the short-circuit current is turned-off after approx. 40 µs. During the short circuit pulse the device has been heated to a level that a high leakage current is flowing after the turn-off event. At such high temperatures additional charge carriers are generated thermally, see Eq. (2.6), and the leakage current increases further. After approx. 100 µs or more, the device is destroyed by overheat caused by the high leakage current.

Long time tests with repetition of short circuit events led to the conclusion that short circuit can be repeated up to ten-thousand times without destruction of the device [Sai04]. This holds true as long as the deposited energy stays smaller than a certain critical energy E_C. A summary of the repeated short circuit tests of a 600-V IGBT is shown in Fig. 13.39.

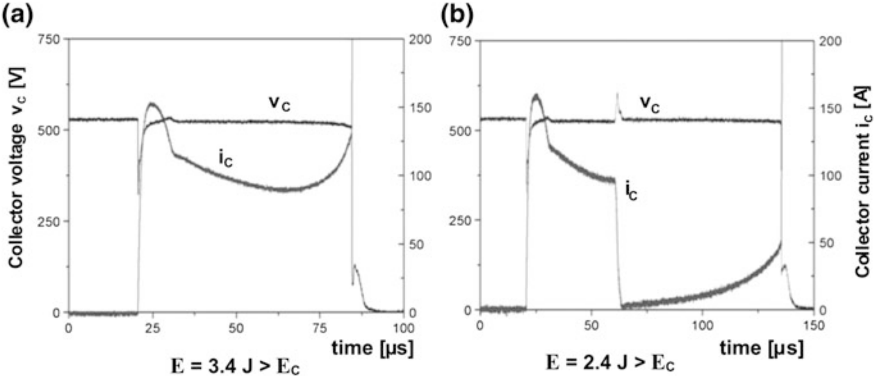


Fig. 13.38 600 V IGBTs in short circuit at V_{bat} = 540 V, T = 125 °C. Course of current and voltage at destruction of IGBTs at a deposited energy beyond the critical energy E_C. Figure from [Sai04]

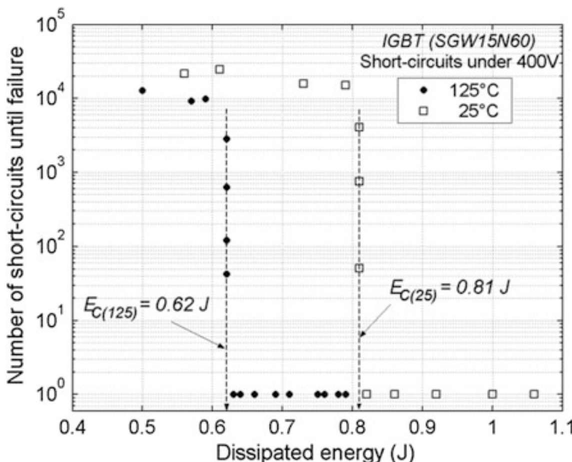


Fig. 13.39 IGBT Ruggedness at repeated short circuit events. 15-A 600-V IGBT, V_{bat} = 405 V. Figure from [Lef05] © 2005 IEEE

A defined limit for the dissipated energy was found and this critical energy E_C is lower for the T = 125 °C compared to T = 25 °C. Above the limit E_C, IGBTs are destroyed after one event by overheating. Since heat transport out of the device is small for the requested short time of the short circuit load, the temperature increase can be calculated according to the thermal capacity of the device, compare Eq. (11.8), treating the deposited energy E in (13.24) as thermal energy Q_{th}, by

$$\Delta T_{SC} = \frac{E}{C_{th}} = \frac{V_{bat} \cdot I_{SC} \cdot t_{SC}}{c \cdot \rho \cdot d \cdot A} \qquad (13.25)$$

The evaluation of different IGBTs and the following calculation of the temperature increase for the volume of the semiconductor in which the electric field occurs resulted in temperatures in the range of 600 °C [Sai04]. For IGBTs with different thicknesses, a similar final temperature was found also.

The results in Fig. 13.39 show that the failures in short circuit are purely thermal for the investigated 600-V IGBTs. Additionally, special attention was taken to find the ageing mechanisms of IGBTs which fail below E_C, but after a large number of short circuit pulses. No trend in leakage current and threshold voltage was found. However it was observed that the forward voltage drop V_C increases and the short circuit current I_{SC} decreases with the number of cycles. Failure analyses showed the increase of the resistivity of the Al metallization layer, a strong degradation of the die metallization by Al reconstruction after approx. 10,000 cycles (Fig. 13.40), and also a strong degradation of the bond wire attach [Lef08].

The degradation of metallization may lead to inhomogeneous current distributions and finally local overheating. The observed aging mechanisms have similarities to power cycling, where also the effects of aging of the metallization layer and bond-wire lift-off are observed [Lut08], see Sect. 12.7.2.

Since the main destructive effects are due to temperature, the critical energy E_C, which the device can take up as heat, depends on its thermal capacity. Modern IGBTs are designed with narrow widths w_B for the low-doped n-base layers to reduce the overall losses. Additionally the reduced voltage drop V_C during forward conduction allows higher rated currents for a given device area. Therefore the area as well as the thickness of the IGBT die is reduced, and the thermal capacity is decreased correspondingly.

An example of the manufacturer Infineon for the die area of different 1200 V 75A IGBT-generations is shown in Fig. 13.41. The thickness of the dies for different IGBT generations and three different voltage ratings is shown in Fig. 13.42. For the 1200 V, 75 A IGBT chip, the area is reduced down to 44% and the

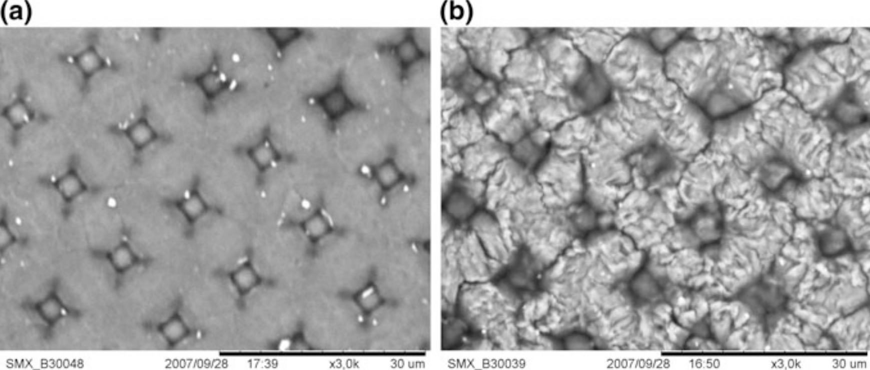


Fig. 13.40 Reconstruction of Al metallization at repeated short circuits: **(a)** before test **(b)** after 24,600 short-circuit cycles. Figure from [Lef08]

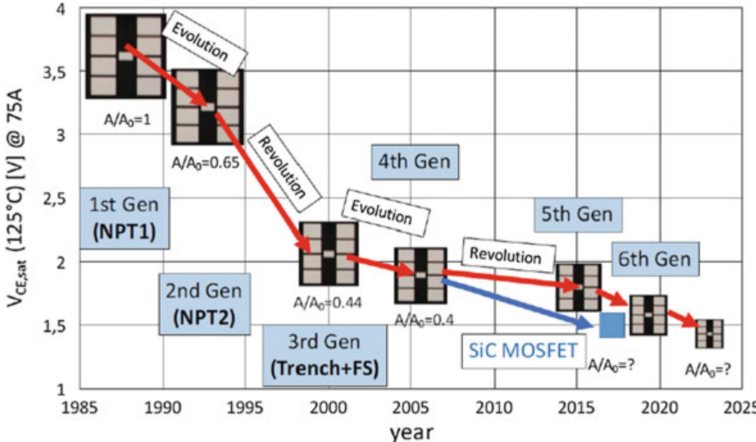


Fig. 13.41 Die area and forward voltage drop for different IGBT generations, on example of the manufacturer Infineon. Figure from T. Laska and T. Basler, Infineon (2016)

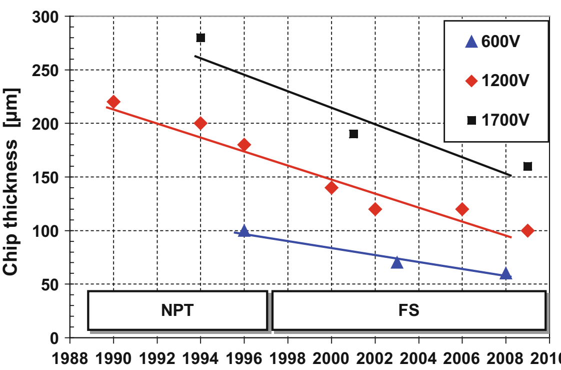


Fig. 13.42 Device thickness for different IGBT generations of the manufacturer Infineon. Figure from Infineon

thickness to 55% compared to the first IGBT generation from 1990. According to Eq. (11.8) this is a reduction of the thermal capacity down to 24%.

With this drastic reduction of the thermal capacity, also the thermal energy, which can be deposited in a device, is reduced in the same way. To maintain ruggedness during short circuit, I_{SC} has to be kept low.

13.6.3 Current Filamentation at Short Circuit

For recent 600 and 1200-V IGBTs the short circuit failure limit is mainly due to thermal reasons and ageing [Lef08]. The temperature increase is different for high voltage IGBTs. Due to the larger thickness of the device and due to the reduced

current density, the temperature increase calculated with Eq. (13.25) is typically smaller. For high voltage IGBTs, it was found: when short circuit limits are determined, failures are not observed after a long t_{sc}, or at short circuit turn-off, or after the turn-off has been completed, as shown in Fig. 13.37. They occur typically during the stationary phase in SC I. An example for such a measurement is shown in Fig. 13.43 [Kop09].

Figure 13.44 shows a summary of short circuit failures for the 6.5 kV IGBT as function of the applied DC-link voltage V_{bat}. At around V_{bat} = 2000 V the capability reaches a minimum and starts increasing again for higher V_{bat}. Note, the ruggedness is higher at 4500 V than at 1000 V!

A minimum for the short-circuit capability is found between 1500 and 2500 V, similar minima were found for a 4.5 kV rated IGBT between 1200 and 1800 V, and at about 1000 V for a 3.3 kV rated IGBT. This special voltage dependency needs new explanations.

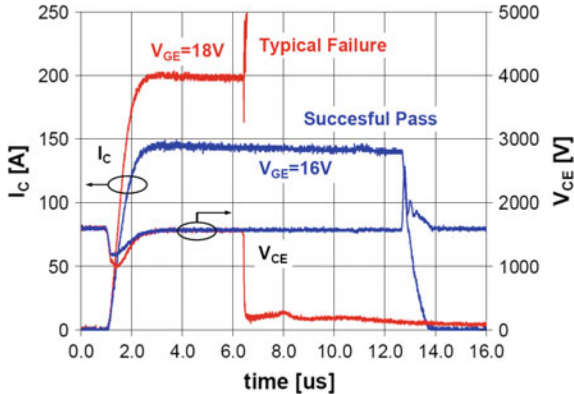


Fig. 13.43 Typical short-circuit pass and fail waveforms, SC I measurement of a 6.5 kV IGBT. Figure from [Kop09] © 2009 IEEE

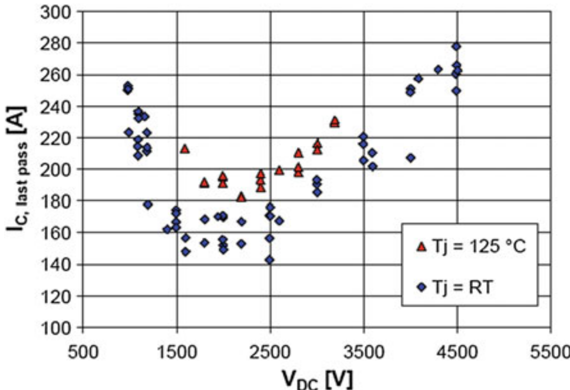


Fig. 13.44 Short-circuit capability for a 6.5 kV IGBT in dependency on the DC-link voltage V_{bat} and starting junction temperature. Figure from [Kop09] © 2009 IEEE

In these high voltage IGBTs with a buffer layer at the collector side, a redistribution of the electric field to a field peak at the collector side under short circuit stress was found [Kop08, Kop09], which results from a high electron density $n > N_D + p$ (Eq. (13.23)), see Fig. 13.45. In the explanation, these field peaks at the nn$^+$-junction were evaluated to be moderate, and because of the increasing ruggedness with increasing voltage second breakdown and Egawa-type failure mechanisms were excluded.

It was found that an increased emitter efficiency of the collector-side p-emitter improves the ruggedness [Kop08, Kop09]. The increased α_{pnp} is, for the given low α_{pnp}, of low effect to I_{SC}. This experimental result also excludes latch-up as failure root cause. The formation of filaments is suggested as failure mechanism. In [Kop09] these filaments, found in numerical simulation, are explained to result from the different drift velocities of electrons and holes. Especially at medium voltage the electric fields are moderate, and the drift velocity for the holes is far below that of electrons, see Chap. 2, Fig. 2.15. Filaments with high local current density lead to IGBT destruction.

The same effect of reduced short-circuit capability at medium voltage is reported in [Bab14] for a 1200 V IGBT. The experimentally found lowest short circuit capability is found at 600 V. Except the different voltage on the *x*-axis, the voltage-dependency is very similar as in Fig. 13.44.

The field redistribution to an electric field peak at the junction form n$^-$-layer to the n-buffer is shown in Fig. 13.45. In the low-field region at the emitter side, a quasi-plasma layer forms. Tanaka and Nakagawa [Tan15] point-out that filaments in the short-circuit mode occur if the electric field at the n$^-$n-junction and the avalanche generation rate exceeds the critical value. In difference, in [Bho16] the simulation results also suggest that short-circuit destruction can be caused by a current filament not being triggered by avalanche generation. The current filamentation is considered to be the main origin of the device failure in both explanations.

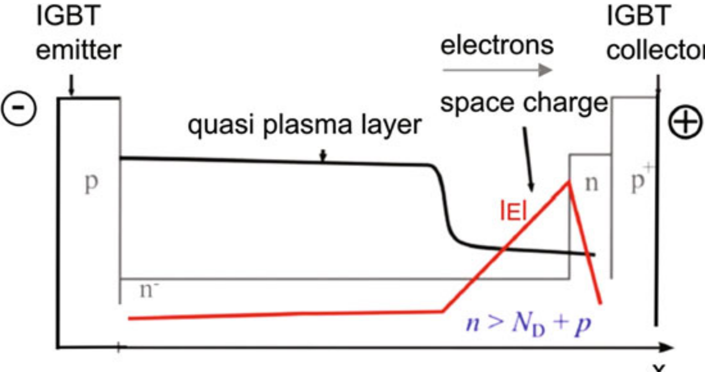


Fig. 13.45 Simplified view of the quasi-plasma layer and the electric field strength during IGBT short circuit with high current. Figure following [Bab14]

Already from the first publications on the effect [Kop08] it is known that increase of the efficiency of the collector-side p-emitter can counteract the filament formation. Also in [Tan16] a lightly doped n-buffer is applied to increase the p-emitter efficiency. However, higher p-emitter efficiency increases the leakage current, and this limits the applicability. A new 2-stage emitter is suggested in [Bho17]. It is formed by an additional buried discontinuous p-layer before the collector layer, see Fig. 13.46. This additional layer act as a hole current amplification stage.

The hole amplification function is shown in Fig. 13.47. In the standard structure, an electric field up to 180 kV/cm grows up in the high-current density filament at the $n^{-}n$-junction. In the new structure, an even lower hole density is found directly at the collector. An enhanced hole density, however, is found in front of the second discontinuous floating p-layer. Because of the high hole density, the condition for formation of a field peak at this position $n > N_D + p$ is only fulfilled at much higher short circuit current, and at comparable current no field peak is built up at this position. The electric field remains rectangular at a low level.

With the novel IGBT structure, the SC capability of modern IGBTs can be improved significantly, because failures due to the current crowding during SC turn-off and during pulse can be suppressed to a considerable extent. The leakage current can be reduced by the lower p-doping of the collector layer in the new structure. Therefore, the failures due to the thermal runaway after a SC event can also be reduced and suitable time for short-circuit pulse can be increased [Bho17]. Up to now, the new structure was only analyzed in simulations and not verified experimentally.

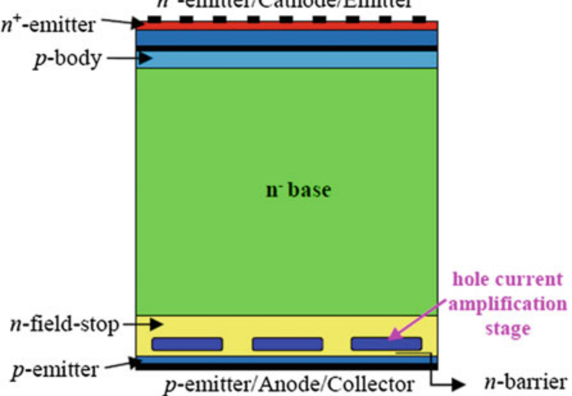


Fig. 13.46 IGBT structure with floating p-amplification stage in front of the p-emitter. Front side IGBT structure simplified. © 2017 IEEE Reprinted, with permission, from [Bho17]

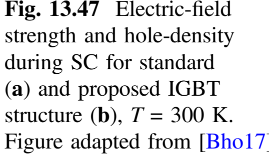
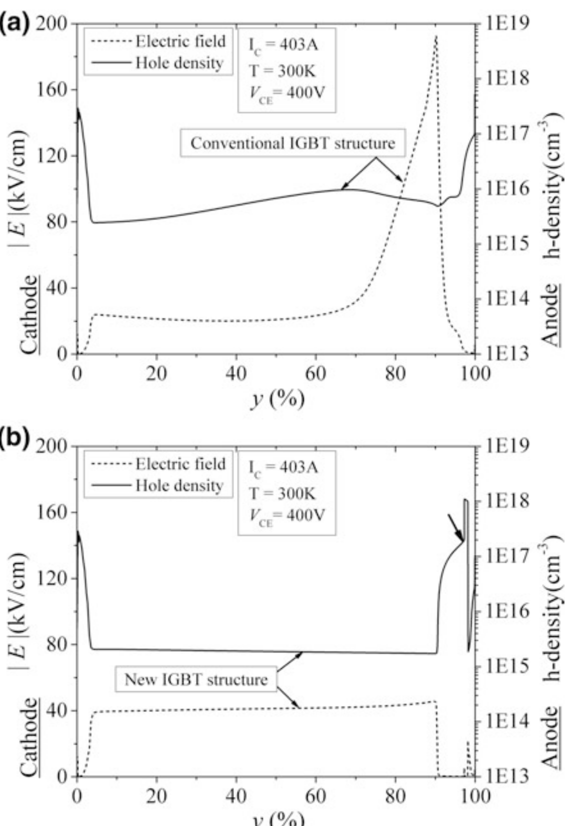


Fig. 13.47 Electric-field strength and hole-density during SC for standard (**a**) and proposed IGBT structure (**b**), T = 300 K. Figure adapted from [Bho17]



13.7 Failure Analysis in IGBT Circuits

Some failure mechanisms occurring in power circuits with IGBTs are discussed here in summary and interrelation. Table 13.1 gives different failure mechanisms, which are divided to failures caused by current and respectively by temperature, by voltage, and by dynamic effects. The failure reasons are printed in cursive letters.

For failures caused by current, a molten zone within the active area of the device is typical. At very high average currents, one finds a destroyed area with a diameter of several millimeters.

If a surge current failure of a diode occurs, the molten area is usually smaller, in the range of 1 mm. For bonded diodes, often a melting of metallization beside the bond feet is observed. Surge current occurs in an application for example when a non-loaded DC link capacitor is connected by a diode rectifier circuit to the grid, and a very high current pulse is generated in the first instant. In this case, an application fault is given. A loading circuit for the DC link capacitor will be of help in this case. Manufacturers of power semiconductors know very well the typical

Table 13.1 Some failure mechanisms in IGBT modules

Current temperature	Voltage	Dynamic Effects
		Applied voltage below the rated voltage
–Too high average current The die shows molten areas with a diameter of several mm. Failures located in the active area	*–Production fault* Failure location starts at the edge	*–Lack of dynamic ruggedness of the freewheeling diode* Diode and transistor in the associated commutation loop are destroyed
–Surge current exceeded Local molten area, size approx, 1 mm, sometimes cracks in the crystal, failure located in the active area	*–Voltage peaks above rated voltage* Failure location starts at the edge	*–Lack of dynamic ruggedness of the freewheeling diode* Only the diode is destroyed, pinhole with diameter <100 µm
–Short circuit capability of IGBT exceeded Only IGBTs destroyed, large part of the emitter area burned off	*–Lack of long term stability of the passivation layer* Failure location starts at the edge	*–Dynamic avalanche 3rd degree* Pinhole with a diameter <100 µm, cracks in the crystal lattice originating

Cursive: *failure reason*. Normal: failure picture

surge current failure pictures of their devices, as shown for example in Fig. 13.6, and they can identify such failures.

If failures are caused by voltage, it is mostly observed that the failure location is at the edge of the device, such as in the junction termination structure. At these positions, the highest electric fields occur at the surface. The junction termination is very important in power device manufacturing and is most sensitive to contamination in the production line, to faults in photolithography due to dust particles etc. If a device has a weak point due to a fault in production, this occurs usually at the edge of the device. While current induced failures are mostly due to faults in the application, this must not be the case for voltage induced failures. Application faults – voltage peaks higher than the rated voltage - as well as production faults must be taken into account.

Failures by dynamic effects are mainly related to switching events. The voltage stays below the rated voltage of the device. At switching events the transistors interact with their freewheeling diode. Figure 13.48 shows the corresponding commutation loops. At power transmission to the load (Fig. 13.48a) IGBT1 commutates with diode D2. If diode D2 fails during its turn-off, the associated transistor in the commutation loop turns on to a short circuit; hence, a short circuit within the bridge with very low inductance is given. Therefore, the IGBT may be destroyed by a short circuit. The same holds for the reverse power flux, where the commutation loop for this is shown in Fig. 13.48b. There IGBT2 commutates with diode D1.

If diodes as well as their associated transistors in the commutation loop are destroyed, the failure reason is usually due to the diode. If the diode fails, the IGBT

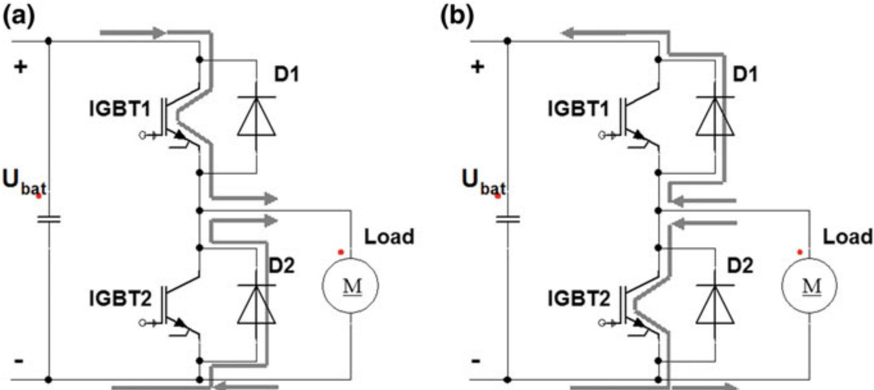


Fig. 13.48 Commutation loops in a half-bridge. **(a)** Power transmission from the DC link to the load. **(b)** Reverse power flux

may be destroyed. On the other hand, if the IGBT fails, this event causes usually no stress for the diode and the diode has no reason to fail. An exception might be, if a spark burns inside the module, which destroys further devices.

With such considerations, sometimes conclusions for failure reasons are still possible even if a module is heavily destroyed.

If the freewheeling diode fails and the IGBT turns-off successfully following short circuit, one finds afterwards a typical pinhole in the diode. For freewheeling diodes rated at 1200–1700 V, these pinholes are a sign of lack of dynamic ruggedness. Freewheeling diodes with higher rated voltages however may be destroyed by a very high current density in the reverse direction and a contemporaneous high voltage. For such failures by dynamic avalanche of the third degree one also finds cracks in the crystal lattice, originating at the pinhole. These cracks are a sign of very high local temperature spots. An example was given in Fig. 13.19.

If only the transistor is destroyed, the failure reason must be searched in the transistor. Short circuit failure is one of the possible mechanisms and is always worth considering. For short circuit failures, burned-off emitter regions with large areas are typical. Note, however, at short circuit III (Sect. 13.6) even a diode failure may occur due to the extreme high dv/dt and voltage spike at diode commutation, while the IGBT may survive [Lut09b].

Another IGBT failure mode is dynamic latch-up, which is caused by single weak cells in the IGBT device. Such failures cannot be found during tests of static parameters at the device manufacturer. For modules with parallel connection of many IGBT dies, the manufacturers execute final tests in an application conform circuit including dynamic turn-off under very high stress conditions in order to find single weak devices and to avoid failures in the targeted application.

Failure analysis is a complex process and a lot of experience is necessary. There are other failure modes, e.g. cosmic ray failures, see Sect. 12.8. Also for such failures pinholes are typical. In most cases one can not conclude a failure reason directly from a failure picture since there are different failure modes which may lead to similar failure pictures. The power circuit and the application condition must be investigated additionally.

With devices connected in parallel, a non-symmetric assembly may lead to oscillations and subsequently overstress the devices at special positions. Often new questions come forth during failure analysis. In conclusion, failure research is very complex, however the results are often extremely valuable.

References

[Bab08] Baburske, R., Heinze, B., Lutz, J., Niedernostheide, F.J.: Charge-carrier plasma dynamics during the reverse-recovery period in p^+-n^--n^+ diodes. IEEE Trans. Electron. Dev. **55**(8), 2164–2172 (2008)

[Bab09] Baburske, R., Domes, D., Lutz, J., Hofmann, W.: Passive turn-on process of IGBTs in Matrix converter applications. In: Proceedings EPE, Barcelona (2009)

[Bab10] Baburske, R., Lutz, J., Heinze, B.: Effects of negative differential resistance in high power devices and some relations to DMOS structures. In: Proceedings 2010 IEEE International Reliability Physics Symposium, pp. 162–169 (2010)

[Bab11] Baburske, R.: Dynamik des Ladungsträgerplasmas während des Ausschaltens bipolarer Leistungsdioden, Dissertation, Universitätsverlag Chemnitz (2011)

[Bab14] Baburske, R., van Treek, V., Pfirsch, F., Niedernostheide, F.J., Jaeger, C., Schulze, H. J., Felsl, H.P.: Comparison of critical current filaments in IGBT short circuit and during diode turn-off. In: Proceedings of ISPSD '14, pp. 47–50 (2014)

[Ben67] Benda, H.J., Spenke, E.: Reverse recovery process in silicon power rectifiers. Proc. IEEE **55**(8), 1331–1354 (1967)

[Bho16] Bhojani, R., Palanisamy, S., Baburske, R., Schulze, H.J., Niedernostheide, F.J., Lutz, J.: Simulation study on collector side filament formation at short-circuit in IGBTs. In: Proceedings of ISPS, pp. 70–76 (2016)

[Bho17] Bhojani, R., Baburske, R., Schulze, H.J., Niedernostheide, F.J., Lutz, J.: A Novel Injection Enhanced Floating Emitter (IEFE) IGBT structure improving the ruggedness against short-circuit and thermal destruction. In: Proceedings of ISPSD '17, Sapporo, pp. 113–116 (2017)

[Bie08] Biermann, J., Pfaffenlehner, M., Felsl, H.P., Gutt, T., Schulze, H.: CIBH diode with superior soft switching behavior in 3.3 kV modules for fast switching applications. In: Proceedings of the PCIM Europe, 367–371 (2008)

[Chm06] Chen, M., Lutz, J., Domeij, M., Felsl, H.P., Schulze, H.J.: A novel diode structure with Controlled Injection of Backside Holes (CIBH). In: Proceedings of ISPSD, Neaples (2006)

[Dom99] Domeij, M., Breitholtz, B., Östling, M., Lutz, J.: Stable dynamic avalanche in Si power diodes. Appl. Phys. Lett. **74**(21), 3170 (1999)

[Dom03] Domeij, M., Lutz, J., Silber, D.: On the destruction limit of si power diodes during reverse recovery with dynamic avalanche. IEEE Trans. Electron Dev. **50**(2), 486–493 (2003)

[Eck94] Eckel, H.G., Sack, L.: Experimental investigation on the behaviour of IGBT at short-circuit during the on-state. In: 20th International Conference on Industrial Electronics, Control and Instrumentation, IECON'94, vol. 1, pp. 118–123 (1994)

[Eck95] Eckel, H.G., Sack, L.: Optimization of the short-circuit behaviour of NPT-IGBT by the gate drive. In: Proceedings of EPE 1995, Sevilla, pp. 213–218 (1995)

[Ega66] Egawa, H.: Avalanche characteristics and failure mechanism of high voltage diodes. IEEE Trans. Electron Dev. **ED-13**(11), 754–758 (1966)

[Fel06] Felsl, H.P., Heinze, B., Lutz, J.: Effects of different buffer structures on the avalanche behaviour of high voltage diodes under high reverse current conditions. IEE Proc. Circuits Devices Syst. **153**(1), 11–15 (2006)

[Fel08] Felsl, H.P., Pfaffenlehner, M., Schulze, H., Biermann, J., Gutt, T., Schulze, H.J., Chen, M., Lutz, J.: The CIBH diode – great improvement for ruggedness and softness of high voltage diodes. In: Proceedings of ISPSD 2008, Orlando (2008)

[Fuh15] Fuhrmann, J., Member, IEEE, Klauke, S., Eckel, H.G.: IGBT and diode behavior during short-circuit type 3. IEEE Trans. Electron Dev. **62**, 3786–3791 (2015)

[Ful67] Fulop, W.: Calculation of avalanche breakdown voltages of silicon p-n junctions. Solid State Electron. **10**, 39–43 (1967)

[Gha77] Ghandhi, S.K.: Semiconductor Power Devices. Wiley, New York (1977)

[Hei05] Heinze, B., Felsl, H.P., Mauder, A., Schulze, H.J., Lutz, J.: Influence of buffer structures on static and dynamic ruggedness of high voltage FWDs. In: Proceedings of the ISPSD, Santa Barbara (2005)

[Hei06] Heinze, B., Lutz, J., Felsl, H.P., Schulze, H.J.: Influence of edge termination and buffer structures on the ruggedness of 3.3 kV silicon free-wheeling diodes. In: Proceedings of the 8th ISPS, Prague (2006)

[Hei07] Heinze, B., Lutz, J., Felsl, H.P., Schulze, H.J.: Ruggedness of high voltage diodes under very hard commutation conditions. In: Proceedings EPE 2007, Aalborg, Denmark (2007)

[Hei08] Heinze, B., Lutz, J., Felsl, H.P., Schulze, H.J.: Ruggedness analysis of 3.3 kV high voltage diodes considering various buffer structures and edge terminations. Microelectron. J. **39**(6), 868–877 (2008)

[Hei08b] Heinze, B., Baburske, R., Lutz, J., Schulze, H.J.: Effects of metallisation and bondfeets in 3.3 kV free-wheeling diodes at surge current conditions. In: Proceedings of the ISPS, Prague (2008)

[Hie97] Hierholzer, M., Bayerer, R., Porst, A., Brunner, H.: Improved characteristics of 3.3 kV IGBT modules. In: Proceedings PCIM Nuremberg (1997)

[How70] Hower, P.L., Reddi, K.: Avalanche injection and second breakdown in transistors. IEEE Trans. Electron Dev. **17**, 320 (1970)

[How74] Hower, P.L., Pradeep, K.G.: Comparision of one- and two-dimensional models of transistor thermal instability. IEEE Trans. Electron Dev. **21**(10), 617–623 (1974)

[Kin05] Kinzer, D.: Advances in power switch technology for 40 V–300 V applications. In: Proceedings of the EPE, Dresden (2005)

[Kop05] Kopta, A., Rahimo, M.: The Field Charge Extraction (FCE) diode – a novel technology for soft recovery high voltage diodes. In: Proceedings of ISPSD Santa Barbara, pp. 83–86 (2005)

[Kop08] Kopta, A., Rahimo, M., Schlapbach, U., Kaminski, N., Silber, D.: Neue Erkenntnisse zur Kurzschlussfestigkeit von hochsperrenden IGBTs, Kolloquium Halbleiter-Leistungsbauelemente, Freiburg (2008)

[Kop09] Kopta, A., Rahimo, M., Schlapbach, U., Kaminski, N., Silber, D.: Limitation of the short-circuit ruggedness of high-voltage IGBTs. In: Proceedings of ISPSD, Barcelona, pp. 33–37 (2009)

[Las92] Laska, T., Miller, G., Niedermeyr, J.: A 2000 V non-punchthrough IGBT with high ruggedness. Solid State Electron. **35**(5), 681–685 (1992)

[Las03] Laska, T., et al.: Short circuit properties of trench/field stop IGBTs design aspects for a superior robustness. In: Proceedings of ISPSD, Cambridge, pp. 152–155 (2003)

[Lef05] Lefebvre, S., Khatir, Z., Saint-Eve, F.: Experimental behavior of single chip IGBT and COOLMOS™ devices under repetitive short-circuit conditions. IEEE Trans. Electron Dev. **52**(2), 276–283 (2005)

[Lef08] Lefebvre, S., Arab, M., Khatir, Z., Bontemps, S.: Investigations on ageing of IGBT transistors under repetitive short-circuits operations. In: Proceedings of the PCIM Europe, Nuremberg (2008)

[Let95] Letor, R.R., Aniceto, G.C.: Short circuit behavior of IGBT's correlated to the intrinsic device structure and on the application circuit. IEEE Trans. Ind. Appl. **31**(2), 234–239 (1995)

[Lin08] Linder, S.: Potentials, limitations, and trends in high voltage silicon power semiconductor devices. In: Proceedings of the 9th ISPS, Prague, pp. 11–20 (2008)

[Lut97] Lutz, J.: Axial recombination centre technology for freewheeling diodes. In: Proceedings of the 7th EPE, Trondheim, p. 1.502 (1997)

[Lut03] Lutz, J., Domeij, M.: Dynamic avalanche and reliability of high voltage diodes. Microelectron. Reliab. **43**, 529–536 (2003)

[Lut08] Lutz, J., Herrmann, T., Feller, M., Bayerer, R., Licht, T., Amro, R.: Power cycling induced failure mechanisms in the viewpoint of rough temperature environment. In: Proceedings of the 5th International Conference on Integrated Power Electronic Systems, pp. 55–58 (2008)

[Lut09] Lutz, J., Baburske, R., Chen, M., Heinze, B., Felsl, H.P., Schulze, H.J.: The nn^+-junction as the key to improved ruggedness and soft recovery of power diodes. IEEE Trans. Electron Dev. **56**(11), 2825–2832 (2009)

[Lut09b] Lutz, J., Döbler, R., Mari, J., Menzel, M.: Short circuit III in high power IGBTs. In: Proceedings EPE, Barcelona (2009)

[Mor00] Mori, M., Kobayashi, H., Yasuda, Y.: 6.5 kV ultra soft & Fast Recovery Diode (U-SFD) with high reverse recovery capability. In: Proceedings ISPSD 2000, Toulouse, pp. 115–118 (2000)

[Mor07] Mori, M., et al.: A planar-gate high-conductivity IGBT (HiGT) with hole-barrier layer. IEEE Trans. Electron Dev. **54**(6), 1515–1520 (2007)

[Mou88] Mourick, P.: Das Abschaltverhalten von Leistungsdioden, Dissertation, Berlin (1988)

[Mue15] Münster, P., Wigger, D., Eckel, H.G.: Impact of the dynamic avalanche on the electrical behavior of HV-IGBTs. In: Proceedings of the PCIM Europe 2015, Nuremberg, pp. 906–915 (2015)

[Mue16] Münster, P., Lexow, D., Eckel, H.G.: Effect of Self Turn-ON during turn-ON of HV-IGBTs. In: Proceedings of the PCIM Europe 2016, pp. 924–931 (2016)

[Nag98] Nagasu, M. et al.: 3.3 kV IGBT modules having soft recovery diodes with high reverse recovery di/dt capability. In: Proceedings of the PCIM 98 Japan, 175 (1998)

[Nic00] Nicolai, U., Reimann, T., Petzoldt, J., Lutz, J.: Application Manual Power modules. ISLE Verlag (2000)

[Nie04] Niedernostheide, F.J., Falck, E., Schulze, H.J., Kellner-Werdehausen, U.: Avalanche injection and current filaments in high-voltage diodes during turn-off. In: Proceedings of the 7th ISPS'04, Prague (2004)

[Nie05] Niedernostheide, F.J., Falck, E., Schulze H.J., Kellner-Werdehausen, U.: Periodic and traveling current-density distributions in high-voltage diodes caused by avalanche injection. In: Proceedings of the EPE (2005)

[Oet00] Oetjen, J., et al.: Current filamentation in bipolar devices during dynamic avalanche breakdown. Solid State Electron. **44**, 117–123 (2000)

[Ohi02] Ohi, T., Iwata, A., Arai, K.: Investigation of gate voltage oscillations in an IGBT module under short circuit conditions. In: Proceedings of the 33rd Annual Power Electronics Specialists Conference, IEEE, vol. 4, pp. 1758–1763 (2002)

[Pen98] Pendharkar, S., Shenai, K.: Zero voltage switching behavior of punchthrough and nonpunchthrough insulated gate bipolar transistors (IGBT's). IEEE Trans. Electron Dev. **45**(8), 1826–1835 (1998)

[Pog03] Pogany, D., Bychikhin, S., Gornik, E., Denison, M., Jensen, N., Groos, G., Stecher, M.: Moving current filaments in ESD protection devices and their relation to electrical characteristics. In: Annual Proceedings—Reliability Physics Symposium, pp. 241–248 (2003)

[Por94] Porst, A.: Ultimate limits of an IGBT (MCT) for high voltage applications in conjunction with a diode. In: Proceedings of the 6th ISPSD (1994)

[Rah04] Rahimo, M., Kopta, A., et al.: Switching-Self-Clamping-Mode "SSCM", a breakthrough in SOA performance for high voltage IGBTs and diodes. In: Proceedings of the ISPSD, pp. 437–440 (2004)

[Rah05] Rahimo, M., et al.: A study of switching-self-clamping-mode "SSCM" as an over-voltage protection feature in high voltage IGBTs. In: Proceedings of ISPSD, Santa Barbara (2005)

[Ron97] Ronan, H.R.: One equation quantifies a power MOSFETs UIS rating. In: PCIM Magazine, August 1997, pp. 26–35 (1997)

[Ros02] Rose, P., Silber, D., Porst, A., Pfirsch, F.: Investigations on the stability of dynamic avalanche in IGBTs. In: Proceedings of the ISPSD (2002)

[Sai04] Saint-Eve, F., Lefebvre, S., Khatir, Z.: Study on IGBT lifetime under repetitive short-circuits conditions. In: Proceedings of the PCIM Europe, Nuremberg (2004)

[Sco89b] Schlangenotto, H., Neubrand, H.: Dynamischer Avalanche beim Abschalten von GTO-Thyristoren und IGBTs. Arch. Elektrotech. **72**, 113–123 (1989)

[Shi59] Shields, J.: Breakdown in silicon pn-junctions. J. Electron. Control **6**, 132ff (1959)

[Sil73] Silber, D., Robertson, M.J.: Thermal effects on the forward characteristics of silicon pin-diodes at high pulse currents. Solid State Electron. **16**, 1337–1346 (1973)

[Sit02] Sittig, R.: Siliziumbauelemente nahe den Grenzen der Materialeigenschaften. ETG Fachtagung, Bad Nauheim, ETG Fachbericht **88**, 9 ff (2002)

[SYN07] Advanced TCAD manual. Synopsys Inc. Mountain View, CA. Available: http://www.synopsys.com (2007)

[Tan15] Tanaka, M., Nakagawa, A.: Simulation studies for avalanche induced short-circuit current crowding of MOSFET-Mode IGBT. In: Proceedings ISPSD '15, pp. 121–124 (2015)

[Tan16] Tanaka, M., Nakagawa, A.: Growth of short-circuit current filament in MOSFET-Mode IGBTs. In: Proceedings of ISPSD '16, pp. 319–322 (2016)

[Tom96] Tomomatsu, Y., et al.: An analysis and improvement of destruction immunity during reverse recovery for high voltage planar diodes under high dI_{rr}/dt condition. In: Proceedings of the ISPSD, pp. 353–356 (1996)

[Wac91] Wachutka, G.: Analytical model for the destruction mechanism of GTO-like devices by avalanche injection. IEEE Trans. Electron Dev. **38**, 1516 (1991)

[Wak95] Wacker, A., Schöll, E.: Criteria for stability in bistable electrical devices with S- or Z-shaped current voltage characteristic. J. Appl. Phys. **78**(12), 7352–7357 (1995)

Chapter 14
Power Device Induced Oscillations and Electromagnetic Disturbances

14.1 Frequency Range of Electromagnetic Disturbances

Every power electronic switching action results in a deviation from the ideal sinusoidal AC current or the ideal homogeneous DC current. Switching events are usually done periodically in time. Every periodic event can be separated into a row of sinus and cosinus terms by means of Fourier transformation. With this tool, the generated frequencies, the harmonics and their intensity can be calculated.

Figure 14.1 shows a rough overview of the disturbances and oscillations caused by power electronics. It is distinguished between disturbances created by switching events in power-electronic converters, the harmonics of the switching frequency on the low- and medium-frequency range, and the device-induced high-frequency disturbances.

At low frequencies, e.g. in phase-commutated converters, the disturbances caused by the input rectifier occur as multiples of the grid frequency of 50 – 60 Hz, their intensity declines proportionally to 1/n. In self-commutated converters with modern power devices, the typical switching frequencies for converters using IGBTs are in the range of 1 – 20 kHz; also in this case, harmonics of the respective switching frequency are to be expected. Higher switching frequencies can be applied with MOSFETs as power devices. In switch-mode power supplies, nowadays frequencies up to 1 MHz and above have become possible.

Device-induced oscillations, on the other hand, result from switching events. Since the switching times of the devices are much smaller than the period of the switching frequency, the electromagnetic disturbances caused thereby occur in a frequency range significantly higher than the switching frequency.

© Springer International Publishing AG 2018
J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8_14

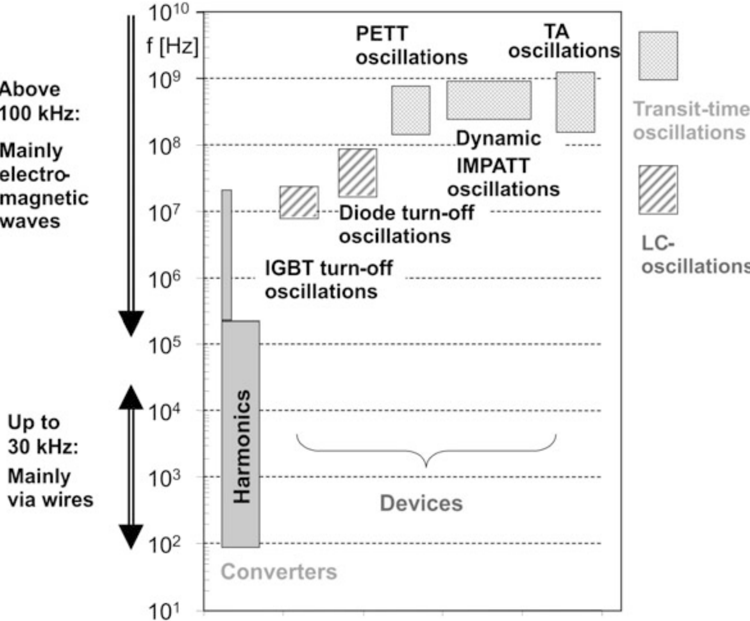


Fig. 14.1 Typical frequency ranges of disturbances caused by power-electronic effects

Electromagnetic disturbances of a frequency < 30 kHz spread mainly via wiring and cables. They disturb the electric grid in the form of grid feedback. Electromagnetic disturbances of frequencies > 100 kHz are radiated off to a large extent as electromagnetic waves, and this can cause incompatibility with other electronic and power-electronic equipment.

Harmonics

Figure 14.2 shows two simplified examples of electric signals, either depicting the current or the voltage curve.

For a rectangular course with point symmetry to π and the amplitude a, as shown in Fig. 14.2a, the harmonics can be calculated using the Fourier transformation

$$y = \frac{4a}{\pi}\left(\sin\,\omega t + \frac{1}{3}\sin\,3\omega t + \frac{1}{5}\sin\,5\omega t + \cdots\right) \tag{14.1}$$

Multiples of the switching frequency $f = \omega/2\pi$ are generated, i.e. the 3rd, 5th, 7th, … harmonics. With increasing ordinal number n, their intensity decreases proportionally to 1/n.

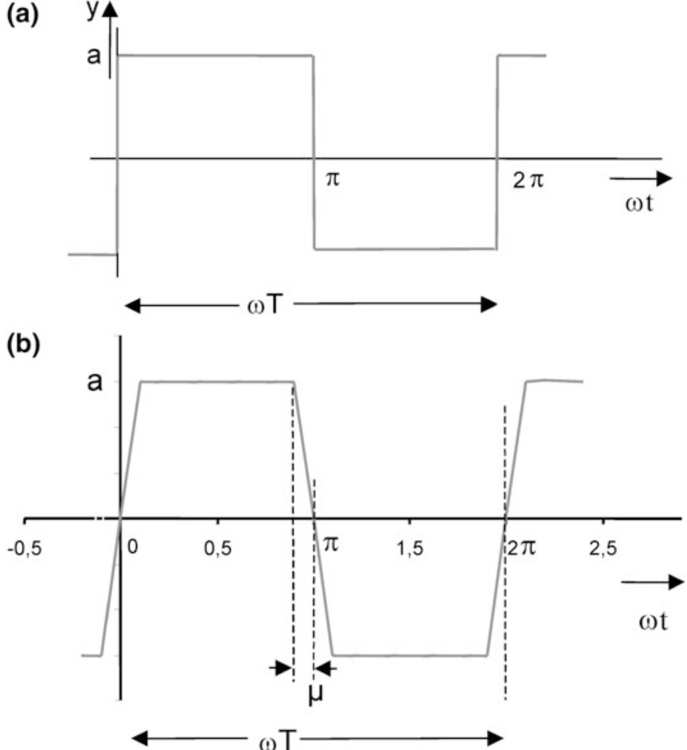


Fig. 14.2 **a** Rectangular course of a signal. **b** Trapezoidal course of a signal

However, for a trapezoidal course with point symmetry to π and the amplitude a as shown in Fig. 14.2b, the Fourier transformation results in

$$y = \frac{4}{\pi}\frac{a}{\mu}\left(\sin\,\mu\,\sin\,\omega t + \frac{1}{3^2}\sin\,3\mu\,\sin\,3\omega t + \frac{1}{5^2}\sin\,5\mu\,\sin\,5\omega t + \cdots\right) \tag{14.2}$$

The amplitudes of the harmonics decrease proportionally to $1/n^2$. For a non-symmetric shape, additional terms will occur. Nevertheless, the faster decline of the harmonics and therefore a trapezoidal course is much more suitable. The slopes *di*/*dt* – represented in Fig. 14.2 with μ – can be adjusted by gate resistors if MOSFETs and IGBTs are used. To reduce harmonics, the switching times are reduced by increased gate resistors. However, this will increase switching losses. A trade-off must be made in many applications between switching losses on the one hand and electromagnetic emissions on the other hand.

Suitable filters are implemented as a countermeasure to improve the electromagnetic compatibility. This shall not be explained further in this chapter; the attention is riveted instead on oscillations created by power devices themselves.

14.2 LC Oscillations

14.2.1 Turn-off Oscillations with IGBTs Connected in Parallel

In power modules, often a lot of single dies are connected in parallel. It is very difficult to give all single dies identical symmetrical conditions in respect to the length of the current-conducting path to the main terminals as well as in respect to the length of the wiring of the drive signals. Thermal conditions must be considered, too. Often trade-offs must be made. Figure 11.33 shows an example of parallel connection of five IGBT dies, in which asymmetric tracks for the main current to the respective die are given. The wires for the drive signals, for which also a symmetrical setup is of importance, are not drawn in the schematic circuit diagram of 11.33b.

Figure 14.3 shows the measurement of a parallel connection of two IGBTs. To create differences, the gate resistor are chosen slightly deviating: for chip 1, a resistor of 6.02 Ω, and for chip 2 of 6.45 Ω [Pal99]. The IGBT chips rated to 100 A are exposed to a forward current of 20 A in each chip. According to the lower gate resistor, chip 1 starts with the turn-off process. This has the consequence that the current in chip 2 increases in the first instance, while the total current is still at the same level. Then, during the current decline, an oscillation between the two chips builds up. A period of 50 ns can be read, which corresponds to a frequency of 20 MHz. The oscillations cannot be seen in the course of the total current in this case. Only the measurement of the single currents shows the oscillations occurring between the chips.

Figure 14.4a shows turn-off oscillations in a first version of a press-pack IGBT housing, see also Fig. 11.4. The current in single chips of the many chips connected in parallel is measured. The current shows a high-frequency oscillation in the range of 10 MHz.

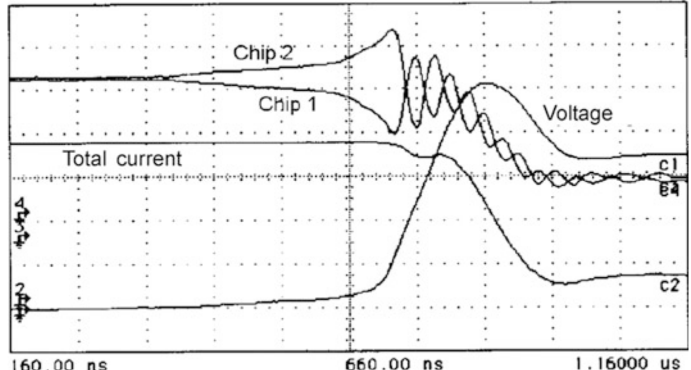


Fig. 14.3 Oscillations of the current in two IGBTs connected in parallel. Gate resistors 6.02 Ω for chip 1 and 6.45 Ω for chip 2. Current 10 A/div, voltage 50 V/div. Figure according to [Pal99] © 1999 EPE

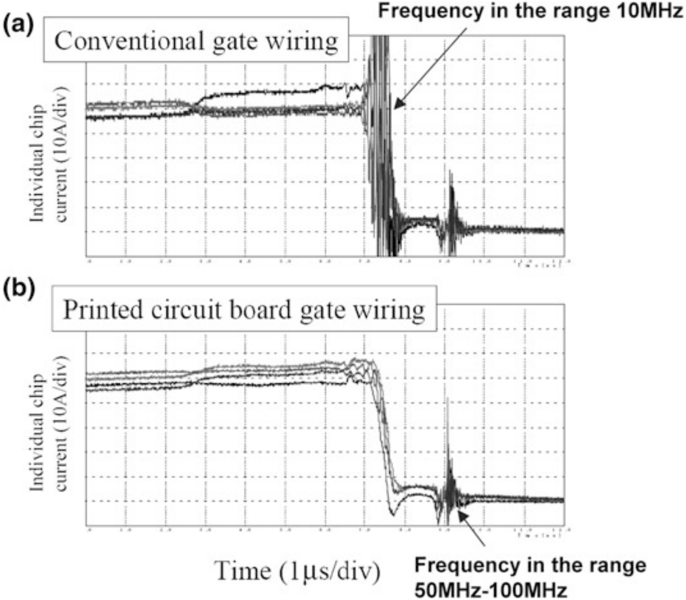


Fig. 14.4 **(a)** Turn-off oscillations in a press-pack IGBT and **(b)** their elimination by a symmetrical arrangement of the gate signal connections with an integrated PCB. Figure from [Omu03] © 2003 IEEE

If the control terminals are realized with a printed circuit board (PCB) which leads to the same conditions for each of the 42 chips connected in parallel, the turn-off oscillations can be eliminated, as can be seen in Fig. 14.4b. Such a PCB, shown in Fig. 14.5, has copper layers on both sides; one side is the potential of the gate signal, the other side is the potential of the control emitter terminal. Very close to every single chip, gate resistors are arranged on the PCB.

The frequency range in which turn-off oscillations are found is in the range between 10 and 20 MHz. This is clearly above the values which are to be expected for the harmonics of switching events (compare Fig. 14.1). Turn-off oscillations must be avoided not only because of electromagnetic emissions. They can additionally increase the turn-off losses of chips and can lead to thermal failures.

Possible countermeasures against turn-off oscillations are:

– The setup of an arrangement as symmetrical as possible. This is also to be considered for parallel connection of discrete devices as well as single modules.
– If this is not possible because of mechanical and other reasons, the gate resistors R_G of IGBTs and MOSFETs can be increased. This counteracts oscillations, but in the same way it increases the turn-off losses. For IGBTs, compare Eq. (10.6) and the discussion of the influence of gate resistors on example of MOSFETs in context with Fig. 9.19.

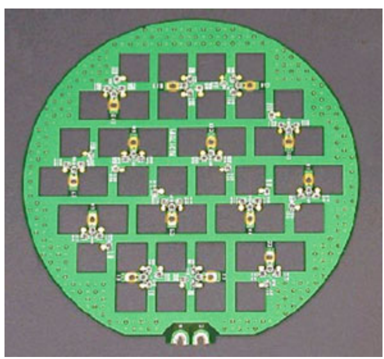


Fig. 14.5 PCB to ensure symmetrical control terminals in a press-pack IGBT. Figure from [Omu03] © 2003 IEEE

14.2.2 Turn-off Oscillations with Snappy Diodes

Fast diodes with insufficient reverse-recovery behavior more often cause oscillations in power circuits than asymmetries in IGBTs connected in parallel. For details on snappy recovery behavior see Chap. 5. Figure 14.6 shows the course of the voltage at turn-off of a snappy diode in a circuit according to Fig. 5.19. The application is a step-down converter in a battery-fed electric vehicle. Instead of the IGBT in Fig. 5.19, a 100-V rated MOSFET is used as power switch. The snap-off of the reverse current in the diode leads to a voltage peak of 100 V. The abrupt snap-off leads to a current overshoot in forward direction, the diode is turned off again, a second and a third voltage peak are generated, and finally the effect ends up in a damped LC oscillation.

The frequency of the LC oscillation generated by snappy diodes is determined by the device capacity C_j and the parasitic inductance L_{par}

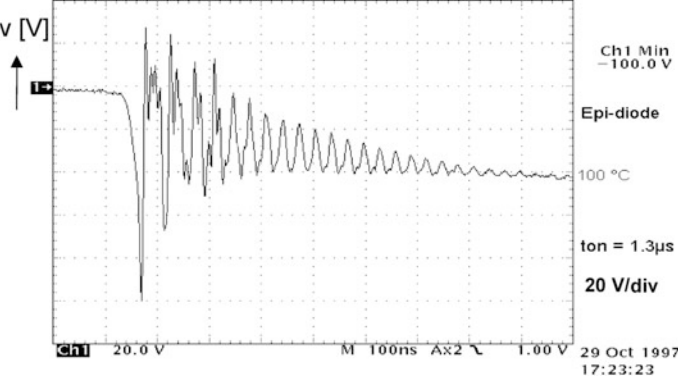


Fig. 14.6 Course of the voltage at turn-off of a snappy diode. Period 30 ns, frequency 33 MHz

$$f = \frac{1}{2\pi}\sqrt{\frac{1}{L_{par}\cdot C_j}} \tag{14.3}$$

The equivalent electrical circuit for such an oscillator is shown in Fig. 14.7 [Kas97]. It must be noted that the capacity $C_j = c_j{\cdot}A$ is dependent on the voltage, see Eq. (3.113). The resistance R_b of the interconnections of the diode acts as damping component; the same holds for $R_{n,p}$ which stands for the base of the diode. During the turn-off process, electrons and holes are removed from the base and $R_{n,p}$ is not linear; in fact, it is strongly varying.

During the LC oscillation, C_j is not constant. For a diode rated to a blocking voltage of 1200 V with a steep doping profile of the p-layer at the pn-junction, C_j can be assumed to be 250 pF/cm^2 for the estimation of the expected frequency of oscillations [Kas97]. Considering the area of the diodes and some typical housings and their respective parasitic inductance, the values given in Table 17 are obtained Table 14.1.

For a diode rated to 100 A and 1200 V in an IGBT module of older construction type with a typical parasitic inductance of 100 nH, the frequency of LC oscillations caused by snappy diodes is expected to be in the range of 48 MHz. With a module of modern architecture, approximately 20 nH are given as parasitic inductance; the frequency moves to a range of 100 MHz. In a high-power module rated to 1200 V, 12 diodes rated 100 A are connected in parallel and a much larger capacity is given. Because of the larger volume of the module, longer interconnections are necessary. Despite of this, the parasitic inductance in such modules has been reduced significantly. Depending on the inductance, frequencies in the range of 5–15 MHz are to be expected for the discussed oscillations.

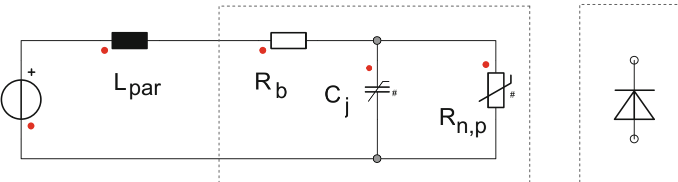


Fig. 14.7 Equivalent circuit for an oscillator consisting of diode and its parasitic components. Figure from [Kas97]

Table 14.1 Estimation of the frequency range of LC oscillations for 1200-V free-wheeling diodes

	C_j	L_{par} (nH)	f (MHz)	$T = 1/f$ (ns)
Bipolar 100-A diode, active area 0.44 cm^2	110 pF	20	107	9.3
		100	48	20.8
1200-A module, diodes	1.32 nF	100	13.9	72
		800	4.9	204

If soft recovery diodes are used, the described oscillations can be avoided. Soft recovery diodes are producible today and they are indispensable for the operation of a modern power-electronic converter.

However, it must be noted that not every oscillation in conjunction with diodes is originated by snappy turn-off behavior. If a non-suited unsymmetrical parallel connection is done, oscillations may occur even if soft-recovery diodes are used. An example is given in Fig. 14.8 [Eld98]. In this example, three diodes are connected in parallel; the diode D1 is located close to the main terminals, D2 and D3 with additional wiring are connected in parallel. The diode D1 with the lowest inductance is commutated with the highest *di/dt*. The reverse-recovery current maximum is first reached in D1. The process continues in D2, and then in D3, with an increased commutation velocity di/dt. It is maximal for D3, because the reverse current is already declining in D1 and D2. During the reverse-recovery process, the current oscillates between the devices. Finally, at the end of commutation, these internal oscillations superimpose to an oscillation of the total current i_D, too.

If a parallel connection is given and oscillations in the reverse recovery of free-wheeling diodes are found, it also has to be investigated whether this is caused by asymmetrical arrangement of the wirings and interconnections.

Oscillations have even been found with a single soft-recovery diode, if the condition is fulfilled that the switching time of the diode t_{rr} agrees with half the period of the resonance of an LC oscillator circuit. In applications with MOSFETs and IGBTs, this can be verified easily: the modification of the gate-resistor R_G of the used IGBT or MOSFET can vary the turn-on slope of the transistor and with it the commutation velocity *di/dt*. So the switching time t_{rr} of the diode is modified, and in this case the oscillations should vanish.

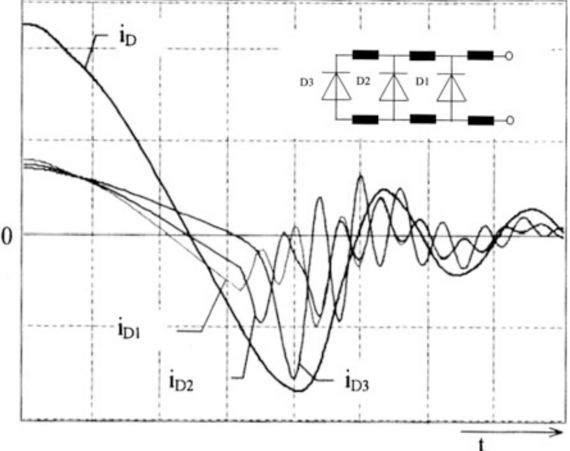


Fig. 14.8 Course of the current during reverse recovery in a parallel connection of diodes with different wiring inductances. 50 ns/div, 50 A/div, 25 °C, V_{bat} approx. 300 V. Figure from [Eld98]

14.2.3 Turn-off Oscillations with Wide Bandgap Devices

If just Si is replaced by SiC in high-current applications, huge problems with oscillations may occur. This is reported for 300 A/1200 V SKiM power module application for a 70 kW inverter with 30 kHz switching frequency [Win12], where Si bipolar freewheeling diodes were replaced with SiC Schottky diodes.

The oscillation is an LC-oscillation between the junction capacity of the Schottky-Diode and the parasitic inductance, where C_j is

$$C_j = \sqrt{\frac{q \cdot \varepsilon \cdot N_D}{2 \cdot (V_j + V_r)}} \cdot A \tag{14.4}$$

and has its maximum $C_j(0\ V)$ when the applied voltage V_r is zero. The value $C_j(0\ V)$ is typically given in the data sheet. Due to the two decades increased N_D^+ with SiC for the same voltage (Fig. 3.19), C_j will be one decade higher in the case of the same active area. The current peak $I_{D,peak}$ caused by the capacity of the Schottky-diode is given by

$$I_D,peak = C_j(0V) \cdot \frac{dv}{dt} \tag{14.5}$$

with dv/dt as voltage slope created by the IGBT. As can be deduced from Fig. 14.9, dv/dt is about 30 V/ns. The Schottky diode used for the measurement in Fig. 14.9 had a capacity $C_j(0\ V)$ of about 754 pF which gives a current peak $I_{D,\ peak}$ of about 90 A, this agrees well with the measured result.

To reduce the oscillations in Fig. 14.9, the gate resistor of the IGBT had to be increased. This leads to increased IGBT turn-on losses. Finally, all advantages of the low-loss SiC diodes were lost if the requirements for low electromagnetic emission are to be fulfilled. Note, that the used SKIM module was already developed for low parasitic inductance. However, for the SiC Schottky diode application it is not sufficient. This shows that high current SiC diode applications are only possible if more effort is taken in new low-inductive packages.

A tail current, which occurs in Si bipolar devices and which damps oscillations, is missing with SiC. This is a challenge for a new packaging technology. Applying wide bandgap devices results in a strong need for low parasitics in the package. An example for acceptable switching behavior of SiC-MOSFETs in a low inductive housing is shown in Fig. 14.10.

The device package also contains, besides the known chip internal capacities C_{gs}, C_{gd} and $C_{ds,}$ the capacities formed by the substrate insulator layers, e.g. Al_2O_3 or AlN, which are displayed in Fig. 14.11b as $C_{\sigma+}$, $C_{\sigma out}$ and $C_{\sigma-}$. Another point to consider is that $C_{\sigma out}$ is recharged with every switching event, and an undesired current is supplied into the heat sink. If the two L_σ and the $C_{\sigma+}$, and $C_{\sigma-}$ are unbalanced, they also generate a current into the heat sink [Fei15].

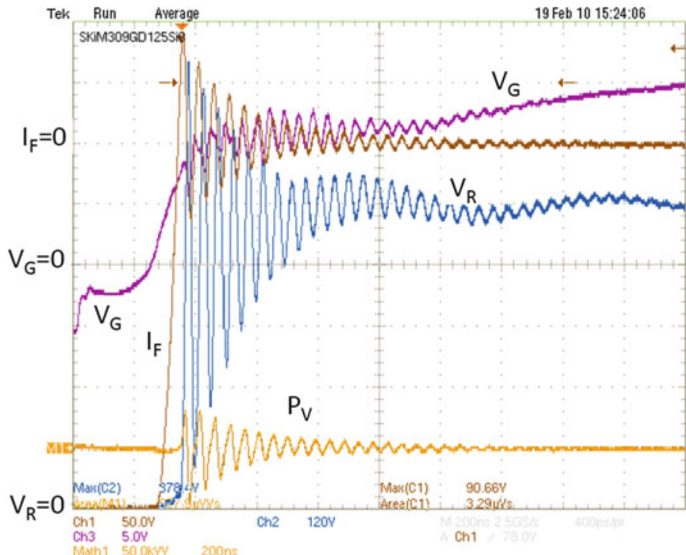


Fig. 14.9 Oscillations at turn-on of an IGBT with SiC Schottky diodes as freewheeling diode. V_R Voltage over the Schottky diode 120 V/div, I_F inverted diode current 50 A/div, V_G gate voltage of the IGBT 5 V/div, 125 °C. Figure from [Win12]

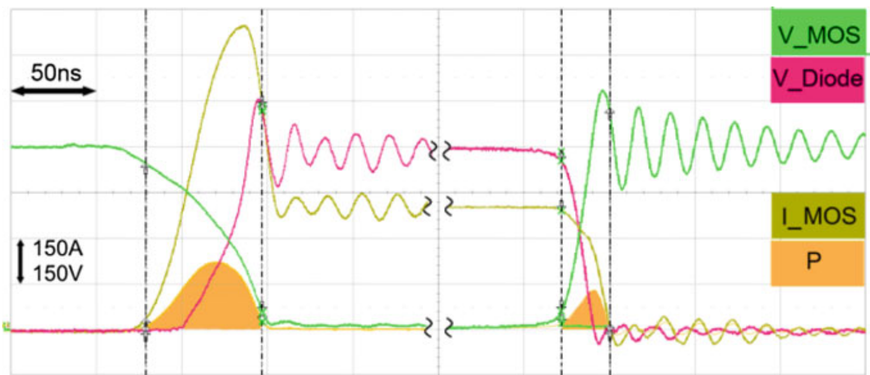


Fig. 14.10 SiC MOSFET with body diodes, 400A 600 V 150 °C dv/dt 40 V/ns. Figure from [Bec16]

Especially the parasitic inductance L has to be very low in a package for SiC and GaN. GaN as well, with the capability of high dv/dt and di/dt, needs low inductive packaging. There is much effort invested today in the development of packages with very low inductance. Special care has to be taken in case of paralleling, since non-symmetric inductivities in front of parallel connected chips will lead to bad dynamic current sharing and internal LC-oscillations, see Fig. 14.8.

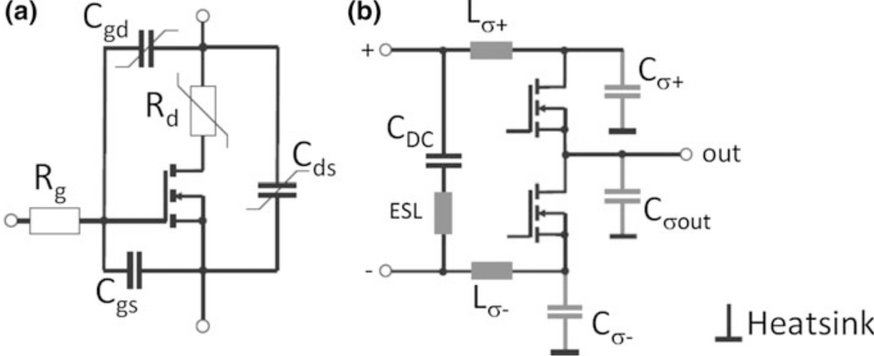


Fig. 14.11 Parasitic elements of a SiC MOSFET **(a)** In the chip **(b)** Switching cell parasitics. Figures from [Fei15] PCIM Europe 2015

14.3 Transit-Time Oscillations

A power device has a lowly doped middle layer with a thickness w_B At turn-off of a bipolar device, the existing charge carriers are removed; a part of them at an instant at which a space charge has already built up. The charge carriers flow through the space charge with at a drift velocity v_{sat} This results in a transit time for which a first approximation can be given with

$$t_T = \frac{w}{v_{sat}} \tag{14.6}$$

The drift velocities for electrons and holes $v_{sat(n)}$ and $v_{sat(p)}$ are given in Eq. (2.38), the width of the space charge w during switching processes is smaller or equal to the base with w_B. The carrier transit time corresponds to a frequency of $1/t_T$. Dependent on the base width, transit-time oscillations occur with a frequency in this range; it is the range between 100 MHz and 1 GHz or even higher, see Fig. 14.1. The occurring frequency depends on the type of effect and on its phase relations. This is shown in the following.

Transit-time oscillations are used for the fabrication of microwave devices which are used as microwave oscillators [Sze81]. With power devices, such oscillations must be avoided since they are a danger for the power device itself, and since their electromagnetic emission can cause unwanted effects in driver circuits and other electronic components of the adjacent environment. The occurrence of transit-time oscillations was observed for three effects in power devices. They are described in the following paragraphs, and measures to avoid them are discussed.

14.3.1 Plasma-Extraction Transit-Time (PETT) Oscillations

PETT oscillations can occur in a bipolar device at turn-off during the tail current interval. They have been observed with IGBTs as well as with soft-recovery free-wheeling diodes [Gut01, Gut02]. Figure 14.12 shows an example with IGBTs.

The oscillations are observed in the course of the gate voltage, but they occur primarily as oscillations of the collector current I_C and collector voltage V_C where they cannot be resolved easily because of their low amplitude. Therefore, the stray effect into the gate signal is mainly observed. The oscillations occur after the device has taken up the voltage and the device is in the interval in which a tail current flows.

An example of the occurrence of PETT oscillations in a soft-recovery free-wheeling diode is shown in Fig. 14.13 [Sie03]. The measurement was executed in a test setup of an IGBT module rated to 600 A and 1200 V. At turn-on of the IGBT, the free-wheeling diode is turned off, the PETT oscillations occur in the tail current interval of the free-wheeling diode. Only soft-recovery diodes can lead to PETT oscillations. However, soft recovery is essential for fast diodes in such applications. Since the resolution of the current measurement of the anode current of the diode is too low, the PETT oscillation is detected in this example with a wire loop which works as antenna and which is placed close to the diode.

The mechanism of oscillation is related to the mechanism of the Barrier Injection Transit Time (BARITT) diode which is intended as microwave oscillator. The BARITT diode has a metal-semiconductor-metal structure or a pn$^-$p-structure. At a voltage applied in blocking direction, the electric field reaches the opposite metal-semiconductor junction or the opposite p-layer, and it releases an injection of carriers there [Sze81]. In contrast to the BARITT effect, with the PETT effect the space charge reaches the remaining charge carrier hill or remaining plasma zone which is still a reservoir of free carries and feeds the tail current.

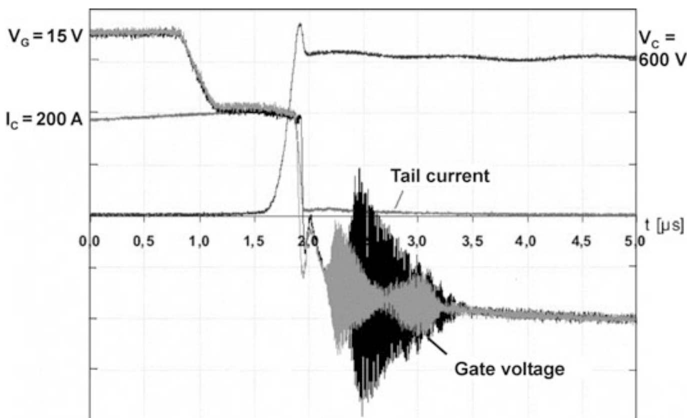


Fig. 14.12 PETT oscillations at turn-off of an IGBT. Figure from [Gut01] © 2001 IEEE

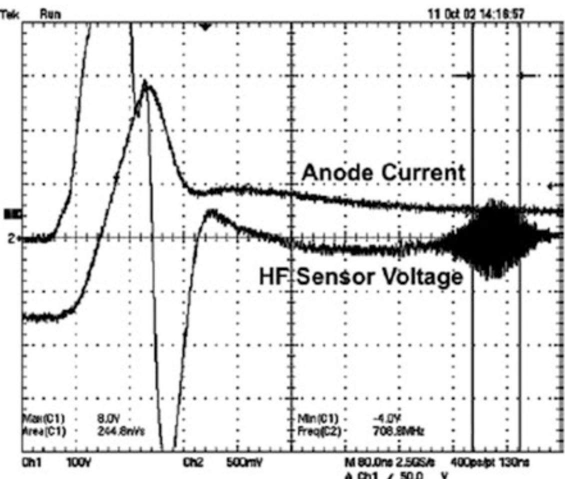


Fig. 14.13 PETT oscillation in the tail current of a soft-recovery free-wheeling diode at turn-on of an IGBT. V_{bat} = 600 V, I_F = 200 A, *di*/*dt* = 4000 A/µs, T = 300 K. Figure from [Sie03] © 2003 EPE

Figure 14.14 shows the process in the device at a PETT oscillation on example of a freewheeling diode. At the anode side (Fig. 14.14, right-hand side), an electric field has built up during the turn-off event; it takes up the voltage.

The charge carrier hill which feeds the tail current is still located at the cathode side (Fig. 14.14, left-hand side), compare Fig. 5.25 or 5.33. The tail current flows through the space charge as hole current. The shape of the electric field is triangular in this interval.

The occurrence of oscillations under such conditions is discussed in detail in [Eis98]. For a PETT oscillation, this is shown in Fig. 14.15. A high-frequency AC voltage $V_{RF} \cdot \sin\omega t$ superimposed to the DC link voltage V_{DC} is assumed (Fig. 14.15a). In the same way as in the BARITT effect, an injected j_{inj} current is generated as the AC voltage has its maximum at $\omega t = \pi/2$ (Fig. 14.15b). This injected current flows through the space charge region with the velocity v_d. The corresponding current density at the terminals of the device j_{inf} is expressed by the

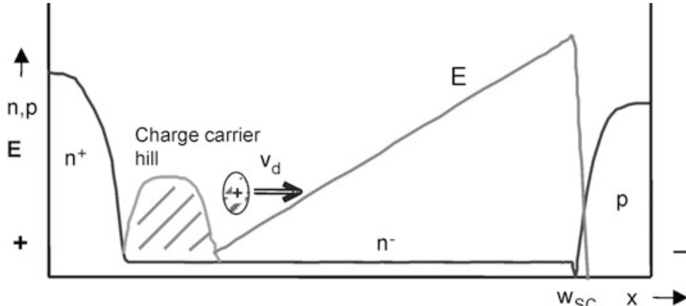


Fig. 14.14 Process in a freewheeling diode at occurrence of a PETT oscillation

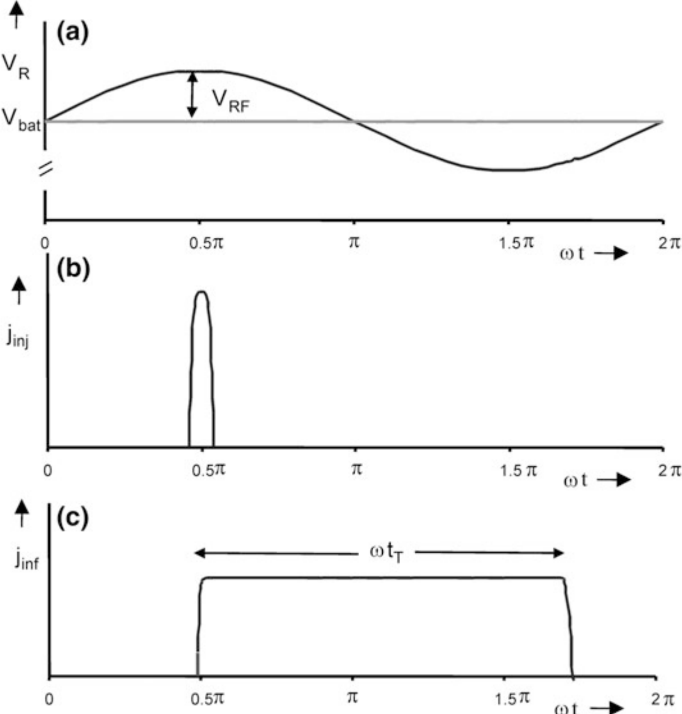


Fig. 14.15 Origin of PETT oscillations: **(a)** High-frequency AC voltage superimposed to the applied DC voltage **(b)** Injection current at $w = w_P$ at the time $\omega t = \pi/2$. **(c)** Current at device terminals. Figure from [Sie06b] © 2006 IEEE

Ramo-Shockley-theorem [Eis98] and shown in Fig. 14.15c. The current flow at the device terminals starts at $\omega t = \pi/2$ and is found in the time interval ωt_T, needed by the carriers to transit the space charge region. The generated RF power is given by

$$P_{RF} = \frac{A}{2\pi} \int_0^{2\pi} j_{\inf}(\omega t) \cdot V_{RF} \sin \omega t d\omega t \tag{14.7}$$

In Figs. 14.15a and c one can see: P_{RF} is positive for $\omega t_T < \pi$, zero at $\omega t_T = \pi$ and negative at $\omega t_T > \pi$. A negative value of P_{RF} means that RF power is generated. The created RF power is maximal for $\omega t_T = 3\pi/2$ and decreases again for $\omega t_T > 3\pi/2$. As long as P_{RF} is of negative sign, the device acts as current source and emits *RF* power.

The transit time t_T is given by [Gut02]:

$$t_T = \int_{w_p}^{w_{SC}} \frac{1}{v_d(w)} dw \tag{14.8}$$

The velocity v_d in the space charge region depends on the strength of the electric field, compare Chap. 2, Eq. (2.39) and Fig. 2.15. The electric field is of a triangular shape in the given case [Sie06b] (see Fig. 14.11).

For the BARITT diode, the transit time is given in first-order approximation [Sze81] by Eq. (14.6) as $t_T = w_{SC}/v_{sat}$, where v_{sat} is the saturation drift velocity of holes under high-field conditions (approximately 10^7 cm/s in silicon). Note that the drift velocity v_d is lower than the saturation velocity v_{sat} for a significant part of the space charge region. Continuing with this simplification and taking into account the point of maximum RF-power generation at $\omega t_T = 3\pi/2$, the frequency of the PETT oscillations is approximated by

$$f_T = \frac{3 \cdot v_{sat}}{4 \cdot w_{sc}} \tag{14.9}$$

as it is given also in [Sze81].

From Fig. 14.15 it follows that excitation of the superimposed AC power is possible in a specific frequency range. It can be concluded that PETT oscillations occur only when the "negative-resistance" behavior found during one period is greater than all other resistive components in the complete circuit. Moreover, the phase shift between the oscillation voltage and the AC current at the device terminals is essential for an occurrence of this kind of oscillation. Furthermore, it can be concluded from Fig. 14.15 that the efficiency of the RF generation is low, since power is always dissipated during the interval $\pi/2$. Low efficiency is also a characteristic of BARITT diodes [Sze81]. Therefore, PETT oscillations only occur if there is a resonance circuit formed by the junction capacitance and the inductance of the bond wires close to the device, whose resonance frequency has to be close to the frequency f_T according to Eq. (14.9).

Additionally, PETT oscillations cannot occur as long as there is a high amount of remaining plasma in the device and the corresponding reverse current is high, since this stored charge acts as damping which hinders the occurrence of the oscillations. Further important parameters are the applied voltage V_{bat}, since this voltage determines w_{SC}, and the temperature, since the drift velocity is temperature-dependent. It is typical for PETT oscillations that they only emerge at very special conditions. If one deviates from these conditions, no more PETT oscillations can be found. Therefore, the possible occurrence of PETT oscillations is easily overlooked in the procedure of the qualification of a power module.

Simulations of the PETT effect have been done already in [Gut02] and [Sie06b]. Detailed simulations in [Fuj12] point out that the velocity of a packet of holes is not constant but varies with the electric field strength in the drift region.

PETT oscillations are radiated-off as electromagnetic waves. This radiation can lead to the effect that the power-electronic equipment violates the requirements of electromagnetic compatibility (EMC). The EMC requirements are fixed in different standards, e.g. by the European standard EN55011 (international standard IEC CISPR 11) [DIN00]. For details see the special literature in this field.

Figure 14.16a, b show an experiment of measuring the electromagnetic radiation of the 600-A 1200-V module, from which the oscillations are shown in Fig. 14.12 [Sie03]. Since no chamber shielded against external electromagnetic radiation was available, first a measurement of existing electromagnetic disturbance of the environment was necessary. Figure 14.16a shows the result of the environmental EMC measurement which lasted 2 h, followed immediately by the subsequent measurements. The different grey-colored bars shown in Fig. 14.16a mark frequency ranges used by broadcasting and telecommunication. Obviously, the largest interfering signals were caused by mobile communication equipment. This

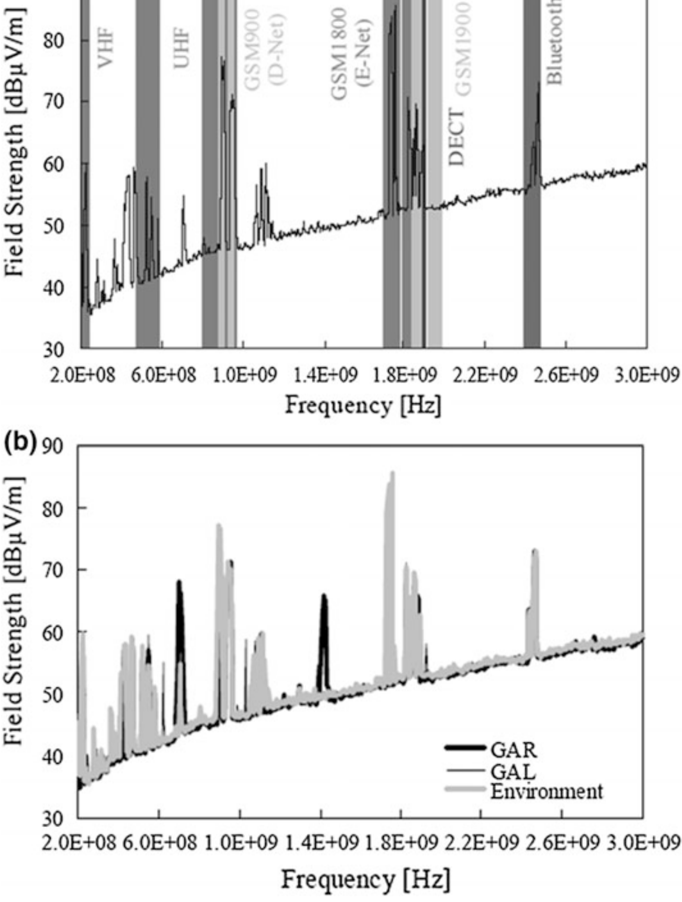


Fig. 14.16 EMC measurement of a PETT oscillation in comparison with environmental background radiation **(a)** Environment without operation of the module **(b)** with operation of the module PETT oscillations at the module GAR at 700 MHz and 1.4 GHz. Figure from [Sie06b] © 2006 IEEE

measurement series is included in Fig. 14.16b to give evidence of additionally generated signals that were caused by the PETT oscillations.

The PETT oscillation during the turn-off (marked as GAR) caused two sharp peaks in the frequency spectrum, appearing at 700 MHz and 1.4 GHz, which could be assigned to the fundamental frequency and the second harmonic.

The width of the space charge of the diode used in the module GAR amounts to approx. 85 μm for the given conditions. If the value of 8×10^6 cm/s is used as drift velocity for holes, a frequency of approx. 700 MHz results from Eq. (14.9), which agrees with the measured value.

Although the spurious radiation caused by PETT oscillation was relatively low in strength, exceeding the EMC limits could easily occur. In particular, this would be expected if more than one power module is used, as is typical for power-electronic equipment, and the radiations of the single modules are summed up.

In another application, PETT oscillations were found in a 1.8 MW high-frequency converter with an operating frequency of around 100 kHz. The setup of the equipment consisted of more than 100 power modules. The onset of the oscillation generated an error signal in the control unit [Sie06b]. Therefore, PETT oscillations must be avoided.

To prevent PETT oscillations, it is therefore not particularly helpful to modify the semiconductor device itself. Every power semiconductor has a space charge if voltage is applied, and it has therefore the capability for oscillations according to Eq. (14.9). However, it is essential to avoid an LC circuit that is in resonance with the transit frequency given in Eq. (14.9).

Figure 14.17 shows the setup of the module GAR, for which PETT oscillations in Figs. 14.12 and 14.17b were found. The oscillations occurred at the diode which was used as free-wheeling diode. It is marked as FWD in Fig. 14.17, left-hand side; in the equivalent circuit in Fig. 14.17 it is the diode at the bottom.

Figure 14.18 shows the impedance of this module on the left-hand side [Sie06b]. The three-dimensional EMC simulator FLO/EMC [FLO04] was used for the calculation. It solves the complete Maxwell equations numerically. The geometry of the module in Fig. 14.17 is used; the space is divided into cells modeled as the intersection of orthogonal transmission lines. There is no possibility to introduce real semiconductors into FLO/EMC. Therefore, a simplified model was used which reproduces the correct junction capacitance or on-state resistance of the devices (IGBT and FWD). For the characterization of the power modules, the excitation in the form of a delta pulse was applied across a FWD. In this way, the electric and magnetic fields and the resulting impedance can be calculated.

Figure 14.18 shows the simulation results for the impedance of the power module GAR (as seen from the FWD where the excitation was applied). A resonance point at a frequency of about 700 MHz resulted, which was in accordance with the oscillation frequency as given by the transit frequency f_T (Eq. 14.9) of the FWD. This resonance point is a necessary condition for the appearance of PETT oscillation, and 3D EMC simulation can be used to predict resonance points of a complex mechanical construction.

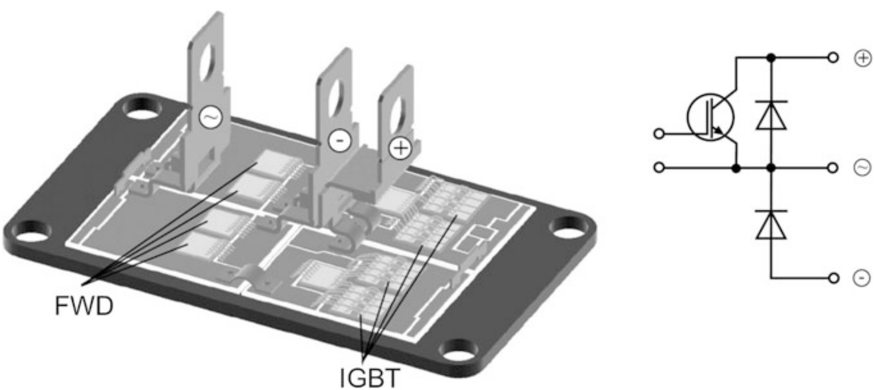


Fig. 14.17 Assembly of the module GAR which showed the PETT oscillations in the measurement in Fig. 14.12, and its internal electrical circuit. Figure from [Sie06b] © 2006 IEEE

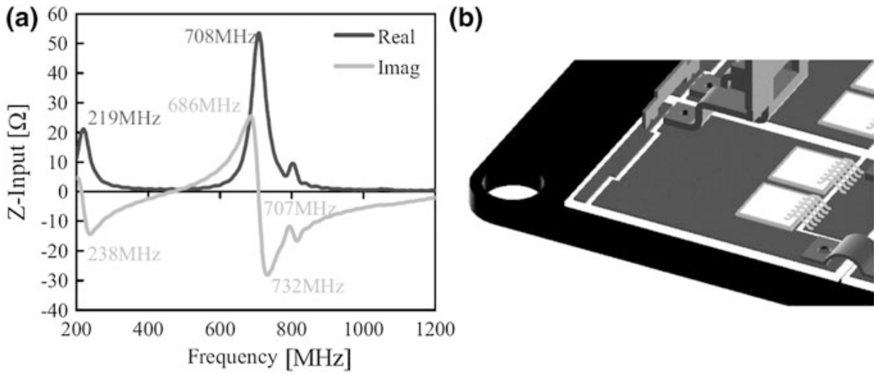


Fig. 14.18 Impedance of the module in Fig. 14.17 (a) Detail of the arrangement of the free-wheeling diodes at which the PETT oscillation occurs (b) Figure from [Sie06b] © 2006 IEEE

However, it must be mentioned that already small modifications in the assembly may shift the occurrence of PETT oscillations to other conditions, or may weaken it or even eliminate it.

The arrangement of the bond wires in the module GAR is shown in Fig. 14.18b. Considering the given chip area and the resulting space charge capacity, only a small inductance is possible for an LC circuit with resonance in the range of 700 MHz. It can be formed only by components in the immediate neighborhood of the chip, e.g. the next conductor paths on the DCB substrate, the bond wires and their arrangement.

An obvious and efficient way for lowering inductance would be to provide additional shorts between the anode contact areas, as published in an older patent application of 1995 [Zim95] and shown in Fig. 14.19b. This results in a clear suppression of the module resonance in the transit-frequency range, as shown in the

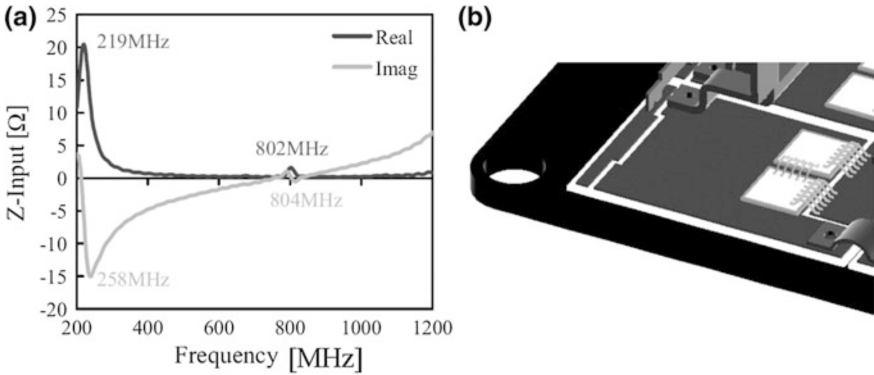


Fig. 14.19 Countermeasure against PETT oscillations: **(a)** Impedance of the module **(b)** detail of the arrangement of the free-wheeling diodes with shorting bond wires. Figure from [Sie06b] © 2006 IEEE

FLO/EMC simulation of this assembly in Fig. 14.19a [Sie06b]. No more resonance point is found in the range of 700 MHz.

The suggestion of [Zim95] is found to be an effective measure against PETT oscillations, in spite of the fact that no knowledge of details of the processes leading to PETT oscillations was available at that time.

To prevent PETT oscillations, the resonance point of the power module must be different from the transit-time frequency f_T governed by the power semiconductor structure. Three-dimensional EMC calculations are useful for this. Ideally, a future simulation system for such tasks should solve the full set of Maxwell equations for the complete construction, and calculate the behavior of the semiconductor devices, e.g. solve the basic semiconductor equations. Although the capabilities of computers steadily increase, this task still seems to be too complex at the moment.

14.3.2 Dynamic Impact-Ionization Transit-Time (IMPATT) Oscillations

IMPATT oscillations have been found as a dynamic oscillatory effect at turn-off of soft-recovery free-wheeling diodes [Lut98]. The attribute "dynamic" is used to indicate that these oscillations occur in connection with a switching event. The dynamic IMPATT oscillation is of high energy, it radiates off disturbances of high intensity and leads to malfunctions of analogue and digital electronic circuits which are present, for example, in drive circuits. A measurement of such an effect is possible in a laboratory setup; such a measurement is shown in Fig. 14.20.

The measurement in Fig. 14.20 was executed in an application in the form of a step-down circuit according to Fig. 5.19. The temperature of the diode was 0 °C. The course of the voltage is plotted inversely. The rated static blocking capability of

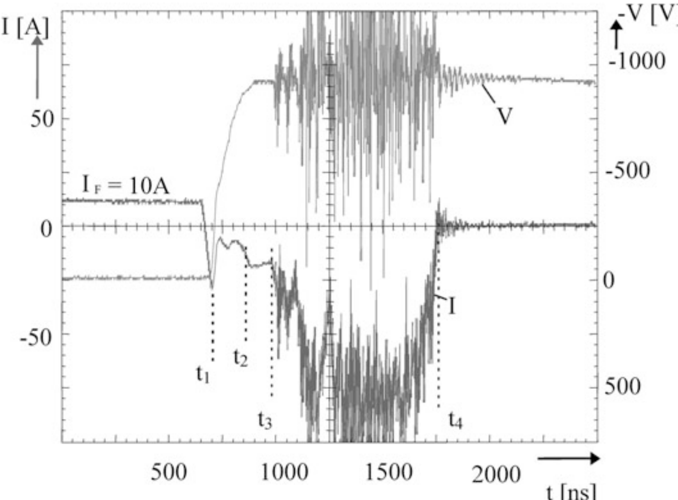


Fig. 14.20 Dynamic IMPATT oscillation of a free-wheeling diode with radiation-induced recombination centers. T = 0 °C. Reprinted from [Lut98] with permission from Elsevier

this diode is 1200 V, its avalanche breakdown voltage is > 1300 V. After the reverse-recovery current peak (time point t1) the reverse current decreases. Between t_1 and t_2, the voltage climbs up close to the value of the applied battery voltage, while a tail current flows in the diode. After t_2, a current hump grows from the tail current. This hump occurs only if the battery voltage is above 910 V. A further increase in the battery voltage to 930 V effects that a high reverse current suddenly shoots up from the current hump whose amplitude is a multiple of the reverse-recovery current peak. A high frequency oscillation is superimposed. After some 100 ns (t_4), the oscillation is finished.

A reduction of the voltage by only one ore two volts, or an increase of the temperature by 1 or 2°, removes the effect.

The mechanism of the dynamic IMPATT oscillation is related to that of the IMPATT diode, a device which is likewise intended as oscillator for microwaves [Sze81]. In IMPATT diodes, the static reverse-blocking capability is exceeded and the device is operated in the avalanche mode. In contrast, the dynamic IMPATT oscillation occurs at a voltage significantly lower than the avalanche breakdown voltage. The dynamic IMPATT oscillation is caused by the K-center, which is created at irradiation of the semiconductor with high-energy particles [Lut98]. The energy levels of the most important centers are shown in Fig. 4.30 in Chap. 4. The energy level of the K-center is located below the middle of the band gap. Recent investigations have found its nature as C_iO_i, a defect consisting of an interstitial carbon and oxygen atom [Niw08]. Its contribution to recombination is low, but its density is higher than that of the density of the OV-centers which determine the recombination in irradiated devices if typical annealing processes are used [Sie06].

The K-center has the characteristics of a temporary donor. At forward conduction, if the device is flooded with free carriers, it is occupied with a hole and it is positively charged. The effective doping then becomes

$$N_{eff} = N_D + N_T^+ \tag{14.10}$$

wherein N_T^+ is the density of positively charged K-centers. After the voltage has changed its polarity, the center is discharged:

$$N_T^+(t) = N_T^+ \cdot e^{-\frac{t}{\tau_{ep}}} \tag{14.11}$$

The time constant τ_{ep} of this discharge process is temperature-dependent; it is short (ca. 100 ns at 400 K) at high operation temperatures; it is in the order of some microseconds at temperatures below 300 K. This results in a temporarily increased doping of the device. The doping determines the onset voltage of avalanche breakdown V_{BD}; Eq. (3.84) can be used for the given situation. The increased doping N_{eff} leads to a strongly reduced value of V_{BD}. If now a fast-switching transistor such as an IGBT applies the battery voltage to the diode in a very short time after the current zero-crossing point of the diode, it finds a device with reduced avalanche voltage, and impact ionization sets on. In Fig. 14.20, the temporarily reduced voltage V_{BD} is reached at the time point t_2, and a current created by avalanche occurs as a hump in the current curve. If the device is driven into the avalanche mode even more, the dynamic IMPATT oscillation starts.

The situation in the device during this process is shown in Fig. 14.21. Between t_1 and t_2 a small tail current flows (see Fig. 14.20); in this state, a charge carrier hill from the remaining plasma still exists close to the cathode, and the shape of the electric field is triangular. Electron packets generated by impact ionization run through the electric field to the right-hand side.

A high-frequency AC voltage V_{RF}.sinωt shall be assumed as superimposed to the DC-link voltage V_{DC}, see Fig. 14.22a. It generates a current pulse j_{inj}. Since the

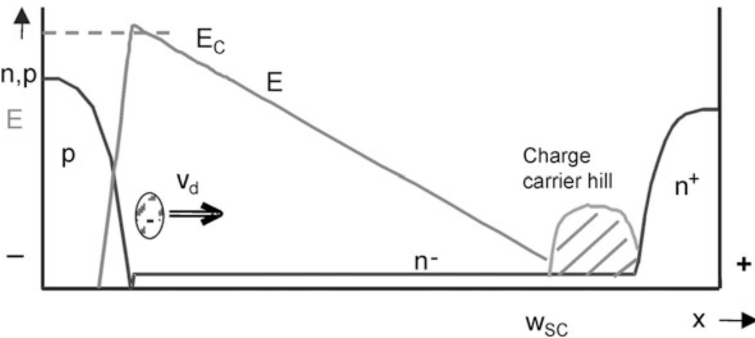


Fig. 14.21 Process in the device at a dynamic IMPATT oscillation

process of impact ionization needs time, the created current pulse is shifted by $\omega t = \pi/2$ to the peak value of the voltage and appears at $\omega t = \pi$, where the AC component of the voltage has its zero-crossing point (Fig. 14.22b). The injected current runs through the space charge at a velocity v_d. Figure 14.22c shows the corresponding current j_{inf} at the terminals of the device according to the Ramo-Shockley-theorem [Eis98]. The current appears at the terminals during the time interval ωt_T which starts at $\omega t = \pi$.

For the generated RF power P_{RF}, Eq. (14.7) holds. P_{RF} is of negative sign and is maximal for $\omega t_T = \pi$, it decreases again for $\omega t_T > \pi$. The diode acts as current source. It emits RF power. For the maximum of generated RF power at $\omega t_T = \pi$, one yields for the frequency of transit-time oscillation at an IMPATT effect [Sze81]

$$f_T = \frac{v_{sat}}{2 \cdot w_{sc}} \tag{14.12}$$

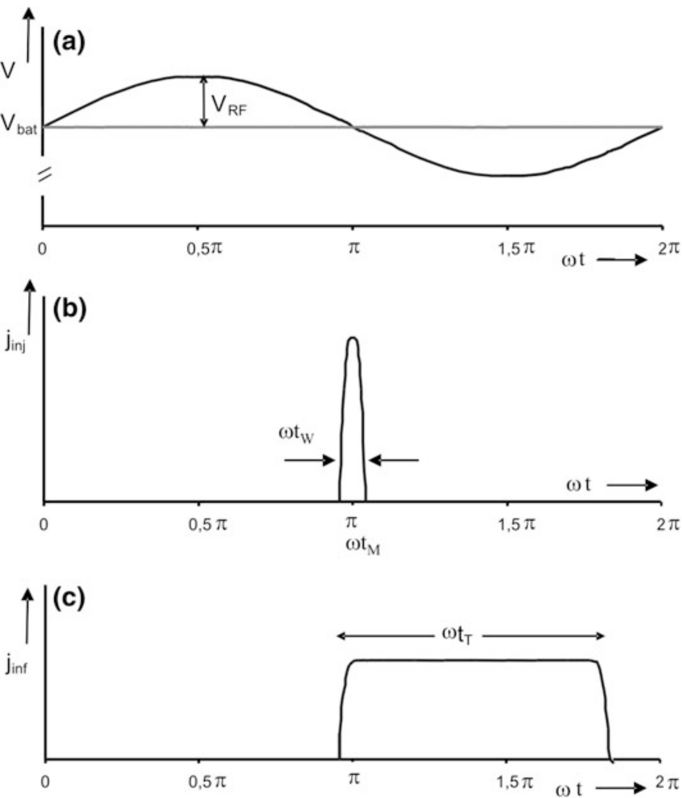


Fig. 14.22 Origin of IMPATT oscillations: **(a)** High-frequency AC voltage superimposed to the applied DC voltage **(b)** Injected current at the time point $\omega t = \pi$. **(c)** Current at device terminals

In contrast to oscillations according to the BARITT effect, there is no phase interval in which the emitted power is damped for the IMPATT effect. The IMPATT effect leads to a high-frequency oscillation of high energy. The RF efficiency of IMPATT diodes can amount up to 30%, as reported in the literature [Eis98]. Additionally, the signal of the IMPATT effect is not in a very narrow frequency band, as depicted in Fig. 14.16 for PETT oscillations, but it contains a lot of noise. This shows obviously that dynamic IMPATT oscillations must be avoided in any case.

IMPATT oscillations can occur if a power semiconductor contains too much K-centers. This can be due to too high electron irradiation doses [Lut98] or high doses of He irradiation projected to an unsuited position in the device [Sie04, Niw08]. For the case of electron irradiation which creates a homogeneous lifetime, a dimensioning rule is given in [Lut98]: Only so many K-centers are allowed that the device can sustain the maximal occurring DC-link voltage in the application (commonly 75% of the rated voltage) at its lowest operation temperature (typ. −40 °C) without occurrence of avalanche. The generation rates of K-centers at irradiation with electrons are given in the literature [Sie06]. If such a rule is observed, the dynamic IMPATT oscillation can surely be avoided.

14.3.3 Transient-Avalanche (TA) Oscillations

A further type of high frequency oscillation was found during the turn-off of 3.3 kV IGBTs at high current. Measurements and simulations indicate that the avalanche generation and transit time effect of carriers within the IGBT leads to this oscillation. The transition time effect takes place during the rise of the collector voltage at turn-off, especially during the dynamic avalanche phase. The range of frequencies is at several 100 MHz. As the oscillation was found at switching with dynamic avalanche, it was named Transient Avalanche Oscillation (TA-Oscillation).

Figure 14.23 shows the measurement of TA oscillations; they are visible in the voltage signal of the wire used as antenna as well as in the waveform of the gate voltage. The oscillations in the gate wiring are a stray-in effect. The oscillations happen at high voltages as the voltage exceeds 2800 V which is the onset of the dynamic avalanche. A large collector current I_C is still flowing and the dV_C/dt is limited by the avalanche process. Oscillations can be found at the beginning, during and at the end of the dynamic avalanche process.

Section 13.4 describes the dynamic avalanche, and for the given situation Eq. (13.13b) can be used

$$V_{av,dyn} = \frac{1}{2} \cdot \left(\frac{8}{B}\right)^{\frac{1}{4}} \cdot \left(\frac{q \cdot N_{eff}}{\varepsilon}\right)^{-\frac{3}{4}} \cdot \left(\frac{\mu_n}{\mu_n + \mu_p}\right)^{\frac{1}{4}} \qquad (14.13)$$

The processes in the IGBT at this instant have similarities to the processes described for diodes in Sect. 13.4.2. In a part of the IGBT-base layer the electric field is built

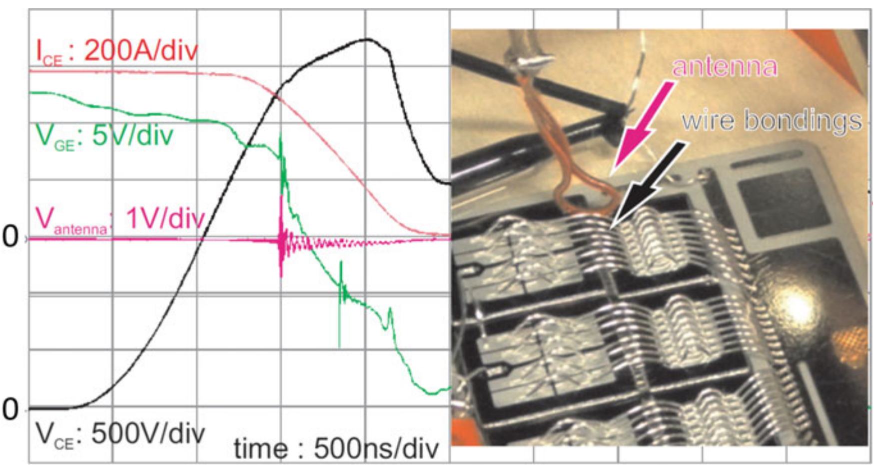


Fig. 14.23 Measurement of TA oscillations at turn-off of 3.3 kV IGBTs. Oscillations visible in an antenna signal as well in the waveform of the gate voltage. Figure from [Hon15]

up, and the current flowing through is carried by holes. This leads to a dynamic avalanche onset well below the rated voltage, see also Fig. 13.12.

The simplified carrier injection mechanism in Fig. 14.22 can be used to calculate the efficiency of RF-power generation in such a transit-time structure. The terminal voltage consists of a DC-part V_{bat} and HF-part V_{RF}. Carriers are injected into the depletion region at ωt_M during the injection angle ωt_W. They need a transit time t_T to drift through the depletion region. The drift angle is ωt_T. As derived from [Eis98], for the RF-efficiency holds, see Fig. 14.22.

$$\eta = \frac{P_{RF}}{P_{DC}} = \frac{-\frac{A}{2\pi}\int_{\omega t_M}^{\omega t_M + \omega t_T} j_{\inf}(\omega t)\cdot V_{RF}\sin(\omega t)d\omega t}{\frac{A}{2\pi}\int_{\omega t_M}^{\omega t_M + \omega t_T} j_{\inf}(\omega t)\cdot V_{DC}d\omega t} \tag{14.14}$$

The solution of (14.15), if additionally the injection angle ωt_W is considered, is

$$\eta = \frac{V_{RF}}{V_{DC}} \cdot \frac{\sin(\omega t_W/2)}{\omega t_W/2} \cdot \frac{\cos(\omega t_M + \omega t_T) - \cos\,\omega t_M}{\omega t_T} \tag{14.15}$$

The three terms in (14.15) are abbreviated as η_1, η_2, η_3. The second term η_2 is for a short $\omega t_W < \pi/5$ close to 1, and for the third term η_3, the transit time dependent part of IMPATT and PETT structures, the solution is drawn in Fig. 14.24. Although the peak values of η_3 decrease with the increase of the drift angle, the second peak value of IMPATT is still as high as the first peak value of PETT. For a certain width of the depletion region w_{sc} each peak value of η_3 in Fig. 14.24 corresponds to a transit time eigenfrequency (f_{1IM}, f_{2IM}, f_{1PE}). Signals at these frequencies would be amplified most effectively.

The resulting frequencies $f_{1IM}, f_{2IM}, f_{1BA}$ can be gained with

$$f_T = \frac{\omega t_T \cdot v_{sat}}{2\pi \cdot w_{sc}} \tag{14.16}$$

As integral effective drift velocity for electrons (IMPATT) $v_{sat,n} = 8.93 \times 10^6$cm/s and for holes (PETT) $v_{sat,p} = 6.44 \times 10^6$cm/s were used. Figure 14.24 shows the dependency of $f_{1IM}, f_{2IM}, f_{1BA}$ on the width of the space charge. For f_{1IM} the result is similar as Eq. (14.12), for f_{1PE} the result is similar to Eq. (14.9). Additionally, f_{2IM} is shown as first harmonics of f_{2IM}.

The TA oscillation usually needs a resonance circuit which is formed by the junction capacity of the devices C_{IGBT} and C_{diode}; further parasitic capacities of the substrate and the wiring close to the dies, see Fig. 14.26a.

Three possible resonance circuits can be identified in Fig. 14.24. The junction capacity of both devices is voltage dependent according to

$$C_j = \frac{\varepsilon \cdot A}{w_{sc}} \tag{14.17}$$

wherein the width of the space charge w_{sc} is voltage dependent, compare Chap. 3.1 Eq. (3.19). The extension of the depletion region leads to the reduction of the chip capacitance both in the IGBT and the diode, which means an increase of the resonance frequencies in all the resonance circuits. Due to the turn-off of the IGBT with high plasma density, the extension of the depletion region of the IGBT is slower than that of the diode. The resonance frequencies shown in Fig. 14.24 are roughly determined with the assumption of 2 times higher w_{sc} for the same voltage

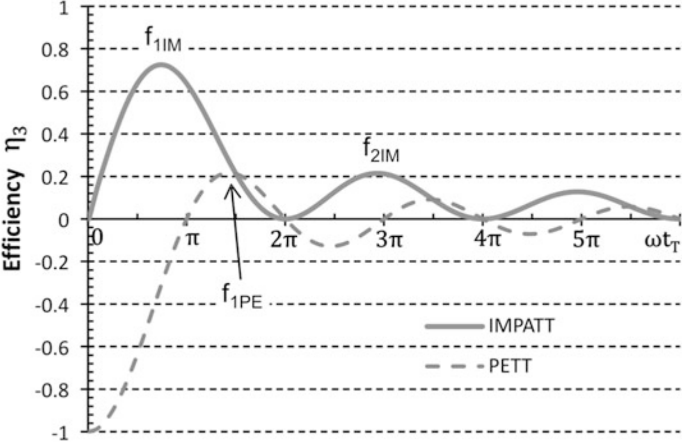


Fig. 14.24 Efficiency η_3 of IMPATT and PETT as function of ωt_T. Figure adapted from [Hon14, Hon15]

Fig. 14.25 Eigenfrequencies of the depletion region: f_{1IM}, f_{2IM}, f_{1BA} (see Fig. 14.24); eigenfrequencies of resonance circuits (s. Figure 14.26): f_{RC1}, f_{RC2} and f_{RC3}. © 2014 IEEE Reprinted, with permission, from [Hon14]

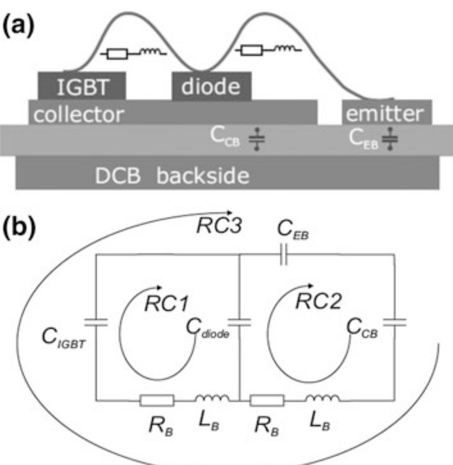


Fig. 14.26 **(a)** Substrate with direct wire bonding between IGBT and diode **(b)** Resonance circuits (RC1, RC2 and RC3) of the setup in a). © 2014 IEEE Reprinted, with permission, from [Hon14]

in the diode, and the depletion region reaches the field stop layer at 400 µm in the considered 3.3 kV diode. Oscillations would be raised under conditions close to the cross points in Fig. 14.25.

Mismatching the resonance circuit and the transit time frequencies can be applied as a countermeasure to the TA-Oscillation. Apart from this, TA-Oscillation can be eliminated by a suitable driver concept: Shortly before the avalanche, an additional short driver pulse is applied to charge the gate and open it again. Partial opening of the MOS-channel can be achieved, delivering an electron current from the MOS-channel. The hole current and hole density at emitter side can thereby be reduced. The corresponding reduction of the peak of the electric field at the blocking junction would suppress the avalanche generation.

14.3.4 Summarizing Remarks on Transit-Time Oscillations

Some aspects of the discussed transit-time oscillations are summarized in Table 14.2. Regarding intensity, the dynamic IMPATT oscillation performs worst, followed by PETT and TA oscillations. However, TA-oscillations are most difficult to avoid completely since the dynamic avalanche occurs at high voltage bipolar Si devices. In cases found up to now, their intensity is low and a resonance circuit is required. However, the IMPATT effect has a high RF efficiency and it cannot be ruled out that under special circumstances, e.g. a strong dynamic avalanche, they even occur without an external resonance circuit with the same resonance frequency.

Even a further type of transit time oscillations, the so-called TRApped Plasma Avalanche Triggered Transit (TRAPATT) oscillation is possible with power devices. It is explained as occurring during avalanche in a device with trapezoidal electric field shape (PT-Design). At a current high enough in the avalanche mode, a plasma of electrons and holes can be formed [DeL70] and an avalanche zone is running through the device.

The RF efficiency of TRAPATT oscillations is found to be very high, even 75% have been obtained [Sze8], which is far above the IMPATT oscillations. In [Kas95] TRAPATT oscillations are explained with numerical simulation of a diode with a

Table 14.2 Summarizing comparison of transit-time oscillations

	Dynamic Impact Ionization transit time (Impatt) Oszillations	Plasma Extraction Transit Time (PETT) Oszillations	Transient Avalanche (TA) Oscillations
Amplitude	200–400 V	<60 V	10 V …
Intensity	Strong radiation. Drivers etc. stop function	Easy overseen at qualification, however disturbs EMC	Hard to find at qualification, however disturbs EMC
Occurrence	Above a defined threshold voltage, which decreases with temperature	May occur at a certain voltage and temperature, can be avoided when changing the conditions	At very special conditions, hard to predict
LC resonance circuit involved?	Not necessary	Yes, necessary	Yes, in most cases
Root cause	Positively charged deep K-centers decrease avalanche onset voltage	Remaining plasma layer injects carriers discontinuously	Mainly dynamic avalanche, PETT mechanism involved
Countermeasure	Device design, limit amount of deep K-centers	Detune LC resonance circuit	- Detune resonance circuit - Short-time electron injection by channel

narrow base at snap-off of the reverse current. The high *di/dt* at snap-off generates a voltage peak larger than the breakdown voltage, which is clamped by the avalanche of the diode. The shown amplitude of the voltage oscillation is almost as high as the applied DC voltage. However, this has occurred in simulations. Snappy diodes are ruled out in most power-electronic applications, see Sect. 14.2.2. Soft recovery behavior is necessary.

Investigations of superjunction MOSFETs with the transmission line pulse (TLP) method in [Chi16] report TRAPATT oscillations at the boarder of the avalanche capability of the device. With the TLP method, short-time high voltage high-current pulses in the avalanche mode are applied, and high-frequency oscillations were found in a short time interval.

However, also at the turn-off of IGBTs voltage peaks are generated which can overshoot the avalanche breakdown voltage. Therefore, it is not ruled out that further oscillation effects will be found.

References

[Bec16] Beckedahl, P., Buetow, S., Maul, A., Roeblitz, M., Spang, M.: 400 A, 1200 V SiC power module with 1nH commutation inductance. In: Proceedings CIPS 2016, VDE Verlag GmbH (2016)

[Chi16] Chirilă, T., Reimann, T., Rüb, M.: Dynamic avalanche in charge-compensation MOSFETs analyzed with the novel single pulse EMMI-TLP method. In: Proceedings of 2016 IEEE International Reliability Physics Symposium (IRPS), pp. 1–5 (2016)

[DeL70] DeLoach B.C., Scharfetter D.L.: Device physics of TRAPATT oscillators. IEEE Trans. Electron Devices **17**(1), 9–21 (1970)

[DIN00] DIN EN 55011 – Industrielle, wissenschaftliche und medizinische Hochfrequenzgeräte; Funkstörungen – Grenzwerte und Messverfahren, VDE-Verlag GmbH, Berlin (2000)

[Eis98] Eisele, H., Haddad, G.: Active microwave diodes. In: Sze, B.M. (eds.) Modern Semiconductor Device Physics. Wiley, New York (1998)

[Eld98] El-Dwaik, F.: Ein Beitrag zur Optimierung des Wirkungsgrades und der EMV von Wechselrichtern für batteriegespeiste Antriebssysteme, Dissertation, Chemnitz (1998)

[Fei15] Feix, G., Hoene, E., Zeiter, O., Pedersen, K.: Embedded very fast switching module for SiC Power MOSFETs. In: Proceedings PCIM Europe, pp. 104–110 (2015)

[Fuj12] Shigeto Fujita, S.: Simulation study on insulated gate bipolar transistor turn-off oscillations. Jpn J Appl Phys **51**, 054101 (2012)

[FLO04] Flomerics Ltd.: FLO/EMC Reference Manual Release 1.3, 2004

[Gut01] Gutsmann, B., Silber, D., Mourick, P.: Explanation of IGBT tail current oscillations by a novel plasma extraction transit time mechanism. In: Proceeding of the 31st European Solid-State Device Research Conference, pp. 255–258 (2001)

[Gut02] Gutsmann, B., Mourick, P., Silber, D.: Plasma extraction transit time oscillations in bipolar power devices. Solid-State Electron. **46**(5), 133–138 (2002)

[Hon14] Hong, T., Pfirsch, F., Bayerer, R., Lutz, J., Silber, D.: Transient avalanche oscillation of IGBTs under high current. In: Proceedings ISPSD (2014)

[Hon15] Hong, T.: Transient avalanche oscillation of IGBTs under high current, Ph.D. thesis Chemnitz (2015)

[Kas95] Kaschani, K.T., Sittig, R.: How to avoid TRAPATT oscillations at the reverse recovery of power diodes. In: International Semiconductor Conference, CAS'95 Proceedings, pp. 571–574 (1995)

[Kas97] Kaschani, K.T.: Untersuchung und Optimierung von Leistungsdioden, Dissertation, Braunschweig (1997)

[Lut98] Lutz, J., Südkamp, W., Gerlach, W.: IMPATT oscillations in fast recovery diodes due to temporarily charged radiation induced deep levels. Solid-State Electr. **42**(6), 931–938 (1998)

[Niw08] Niwa, F., Misumi, T., Yamazaki, S., Sugiyama, T., Kanata, T., Nishiwaki, K.: A Study of correlation between CiOi defects and dynamic avalanche phenomenon of PiN diode using he ion irradiation. In: Proceedings of the PESC, Rhodos (2008)

[Omu030] Omura, I., et al.: Electrical and mechanical package design for 4.5 kV ultra high power IEGT with 6kA Turn-off capability. In: Proceedings of the ISPSD, Cambridge (2003)

[Pal99] Palmer, P.R., Joyce, J.C.: Causes of parasitic current oscillation in IGBT modules during turn-off. In: Proceedings of the EPE, Lausanne (1999)

[Sie03] Siemieniec, R., Lutz, J., Netzel, M., Mourick, P.: Transit time oscillations as a source of EMC problems in bipolar power devices. In: Proceedings of the EPE, Toulouse (2003)

[Sie04] Siemieniec, R., Lutz, J., Herzer, R.: Analysis of dynamic impatt oscillations caused by radiation induced deep centres with local and homogenous vertical distribution. In: IEEE Proceedings Circuits, Devices and Systems, vol. 151(3), pp. 219–224 (2004)

[Sie06] Siemieniec, R., Niedernostheide, F.J., Schulze, H.J., Südkamp, W., Kellner-Werdehausen, U., Lutz, J.: Irradiation-induced deep levels in silicon for power device tailoring. J. Electrochem.l Soc. **153**(2) G108-G118 (2006)

[Sie06b] Siemieniec, R., Mourick, P., Netzel, M., Lutz, J.: The plasma extraction transit-time oscillation in bipolar power devices – mechanism, EMC effects and prevention. IEEE Trans. El. Dev. **53**(2) 369–379 (2006)

[Sze81] Sze, S.M.: Physics of semiconductor Devices. Wiley, New York (1981)

[Win12] Wintrich, A.: "Herausforderungen beim Einsatz von SiC in Hochleistungsmodulen", ELEKTRONIKPRAXIS Leistungselektronik & Stromversorgung, Mai (2012)

[Zim95] Zimmermann, W.: Sommer KH, Patent DE 19549011C2 (1995)

Chapter 15
Integrated Power Electronic Systems

15.1 Definition and Basic Features

The expression 'power electronic system' is used in different contexts with different meanings. Whereas the term "power electronic system" in the Introductory Chap. 1 refers to a complete converter system, it should be clear to the reader that such converter systems comprise many sub-systems. In the context of integrated power electronic systems, the term "power electronic system" will be used for an assembly of those components that are necessary to perform an energy conversion task and that can be integrated with a given integration technology.

Figure 15.1 illustrates this definition with an example of a hybrid propulsion system.

In a hybrid vehicle, the main power electronic function is the transformation of energy from a battery to an electric traction motor and possibly vice versa. The fundamental switching processes are taking place in the silicon power devices, which are not shown in Fig. 15.1. On this level of abstraction, the principle item is a three-phase converter module with an internal circuit as shown in the introductory chapter (see Fig. 1.5) and is illustrated here in Fig. 15.2. This converter module contains the power electronic devices, six IGBTs and six freewheeling diodes, as well as all the electrical contacts and the interface to a heat sink. Such converter topologies, as shown in Fig. 15.2, are available in a single power module package.

The next level of functionality is reached by adding gate drives for the six IGBTs, sensors for temperature and potentially current sensors and protection logic circuits. The integration of these functions in one package unit is named Intelligent Power Module (IPM). IPM modules are commercially available for small and medium power ranges.

Subjoining a DC-link energy storage, the DC-link charge circuit and the auxiliary power supply as well as the heat sink to the assembly leads to the next level of

© Springer International Publishing AG 2018
J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8_15

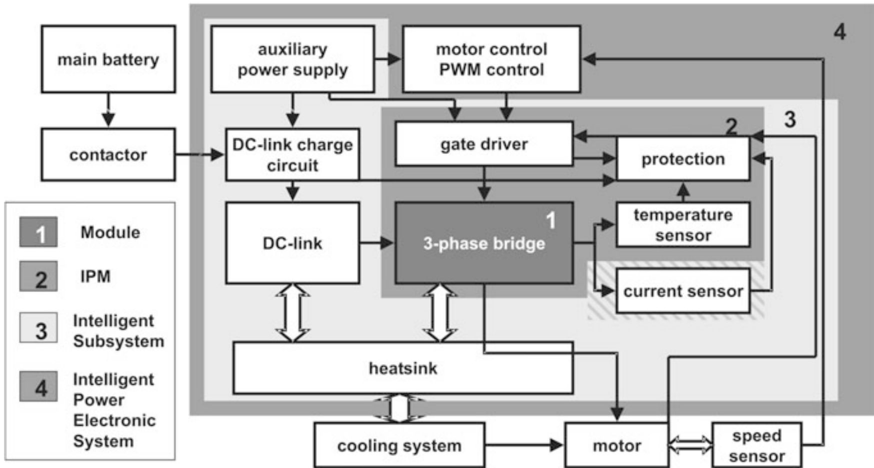
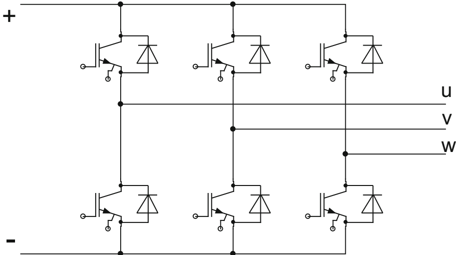


Fig. 15.1 Illustration of the terminology in power electronic systems using the example of automotive hybrid traction drives. Schematics from W. Tursky, Semikron

Fig. 15.2 Three-phase converter circuit diagram



functionality: the intelligent sub-system. For specific applications, integrated custom specific intelligent sub-systems are available on the market.

The only element missing to form a complete power electronic system is a digital controller, which typically comprises a 'Digital Signal Processor' (DSP) or microprocessor, and fast logic devices, such as Field Programmable Gate Arrays (FPGAs), to generate the Pulse Width Modulation (PWM) signals and to control all aspects of the power electronic system by suitable software. The external heat exchanger as well as the source (battery with safety contactor) and the drain (motor with speed sensors) of the energy are not considered as part of the power electronic system.

A power electronic system must comprise the following features:

- *Full electrical functionality*: The power electronic circuit with the associated driver circuits, the sensor elements to monitor the operating status of the converter (output current, DC-link voltage, reference temperatures) and additionally start-up circuits, protection circuits and auxiliary power supply circuits.

- *Thermal management components*: The extraction of heat from the silicon power devices is of highest priority for the operation of power electronic systems. The efficient function of these components should be permanently monitored to avoid thermal overload conditions.
- *Hard- and software for PWM and control*: The PWM and control algorithms are of fundamental relevance for the efficient and reliable performance of a power electronic system. They can enhance the energy efficiency of an application by selecting optimal operation conditions of the motor and they can protect the power electronic system from unexpected stress conditions and external failures like motor short circuits.

The system approach is an essential prerequisite for effective system optimization. Focusing on a single aspect of the system optimization, for example the increase of the power density per unit volume, can lead to a suboptimal system with respect to the reliability, as not all factors are taken into account.

A general strategy to increase the power density without sacrificing the system reliability is the system integration. Integration of functions can eliminate connections and interfaces and can sometimes exploit synergy effects by combining more than one function in a single functional element. Examples for such synergy effects are frequently found in the integration of passive components; in multi-layer printed circuit boards (PCBs), capacitors, inductors and transformers can be simultaneously integrated into a single building block [Dic08, Waf05]. This concept has a high potential for improvements in power electronic systems in the future.

An effective concept is the monolithic integration. Today, components are available for small power, which have all power electronic, control and logic functions integrated on a single chip. Hybrid integration, which assembles different components on a single substrate, is another way of system integration. Common to all strategies of integration is the reduction of footprint, the system volume and the weight, while simultaneously reducing interconnections and interfaces.

A high level of integration reduces the effort necessary for the system assembly on the one hand, but increases the complexity of the components on the other hand. The increased complexity of highly integrated components requires a high quality of the manufacturing process and demands comprehensive test procedures to verify the component functionality. This enhanced effort in the manufacturing process is only commercially reasonable when the production quantities are large enough. Furthermore, the integration of control and protection functions will impede the application of established test procedures to validate the system reliability. New concepts for accelerated test to derive lifetime models and for screening test beyond the specification limits must be developed.

The design of power electronic systems requires expert knowledge in several engineering disciplines, i.e. mechanical engineering, electrical engineering, material science, computer science, etc. It requires a cooperation of engineers from all these fields for a successful design of a new power electronic system. A single engineer alone can hardly cover all aspects of such system design, so teamwork is mandatory.

15.2 Monolithically Integrated Systems – Power IC's

Monolithic integration is the assembly of different functions on a single silicon chip. Sensor elements, analogue and digital circuits as well as power electronic devices are combined in a single integrated circuit.

The standard CMOS technology is the basis for the majority of integrated systems. The abbreviation CMOS stands for Complementary MOSFET and indicates a technology adapted from n-channel and p-channel MOSFETs. Additional functional elements like sensors, storage cells, power devices and the necessary interconnections can be added in the same technology by supplemental process steps and mask levels.

One of the first technologies that successfully implemented this process extension is the Smart SIPMOS technology [Pri96]. Figure 15.3 shows an application example combining logic and power circuits on a single chip. The vertical power transistor (power N-MOS) on the right hand side in Fig. 15.3 has already been discussed in Chap. 9. The compatibility of its structure to the CMOS structures on the left hand side is evident. Source structures, gate oxides and gate contacts, passivation layers and contact metallization layers can extensively be inherited from the CMOS technology.

The mutual insulation of the different elements in the circuit shown in Fig. 15.3 is realized by pn-junctions, therefore this concept is called 'junction insulation'. The blocking capability of devices based on this technology is with justifiable effort limited to 100–200 V [Gie02]. A closer look on Fig. 15.3 reveals a multitude of parasitic structures in the circuit: npn- and pnp-transistor structures and even pnpn-structures representing thyristor type parasitic elements. During operation at significantly high current densities and particularly at elevated temperatures and

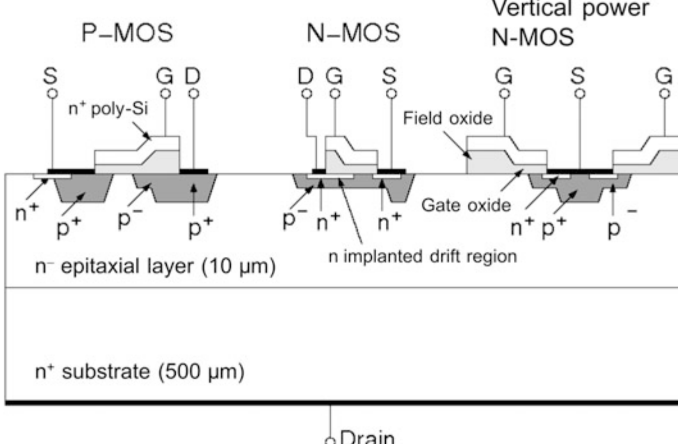


Fig. 15.3 Self-insulated vertical DMOS transistor integrated with CMOS logic elements. Drawing similar to [Tih88b]

high voltages, the interactions between different devices can provoke a latching of the parasitic thyristor structures. This effect is the dominant failure mode of integrated circuits; it limits the permitted current and voltage levels and primarily constrains the operation temperature range.

Adapted to the high voltage levels is the technology presented by ST Microelectronics as 'vertical intelligent power technology', described in [And96]. Figure 15.4 illustrates an exemplary application. A vertical power MOSFET is shown on the left side, which serves as output stage. This device was discussed in Chap. 9. The elements of the logic circuit are insulated by several pn-junctions. These pn-junctions are formed (beginning with the bottom n^+ substrate) by a first n^- epitaxy layer, local p and n^+ buried implantation layers, a secondary n type epitaxy layer and in lateral direction by deep diffusion p columns. The epitaxial growth process is interrupted for the implantation of the buried p and n^+ islands. During the resumed epitaxial growth, the elevated process temperatures initiate diffusion in the buried layers, so that they reach their projected dimensions at the end of the process. This elaborate multilayer insulation technique insulates the logic circuit elements for blocking voltages up to 600 V.

The collector of the npn bipolar transistor is connected to the buried n^+ collector island by a vertical n^+ column. This typical design structure for a high-gain npn-transistor is therefore denoted as 'pseudo vertical' device.

The technology depicted in Fig. 15.4 represents a platform for the integration of manifold structures and functions of analogue and digital circuits. The fabrication of the deep p zones for the electrical separation of the functional elements is rather elaborate. The deep diffusion of these columns from the surface is accompanied by

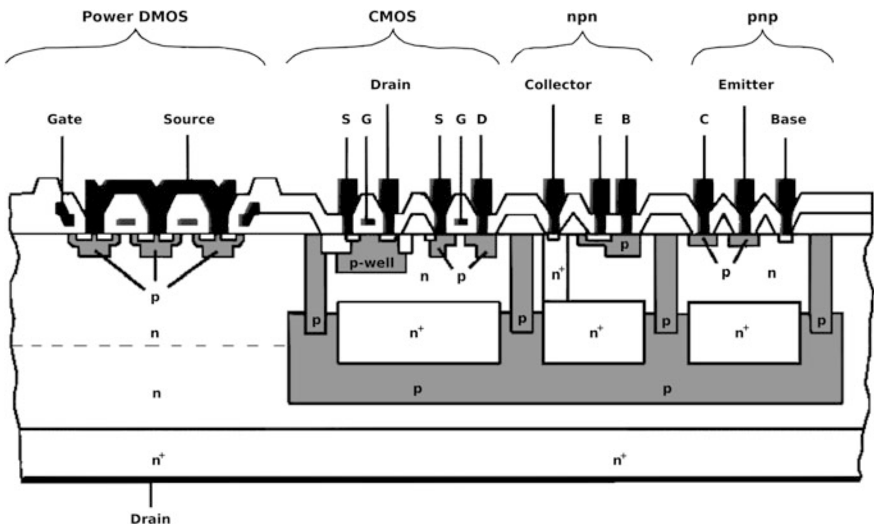


Fig. 15.4 'Vertical intelligent power technology' from ST Microelectronics. Figure prepared by R. Herzer based on ST Microelectronics datasheets

a lateral diffusion with an aspect ratio of 0.8. For a 10 μm deep column, a lateral diffusion of 8 μm must be taken into account. This technique therefore requires relatively wide separation areas and thus provokes a loss of area available for active devices.

Another drawback is the generation of a leakage current with voltage applied to the pn-junctions, which is increasing for higher temperatures. The leakage currents limit the maximum blocking capability and restrict the operation temperature range as well. They comprise the potential of latching (the turn-on of parasitic thyristor structures), which results in the destruction of the circuit. Pn-insulated technologies are limited for high voltages to temperatures of maximal 150 °C.

These constraints can be eliminated by a dielectric insulation technology, where oxide layers are implemented for the electrical separation of different devices. These oxide layers feature bidirectional insulation and generate much smaller leakage currents, so that a higher blocking capability can be realized with comparatively thin layers even at high operation temperatures.

An example of this so-called 'silicon on insulator' (SOI) technology in conjunction with a trench technology is shown in Fig. 15.5. A wafer bonding process, in which two oxidized silicon wafers are merged together, produces the substrates for this technology. The top wafer was prepared with a diffused n^+ surface region prior to the bonding process. After the wafer bonding process, the top layer is grinded down to the desired thickness and afterwards polished. A subsequent etching process forms the deep trenches. An implantation process followed by a diffusion process generates the n^+ layers on the trench sidewalls. Then the trenches are filled with SiO_2 and finally a plane surface is created by mechanical treatment. Now, the actual fabrication of the active devices can start.

Figure 15.5 shows a cross section of a pseudo-vertical n-channel MOSFET (see Chap. 9 for reference) in SOI-technology for blocking voltages of 600 V, suitable for 230 V grid applications. The deep n^+ layer serves as the drain contact, which is routed via the vertical n^+ regions of the trench sidewalls to the surface contact on the right edge of the cell. In the adjacent dielectrically insulated cells, other independent power devices or arbitrary logic circuit structures (CMOS, bipolar elements) can be placed.

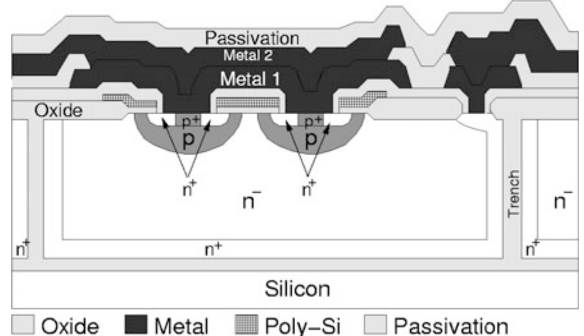


Fig. 15.5 Schematic cross section of a pseudo-vertical n-channel MOSFET in silicon-on-insulator thick film technology in conjunction with trench insulation technique from [Ler02]

The production effort for SOI-substrates is comparatively high. However, this technology effectively prevents any interaction between different elements of the circuit even at high current and voltage levels and under elevated temperatures. The crosstalk between elements of the circuit – another limiting factor for an increasing integration density – is also reduced by the dielectric insulation. The SOI technology allows a high packing factor and exhibits a better exploitation of the wafer area due to the smaller insulation regions. SOI devices are immune to latch-up problems generated by parasitic pnpn-structures. As a result, the integrated devices remain functional up to operation temperatures of 200 °C.

Monolithic integration has made tremendous progress during the last years with respect to packing density, blocking capability and temperature stability. However, the conflicting requirements of high voltage and temperature stability and the immunity against crosstalk effects on the one hand and the demand for increasing packing factors on the other are a challenge for further progress. Today, systems with a blocking capability up to 1200 V and current up to 10 A are integrated in smart power ICs. For higher voltages and currents, hybrid or discrete solutions are preferred.

15.3 GaN Monolithic Integrated Systems

GaN power devices are lateral devices as described in Chap. 10. Nowadays, they are fabricated in AlGaN/GaN-on-Si technology as described in Sect. 4.11. This technology also offers new opportunities in terms of monolithic integration. Several components can be placed side-by-side on a single chip. This is enabled by isolated areas in between as well as by the property of lateral devices that all terminals of components can be interconnected on the top side of a chip [Rei16].

Figure 15.6 shows a lateral GaN HEMT (see Sect. 9.13) with integrated free-wheeling diode. Schottky contacts, which are combined with the source connected field plate, provide a path for reverse conduction that is independent of the state of the gate. The area-consuming depletion zone is used for both components, for the HEMT and the antiparallel diode. Therefore, the intrinsic structure requires little additional chip area and does not contribute significant parasitics.

The integration of a half-bridge configuration with two HEMTs and two Schottky diodes into one single chip is shown in Fig. 15.7.

This device realizes the function of four conventional chips in one die. Due to the high field strength of GaN, very narrow isolation distances are achievable. A question still under debate for this type of integration is the optimal potential to be applied to the substrate. In a single HEMT, the Si-substrate (Fig. 15.6) is usually at the same potential as the source electrode. However, both source points in Fig. 15.7 are at different potential. The problem of the backside coupling is investigated in [Wei16] and a trade-off solution is presented. To avoid the current collapse issues (see Sect. 9.13), the substrate is connected to $V_{bat}/2$ by an external

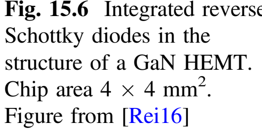


Fig. 15.6 Integrated reverse Schottky diodes in the structure of a GaN HEMT. Chip area 4 × 4 mm^2. Figure from [Rei16]

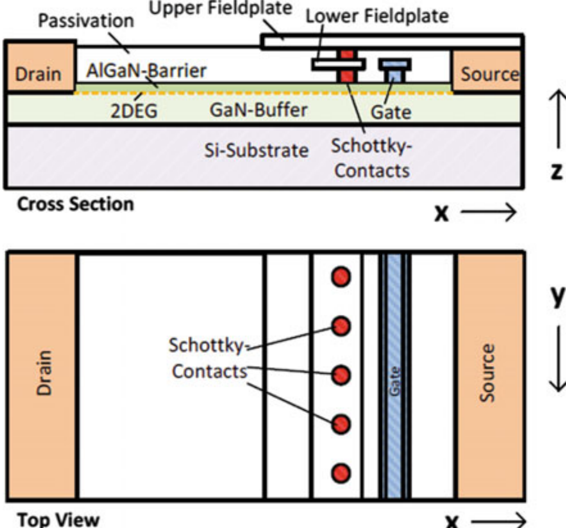
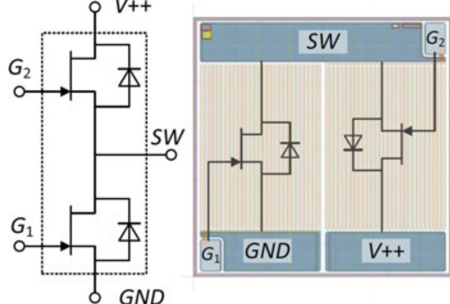


Fig. 15.7 Circuit and chip layout of a monolithic half-bridge with integrated reverse Schottky diodes. Chip area 4 × 4 mm^2. Figure from [Rei16]



resistive network with two resistors R_{DIV} = 110 kΩ between positive supply voltage and ground.

A very advantageous feature of GaN devices is the low gate charge. It results in 20-times lower drive losses than comparable silicon devices [Kin16], which makes low power driving electronics possible even for high switching frequencies. Driver and logic functions can be integrated in the chip. An example is shown in Fig. 15.8.

The realized 650 V GaN HEMT is of enhancement mode. The driver losses are smaller than 35 mW at 1 MHz switching frequency. The integration of driver on the chip is reducing the loop inductance in the gate circuit to a minimum, which is a significant advantage for fast switching. Delay time t_d and t_{on}, t_{off} are in the range of 10–20 ns. The device shows a high *dv*/*dt* immunity and is stable up to 200 V/ns [Kin16]. Further functions can be realized as well. A 650 V half-bridge power IC with monolithic integration of 650 V eMode GaN FET, driver, logic, level-shift, bootstrap and protection functions is reported in [Kin17].

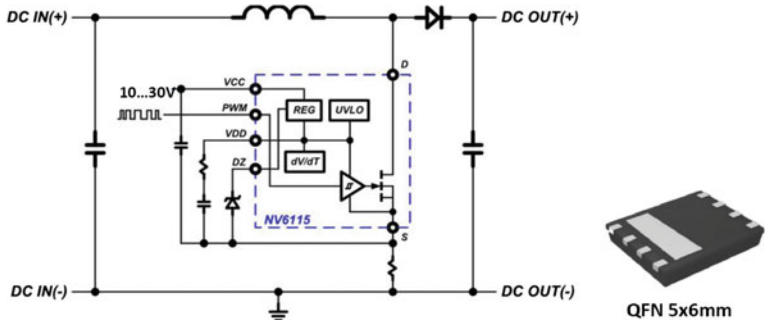


Fig. 15.8 GaN Power IC with monolithic integration of GaN FET, GaN Driver and GaN Logic. Figure adapted from [Kin16a]

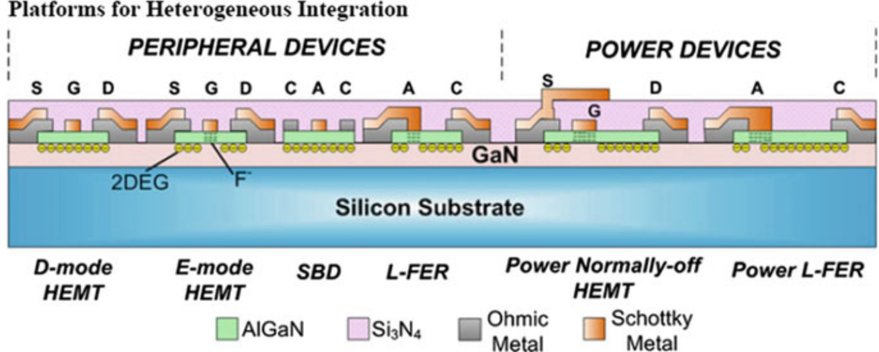


Fig. 15.9 Platform for monolithic integration in GaN-on-Si technology. Figure from [Che17] © Springer 2017

Figure 15.9 shows simplified a technology cross-section of different semiconductor structures and functions which can be realized in the lateral GaN-on-Si technology [Che17]. The d-mode (depletion-mode) HEMTs are of normally-on type, the e-mode (enhancement mode) HEMTs are of normally-off type, which is realized by a local fluorine implementation below the gate, compare Fig. 9.35.

Different functions used for analogous and logic devices are possible. Beside the power HEMT transistor, Schottky diodes and diodes (Lateral Field-Effect Rectifier, L-FER) are realized. Due to the lack of p-channel GaN devices, the inverter configuration in GaN ICs is the so-called direct-coupled FET logic which uses both e-mode and d-mode n-channel HEMTs [Che17].

An integrated enhancement/depletion-mode GaN HEMT is shown in [Cai06]. The function corresponds to the C-MOS inverter. The schematics are compared in Fig. 15.10.

A GaN-based IC with full-pulse width modulation (PWM) function was presented in [Wan15]. The circuit is able to generate a 1 MHz PWM signal with a duty

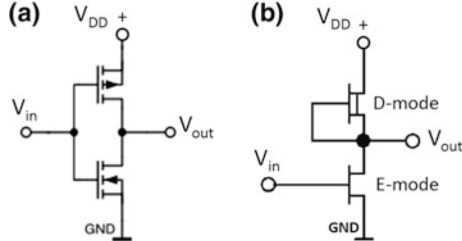


Fig. 15.10 **a** Classic C-MOS inverter used in ICs consisting of p-and n-channel MOSFET. **b** GaN inverter IC consisting of e-mode and d-mode HEMT, adapted following [Cai06]

cycle modulated over a wide range. The IC can operate at temperatures up to 250 °C. Such a circuit could be integrated with GaN power devices into a monolithic power IC in the future.

Monolithic integration in combination with high switching frequency enables small passive components and very compact power electronic systems. Significant progress in this field is to be expected.

15.4 System Integration on Printed Circuit Board

Discrete passive components in their conventional individual package consume a considerable part of the volume of, for example, grid-connected power electronic systems. Their main function is the preservation of the high quality of power grids. The progress of power electronic devices enables the transition to higher switching frequencies and therefore reduces the necessary capacitance and inductance values. This trend facilitates the integration of these passive components on printed circuit boards (PCB), which are a common platform for power electronic systems [Waf05]. The 'embedded passive integrated circuit' (emPIC) technology allows compact system designs and a high power density per unit volume [Pop05].

Layers with specific properties are required to design such a system, as can be seen in Fig. 15.11 [Waf05]. The application of printed electrical resistors on so-called prepregs (short form for pre-impregnated fibers, a semi-finished part in the PCB production process) is a state-of-the-art industrial process, although the alignment tolerances requested by the emPIC technique is a requirement that is not easy to fulfill. The challenges for the research are to develop suitable layers with high dielectric constants and layers with high magnetic permeability.

Special PCB prepregs loaded with particles of a high dielectric material can be used to form capacitors. A commercially available layer material named 'C-Lam' features a relative dielectric constant of $\varepsilon_r = 12$. Applied as a dielectric layer with 40 µm thickness between two layers of copper, the so formed component has a capacity of 0.26 nF/cm^2 with a dielectric loss factor of 0.02 and a frequency response up to 1 GHz.

Embedding ferrite particles in a polymer matrix allows the fabrication of magnetic layers. MagLam is a brand name for a commercially available layer material,

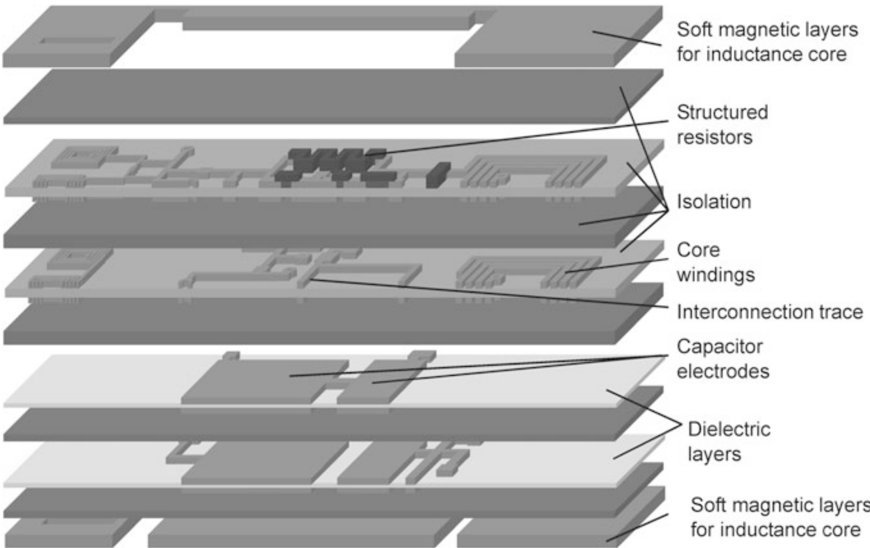


Fig. 15.11 Explosion view of the layers in a high integration emPIC PCB [Waf05]

which is compatible with the PCB production process based on prepregs. MagLam exhibits a relative permeability of $\mu_r = 17$, the saturation flux density is 300 mT and the frequency application range exceeds 10 MHz. Figure 15.12 illustrates, that an integrated transformer with a closed magnetic core can be created by appropriate structuring of the layers. During the lamination process, in which the layers are pressed together under elevated temperatures, material from the magnetic layers is pressed in the vertical duct and forms the closed magnetic core.

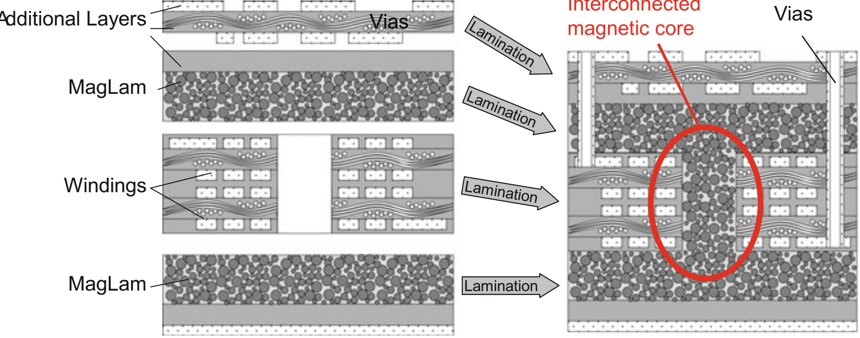


Fig. 15.12 Creation of a closed magnetic core with the ferrite polymer compound 'MagLam', which is compatible to the PCB lamination process

Another possible solution for the implementation of magnetic layers is 'μ-metal', which can be integrated as thin foils of 50 μm thickness. A relative permeability μ_r of more than 10,000 can thus be achieved. However, this material is electrically conductive and is therefore less suitable for high-frequency applications. The μ-metal foils can be laminated and structured as conventional copper layers and applied to thin flexible substrates as 'flexfoil' (polyimide), elastic coils can be created [Waf05b].

An example of a complete system based on the emPIC-technology is depicted in Fig. 15.13. It shows an AC/DC converter with a resonant topology, which delivers an output power if 60 W from a 230 V supply grid with an efficiency of up to 82% [Waf05]. The power MOSFETs and the driver ICs are assembled in discrete packages on the surface. Most of the other passive components have been integrated in the printed circuit board. The footprint of the systems has the size of a credit card.

Design and optimization of passive integration requires the consideration of all interactions between different layers and elements to prevent undesirable effects. A three-dimensional simulation of the electro-magnetic fields inside the PCB by solving the Maxwell equations is a powerful tool to avoid these problems. Software tools for this task, which were already discussed in Sect. 14.3, become more and more available. The analytical approach is nonetheless essential for a fast conceptual design especially for integrated transformers [Waf05].

The integration of passive components represents a tremendous progress in power electronic system design. The PCB, formerly only used as assembly platform and wiring element, advances to a functional component of the system. The traditional handicaps of the PCB traces like parasitic inductance and parasitic capacitance are transformed into functional elements. The number of solder joints is dramatically reduced and the compact design with less externally assembled components makes the system less sensitive to mechanical vibration and shock. Those factors increase the system reliability. The passive integration has a high potential for system improvements in the future [Nee14].

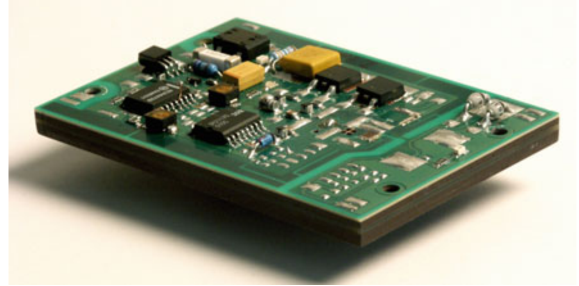


Fig. 15.13 Slim line 60 W AC/DC converter for 230 V supply voltage—transformers and the majority of the capacitors are integrated in the PCB

15.5 Hybrid Integration

Integration of power electronic systems constitutes a particular challenge, because the extraction and dissipation of the heat, generated in components during operation by power losses, implies narrow constraints to miniaturization. While microelectronics has achieved a tremendous progress over a period of several decades—continuously doubling the number of transistors on a single chip according to 'Moore's Law' every two years—by confining the possible elements to a small set of standard elements on one hand and in scaling down these standard elements to smaller and smaller sizes on the other hand, power electronics can adapt this strategy only rudimentary.

Three major paradigm shifts were responsible for the revolutionary progress in information technology. The first step was the reduction of information to binary elements—a sequence of 0s and 1s—and thus the standardization of data. The second step was the introduction of the CMOS technology, consisting of CMOS inverter structures and DRAM storage cells, as illustrated in Fig. 15.14a [Bor05]. Every microelectronic system was constructed by combining these basic elements to complex circuits and the progress was focused on the miniaturization of these elements. The third step was the 'Very Large Scale Integration' (VLSI) technology.

Can similar paradigm shifts in power electronics lead to a progress in integration comparable to the enormous success in microelectronics? The equivalence to the digital concept in information technology is the standard Pulse Width Modulation (PWM) technique in power electronics. The energy flow on the input side is chopped into single packets and is combined to the desired energy flow on the output side, which could be either a DC current with controlled amplitude in a DC

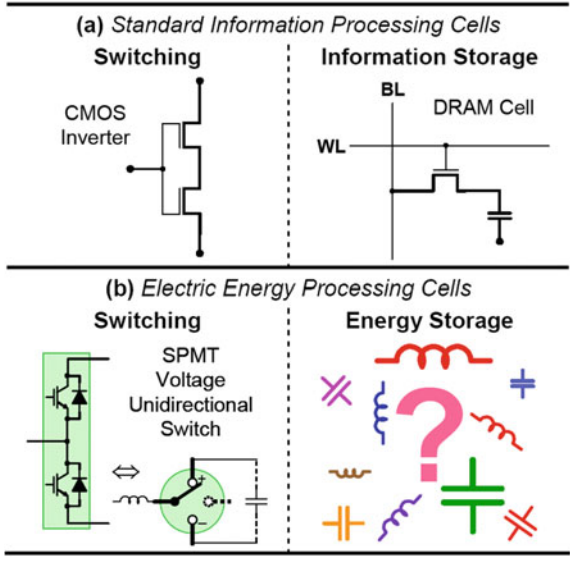


Fig. 15.14 Standardized modularization in **a** information technology and **b** power electronics according to [Bor05]

regulator, or else a sinusoidal current and voltage characteristic of selectable frequency in an AC inverter. There is also an equivalence for the second step: The basic topology of two switching devices and two anti-parallel freewheeling diodes in half-bridge or phase-leg configuration as shown in Fig. 15.14b (left) is a standard topology in power electronics. This configuration is used in the vast majority of power electronic applications.

The unsolved problem is step three in this analogy: the equivalence for a standardized storage cell like the DRAM in information technology, which would be a standard energy storage cell in power electronics (Fig. 15.14b, right). A variety of single components in various technologies and package outlines are available without a noticeable trend for standardization. The final step in system assembly, the wiring of power devices and passive components, is elaborate and the passive components are to a great extent responsible for volume and weight of a system. A continuous progress by increasing integration density as in information technology is still not conceivable for power electronics.

State-of-the-art power electronic devices can be operated today at high switching frequencies, especially power MOSFETs. This facilitates the reduction of the capacitive and inductive components. Materials with a high relative dielectric constant ε_r or a high relative permeability μ_r are available as discussed in Sect. 15.4. However, additional boundary conditions have to be accounted for besides the electrical requirements. The heat generated by power losses in passive components must be efficiently dissipated. Additional heat produced by the nearby power devices can significantly increase the operation temperature of the passive components, so that high temperature ratings are necessary. Finally, the stress provoked by differences in coefficient of thermal expansion (CTE) in combination with high operation temperature is a challenge for the reliability of the components. These additional requirements impede the progress of integration technologies.

While the integration of passive components is still in a state of infancy, manufacturers of power modules have developed a different strategy: the integration of driver circuits into power modules. These ‘intelligent power modules’ (IPM) facilitate the system design and have reached a certain standard in the small power range.

As was already discussed in Sect. 11.2, the transfer mold technology based on copper lead frames is an ideal platform to integrate drivers in the low power range (see Fig. 11.11). With increasing power losses, the heat transfer through the contact leads is not sufficient. Integrated cooling structures enhance the heat dissipation of the package. Figure 15.15 illustrates this improvement for a DIP-IPM power module from Mitsubishi. This package is suitable for an output power of 1.5 kW with phase currents up to 20 A.

Higher output power levels require even more heat extraction capability. Implementing ceramic substrates allows designing IPM modules for output power of up to 22 kW. The example in Fig. 15.16 shows a power module without base plate, comprising a three-phase input rectifier, a three-phase inverter, a brake chopper and a 7-channel SOI driver (refer to Fig. 15.5). The challenge in this

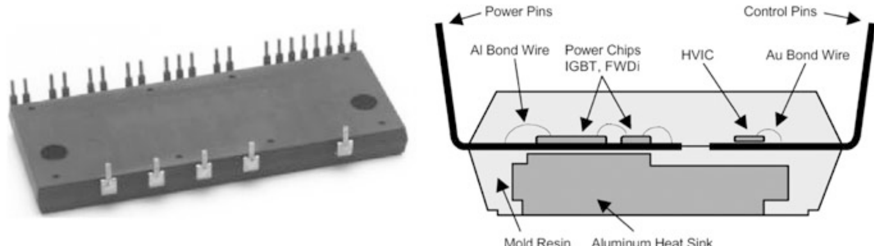


Fig. 15.15 Photographic image (left) and cross section (right) of a Mitsubishi DIP-IPM – the implemented heat sink structure increases the heat extraction [Mot99]

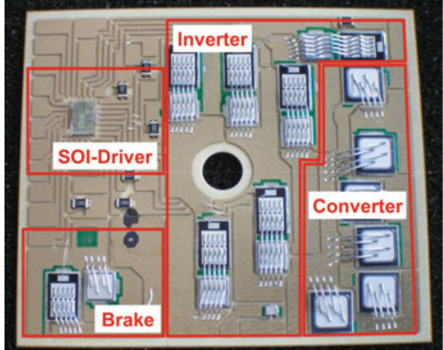


Fig. 15.16 IPM module without base plate from Semikron – an input converter, an inverter, a brake chopper and a SOI driver assembled on a single substrate [Grs08]

design is the production of the small current tracks on the DBC substrate for the multiple SOI contacts. However, the excellent thermal contact of the SOI chip to the heat sink enables the device to dissipate more power losses, so that the output stages can produce higher gate currents and therefore are able to drive even larger chips.

Advanced level shifters are integrated in the SOI chip for the TOP and BOT gate drivers, which compensate shifts in the reference potential of both polarities. This feature makes the driver immune to static and dynamic reference potential changes up to $\pm$20 V.

Another approach for the IPM design is the integration of a common PCB driver into the package of a classical base plate module as shown in Fig. 15.17. The PCB is connected to the power circuit by soldered posts, which position the PCB below the top cover of the module housing.

These IPMs are typically equipped with a temperature sensor. The driver provides short-circuit detection and several additional protection functions for safe operation.

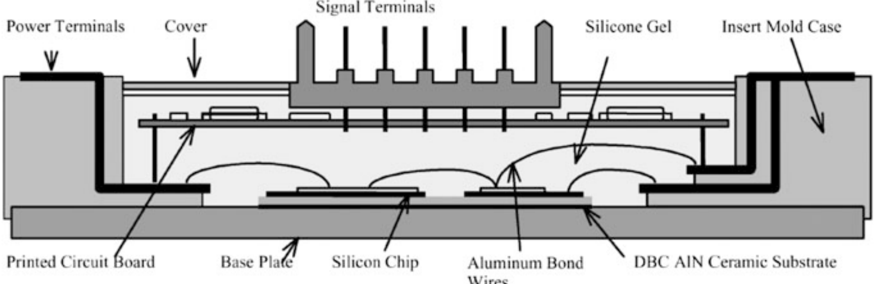


Fig. 15.17 Mitsubishi-IPM for engine ratings up to 30 kW [Mot93]

The next step in increasing complexity implements a micro-controller together with the control algorithms for PWM in a unique package. Thus, a motor control unit emerges that contains a control circuit board, which includes a computer chip with the appropriate control software (Fig. 15.18) [Ara05]. The application engineer receives a complete system as a black box, which can be adapted to the application demands simply by transmitting suitable control sequences.

The integration of the DC-link capacitors upgrades the package to a complete power electronic system, as shown in Fig. 15.19. The example shows a single housing equipped with a three-phase MOSFET inverter, current sensors, DC-link bus bars and DC-link capacitors, driver board and a micro-controller. The sealed package design makes this compact architecture ideally suited for the application in industrial service vehicles.

While these concepts of integration aim to design a high-volume series product applicable to a multitude of applications, the SKiiP platform introduced by Semikron focuses on a flexible design, which can be easily adapted to different requirements [Scn02]. Figure 15.20 exemplifies this concept with a module from the SKiiP family, comprising a three-phase inverter circuit with compensated current sensors for each output terminal. The load and control contacts between the DBC substrate and the PCB (not shown) are accomplished by spring contacts. The control PCB and the DC-link are designed by application engineers and can be

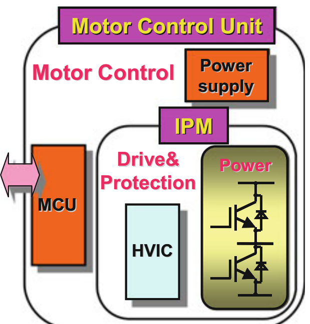


Fig. 15.18 Extension of an IPM to a complete motor control unit by adding a micro-controller and the power supply [Ara05]

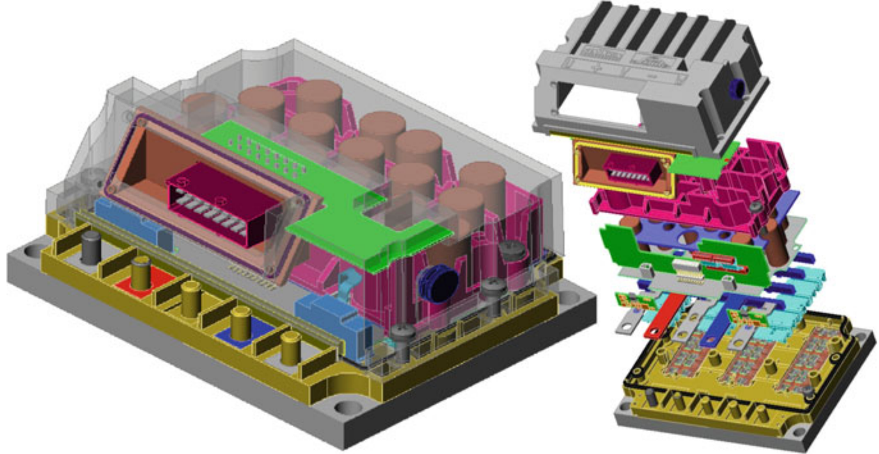


Fig. 15.19 Integrated power electronic system in one compact package with current and temperature sensors, DC-link capacitors, driver and micro-controller—the power rating is 13 kW

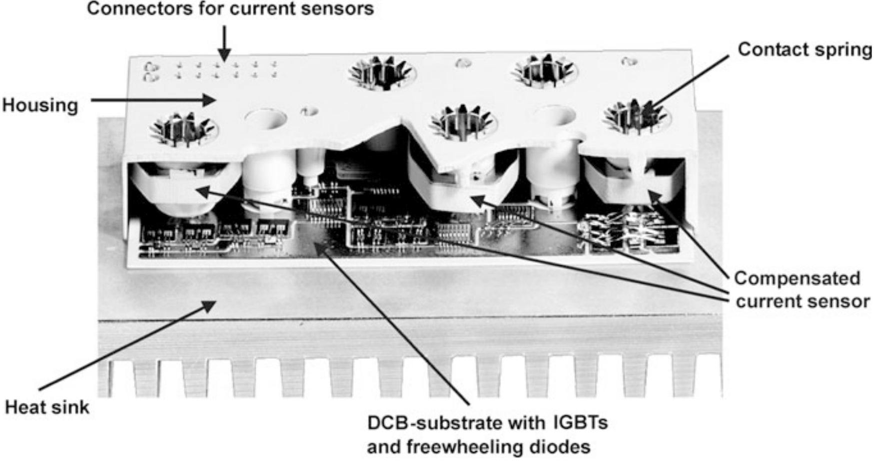


Fig. 15.20 Semikron SKiiP system with integrated compensated current sensors

adapted to the specific requirements of the application. The module design can easily be customized by different DBC layouts and by implementation of suitable power devices. This flexibility allows for a commercially successful module production for moderate quantities.

Most of the above presented examples of hybrid integration have successfully penetrated the market. They simplify the system design for application engineers by solving the problems of supplying and controlling the load current and the extraction of heat. Nonetheless, essential aspects of integration remain unsolved:

- The package outline of passive components aggravates their efficient integration. Progressive foil capacitors and transformers integrated in PCBs are capable to open new opportunities in the future.
- The requirement for an excellent thermal contact restricts the mounting plane of the power devices to the 2-dimensional surface of the DBC substrate. If the load current is routed only by current tracks on the substrate, it is difficult to provide short current paths, which supply paralleled chips with identical paths lengths in order to minimize parasitic effects and to prevent unbalance phenomena between parallel chips. Multi-contact internal bus bars in pressure contact systems could help to overcome this limitation in future architectures [Scn09].

A promising strategy is the exploitation of parasitic effects to implement new features by integration. An excellent example of this strategy is illustrated in Fig. 15.21, which shows a bus bar structure with an integrated high-frequency filter [Zha04].

For a power converter, bus bars are required to conduct the load current between the converter and the energy source. As state-of-the-art, the DC supply is implemented by conductive sheets or bars, which are separated by an intermediate insulation layer, thus forming a parasitic capacity adjacent to the switching devices.

The structure in Fig. 15.21 implements a $BaTiO_3$ dielectric insulation layer between the +DC and −DC bus bars. The bus bars in this example are formed by a thin Ni-layer on one side and a thicker Cu-layer on the other side, separated by an Al_2O_3 ceramic insulator. They are assembled with the Ni-layers towards the $BaTiO_3$ layer and the Cu-layer to the outside to form a DC bus bar structure. Low frequency currents are conducted primarily by the copper layer due to its lower electrical resistance. High-frequency currents on the other hand are squeezed into the Ni-layer by skin- and proximity effect. Therefore, the high-frequency content is damped stronger and the structure forms a low-pass filter. The equivalent circuit of this bus bar structure is a capacitor with a frequency-dependent resistor, which is increasing with the frequency. This example demonstrates that a skillful arrangement of different elements can make use of physical effects—which are usually considered as a handicap—to integrate a low pass filter into a bus bar structure with little additional effort.

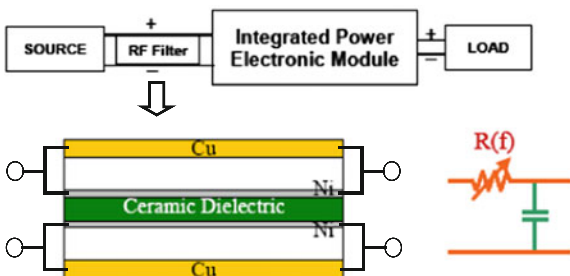


Fig. 15.21 Bus bar structure with integrated high-frequency filter [Zha04]

Combining multiple functions into single functional element is identified as synergy. Synergy effects represent a non-linear progress in the process of integration. They cannot be scheduled in a project plan or a road map, but when they are discovered, they generate an unexpected drive for further progress. Possible areas to search for synergy effects are:

– Synergy effects on the device level: Electrical functions of different devices are integrated in a single chip. An example is the reverse conducting IGBT, which merges the traditional IGBT and the freewheeling diode into one single chip.
– Synergy effects on the package level: Different functions are merged into one functional element. An example is the integration of passive components into the PCB.

A survey of the existent opportunities in integration for a 12 V/42 V DC–DC converter is presented in [Pop04]. Requirements with respect to cooling performance, electrical function and electro-magnetic interference are taken into account. This survey proposes a design, in which all elements are assembled on a single thermally conductive rail, which allows to reduce the number of components from 38 to merely 19. This would diminish the required system volume to approximately one-eighth. The survey constitutes an indication of the conceivable miniaturization of power electronic systems by integration.

While the integration of driver electronics and sensor functions has reached a status of maturity in the successful concept of the IPM, the integration of passive components remains in a state of infancy. Details of research results in this area are currently being implemented in new products, but they remain incoherent. The thermal limit of power electronic systems impedes a fast progress as accomplished in the area of microelectronics.

The progress in increasing power density and volume reduction of power electronic systems will primarily be depending on the successful development of power devices with reduced on-state and switching losses. There is considerable potential for further progress in this direction, as has been indicated in many respects in this chapter. Improved devices will operate at higher switching frequencies, which will diminish the size of passive components. This will presumably be the major driving factor for future progress. It will continuously facilitate the boundary conditions for system integration, an indispensable requirement for a reduction of system cost and thus a prerequisite for realizing the potential of increasing energy efficiency of advanced power electronic systems.

References

[And96] Andreini, A., Contiero, C., Glabiati, P.: BCD technologies for smart power ICs. In: Murati, B., Bertotti, F., Vignola, G.A. (eds.) Smart Power ICs. Springer, Berlin (1996)

[Ara05] Araki, T.: Integration of power devices – next tasks. In: Proceedings of EPE, Dresden (2005)

[Bor05] Boroyevich, D., van Wyk, J.D., Lee, F.C., Liang, Z.: A view at the future of integration in power electronics systems. In: Proceedings of PCIM, Nuremberg, pp. 11–20 (2005)

[Cai06] Cai, Y., Cheng, Z., Yang, Z., Tang, W.C.-W., Lau, K.M., Chen, K.J.: Monolithically integrated enhancement/depletion-mode AlGaN/GaN HEMT inverters and ring oscillators using CF4 plasma treatment. IEEE Trans. Electron Dev. **53**, 2223–2230 (2006)

[Che17] Chen, K.J.: Fluorine-implanted enhancement-mode transistors. In Meneghini, M., Meneghesso, G., Zanoni, E. (eds.) Power GaN Devices – Materials, Applications and Reliability. Springer, Switzerland (2017)

[Dic08] Dick, C.P., Hirschmann, D., Plum, T., Knobloch, D., De Doncker R.W.: Novel high frequency transformer configurations – amorphous metal vs. ferrites. In: Proceedings of IEEE PESC, 2008, pp. 4264–4269 (2008)

[Gie02] Giebel, T.: Grundlagen der CMOS-Technologie. Teubner Verlag, Stuttgart (2002)

[Grs08] Grasshoff, T., Reusser, L.: Integration of a new SOI driver into a medium power IGBT module package. In: Proceedings of PCIM Europe, Nuremberg (2008)

[Kin16] Kinzer, D., Oliver, S.: Monolithic HV GaN Power ICs. IEEE PELS Power Electron. Mag. **3**(3), 14–21 (2016)

[Kin16a] Kinzer D., Driving for zero switching loss power solutions. Presentation at PCIM Asia, Shanghai (2016) https://www.navitassemi.com/pcim-shanghai-keynote/

[Kin17] Kinzer, D.: GaN Power ICs at 1 MHz+: Topologies, Technologies and Performance, APEC 2017, PSMA Industry Session, Semiconductors (2017)

[Ler02] Lerner, R., Eckoldt, U., Knopke, J.: High voltage smart power technology with dielectric insulation. In: Proceedings of CIPS, pp. 83–88 (2002)

[Mot93] Motto, E.R.: New Intelligent Power Modules (IPMs) for motor drive applications. In: Proceedings of IEEE IAS, Toronto (1993)

[Mot99] Motto, E.R., Donlon, J.F., Iwamoto, H.: New power stage building blocks for small motor drives. In: Proceedings of Powersystems World Conference '99, pp. 343–349, Chicago (1999)

[Nee14] Neeb, C., Boettcher, L., Conrad, M., De Doncker, R.W.: Innovative and reliable power modules: a future trend and evolution of technologies. IEEE Ind. Electron. Mag. **8**(3), 6–16 (2014)

[Pop04] Popović, J., Ferreira, J.A.: Concepts for high packaging and integration efficiency. In: Proceedings of PESC, Aachen, pp. 4188–4194 (2004)

[Pop05] Popović, J., Ferreira, J.A., Waffenschmidt, E.: PCB embedded DC/DC 42/14 V converter for automotive applications. In: Proceedings of EPE, Dresden (2005)

[Pri96] Pribyl, W.: Integrated smart power circuits technology, design and application. In: Proceedings of ESSCIRC (1996)

[Rei16] Reiner, R., Waltereit, P., Weiss, B., Moench, S., Wespel, M., Müller, S., Quay, R., Ambacher, O.: Monolithically-integrated power circuits in high-voltage GaN-on-Si heterojunction technology. In: Proceedings of ISPS, Prague (2016)

[Scn02] Scheuermann, U., Tursky, W.: IPMs zwischen Modul und intelligenten leistungselektronischen Antriebssystemen. In: Proceedings of Fachtagung Elektrische Energiewandlungssysteme, Magdeburg, pp. 105–110 (2002)

[Scn09] Scheuermann, U.: Power module design without solder interfaces—an ideal solution for hybrid vehicle traction applications. In: Proceedings of APEC 2009, Washington D. C., pp. 472–478 (2009)

[Tih88b] Tihanyi, J.: Smart SIPMOS technology. In: Siemens Forschungs- und Entwicklungsberichte Bd.17 Nr.1, pp. 35–42. Springer, Berlin (1988)

[Waf05] Waffenschmidt, E., Ackermann, B., Ferreira, J.A.: Design method and material technologies for passives in printed circuit board embedded circuits. Special Issue on Integrated Power Electronics. IEEE Trans. Power Electron. **20**(3), 576–584 (2005)

[Waf05b] Waffenschmidt, E., Ackermann, B., Wille, M.: Integrated ultra thin flexible inductors for low power converters. In: Proceedings of PESC, Recife (2005)

[Wan15] Wang, H., Kwan, A.M., Jiang, Q., Chen, K.J.: A GaN pulse width modulation integrated circuit for GaN power converters. IEEE Trans. Electron Dev. **62**, 1143–1149 (2015)

[Wei16] Weiss, B., Reiner, R., Waltereit, P., Quay, R., Ambacher, O., Sepahvand, A., Maksimovic, D.: Soft-switching 3 MHz converter based on monolithically integrated half-bridge GaN-chip. In: Proceedings of IEEE 4th Workshop on Wide Bandgap Power Devices and Applications (WiPDA), pp. 215–219 (2016)

[Zha04] Zhao, L., van Wyk, J.D.: A high attenuation integrated differential mode RF EMI filter. In: Proceedings of CPES Power Electronics Seminar, Blacksburg, pp. 74–77 (2004)

Appendix A
Modeling Parameters of Carrier Mobilities in Si and 4H-SiC

A.1 Mobilities in Silicon

Can be well described by the Caughey-Thomas formula [Cau67]:

$$\mu = \mu_\infty + \frac{\mu_0 - \mu_\infty}{1 + (N/N_{ref})^\gamma},$$

see Fig. 2.12. The parameters μ_0, μ_∞ and N_{ref} at 300 K used for Fig. 2.12 have been determined by fitting the formula to the experimental carrier dependence at 300 K [Thu80a, Thu80b, Mas83]. For the temperature dependence of parameters, the concentration dependence at temperatures between 250 and 450 K and the temperature dependence of resistivity at various doping densities have been used [Li77, Li78, Swi87] taking into account the incomplete ionization around $10^{18}/\text{cm}^3$. The experimental results can be described fairly well using the following temperature dependent parameters:

Electrons:

$$\mu_0 = 1412 \cdot \left(\frac{300}{T}\right)^{2.28} /\text{cm}^2(Vs), \quad \mu_\infty = 66 \cdot \left(\frac{300}{T}\right)^{0.90} \text{cm}^2/(Vs) \tag{A.1}$$

$$N_{ref} = 9.7 \cdot 10^{16} \cdot \left(\frac{T}{300}\right)^{3.51} \text{cm}^{-3} \quad \gamma = 0.725 \cdot \left(\frac{300}{T}\right)^{0.270} \tag{A.2}$$

Holes:

$$\mu_0 = 469 \cdot \left(\frac{300}{T}\right)^{2.10} \text{cm}^2/(Vs), \quad \mu_\infty = 44 \cdot \left(\frac{300}{T}\right)^{0.80} \text{cm}^2/(Vs) \tag{A.3}$$

$$N_{ref} = 2.4 \cdot 10^{17} \cdot \left(\frac{T}{300}\right)^{4.13} \text{cm}^{-3} \quad \gamma = 0.70 \cdot \left(\frac{T}{300}\right)^{0.00} \tag{A.4}$$

© Springer International Publishing AG 2018
J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8

The parameters at 300 K differ considerably from original values in [Cau67] because the early measurements [Irv62] have been noticeably corrected later.

A useful formula for the *field and concentration dependence of mobilities* in silicon has been proposed by Scharfetter and Gummel [Scf69], it reads

$$\mu = \frac{\mu^{(0)}}{\left\{1 + \frac{N}{N/S + N_r} + \frac{(\mathsf{E}/A)^2}{\mathsf{E}/A + F} + \left(\frac{\mathsf{E}}{B}\right)^2\right\}^{1/2}} \tag{A.5}$$

N is again the concentration of donors or acceptors, $\mu^{(0)}$ are the respective Mobilities from Eqs. (A.1) to (A.3) for small $N < 1 \times 10^{14}$ cm^{-3}. The fitting parameters A, B, F, N_r and S for electrons and holes are

	N_r	S	A	F	B
Electrons	3×10^{16}	350	3.5×10^{8}	8.8	7.4×10^{3}
Holes	4×10^{16}	81	6.1×10^{8}	1.6	2.5×10^{4}

A.2 Mobilities in 4H-SiC

The mobilities in 4H-SiC are weakly anisotropic as noted in Sect. 2, but in modeling the mobilities the anisotropy is mostly neglected. According to [Scr94] the dependence on the *total doping density N and temperature* can be described by

$$\mu = \mu_\infty + \frac{\mu_0 \cdot (T/300)^\alpha - \mu_\infty}{1 + (N/N_{ref})^\gamma} \tag{A.6}$$

with the following parameter values:

Electrons: $\mu_0 = 947\ \mathrm{cm^2/(Vs)}$ $\quad \mu_\infty = 0,$
$N_{ref} = 1.94 \times 10^{17}\mathrm{cm^{-3}}$
$\alpha = -2.15$ $\quad \gamma = 0.61$

Holes: $\mu_0 = 124\,\mathrm{cm^2/(Vs)},$ $\quad \mu_\infty = 15.9\,\mathrm{cm^2/(Vs)}$
$N_{ref} = 1.76 \times 10^{19}\mathrm{cm^{-3}}$
$\alpha = -2.15$ $\quad \gamma = 0.34$

Appendix B
Correlates to Recombination Centers

B.1 Effective Degeneracy Factors

Measurements of the emission rates, for example the electron emission rate of an acceptor level, are presented in the form

$$e_n = A\left(\frac{T}{300}\right)^{m'} exp\left(-\frac{\Delta E'}{kT}\right) \tag{B.1}$$

where A and E' are constants and the exponent m' of the power term is usually 2. Although we talk at first of the electron emission rate of an acceptor level, the equations obtained below hold in the same form for the hole emission rate of a donor level. Hence the subscripts 'n,a' or 'p,d' are omitted in equations holding for both cases. The choice $m' = 2$ agrees with the ideal theoretical value of the exponent of the term $c_{n,a}N_c \sim T^m$ in the detailed balance Eq. (2.61), but really m differs often considerable from 2. Equating the right hand sides of (2.61a) and (B.1) one obtains

$$B\left(\frac{T}{300}\right)^{m} g\ exp\left(-\frac{\Delta E}{kT}\right) = A\left(\frac{T}{300}\right)^{m'} exp\left(-\frac{\Delta E'}{kT}\right) \tag{B.2}$$

where $B \equiv c_{n,a}(300)N_c(300)$ and $\Delta E \equiv E_c - E_r$. By this equation the activation energy ΔE can be determined as function of T. With

$$g' \equiv \frac{A}{B}\left(\frac{T}{300}\right)^{-\Delta m} \tag{B.3}$$

where $\Delta m \equiv m - m'$, an effective degeneracy factor g' is introduced, which for $\Delta m = 0$ and $\Delta E = \Delta E' = const$ is equal to the intrinsic degeneracy g introduced in Sect. 2.5. Using the degeneracy factor g' Eq. (B.2) takes the form

$$g\ exp\left(-\frac{\Delta E}{kT}\right) = g' exp\left(-\frac{\Delta E'}{kT}\right) \tag{B.4}$$

© Springer International Publishing AG 2018
J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8

Multiplied with N_C the left hand side is identical with Eq. (2.61a) for $n_r = n_a$, the right hand side equals Eq. (2.79a). Solved for ΔE Eq. (B.4) writes:

$$\Delta E(T) = \Delta E' + kT \ \ln\left(\frac{g}{g'}\right) \tag{B.5}$$

The derivative is:

$$\frac{d\Delta E}{dT} = k\left\{\ln\left(\frac{g}{g'}\right) + \Delta m\right\} \tag{B.6}$$

The second derivative $d^2\Delta E/dT^2 = k\Delta m/T$ is physically irrelevant, since m' has only a meaning in connection with the constant A in (B.1) and for itself is to some extent arbitrary. The measurements represented by (B.1) together with Eqs. (2.61), (2.61a) give hence only information about the *linear* variation of ΔE. If T_0 is a central point of the experimental temperature region one obtains using (B.5) and (B.6) for $\Delta E(T_0)$ and $d\Delta E/dT(T_0)$ the following equation

$$\Delta E(T) = \Delta E' - \Delta mkT_0 - \alpha_T T \tag{B.7}$$

where the negative temperature coefficient α_T is given by

$$\alpha_T = k\left\{\ln\left(\frac{g'(T_0)}{g}\right) - \Delta m\right\} \tag{B.8}$$

Equation (B.8) is used to determine α_T from g' and Δm.
Solved for g' the equation (B.8) reads

$$g' = \underbrace{exp(\Delta m)}_{g_\Delta}\ \underbrace{exp\left(\frac{\alpha_T}{k}\right)}_{\chi}\ g \tag{B.9}$$

If $\alpha_T, \Delta m > 0$, $\Delta E'$ is larger than ΔE according to (B.7), hence a degeneracy factor $g' > g$ is necessary to compensate the use of $\Delta E'$ instead of ΔE. Due to the component $g_\Delta \equiv exp(\Delta m)$, the same measurements can result in different degeneracy factors g', depending on the choice of m'.

The factor

$$\chi \equiv exp\left(\frac{\alpha_T}{k}\right) \tag{B.10}$$

is called 'entropy factor', which has the following reason: The energy gap E_g and the activation energy ΔE are really free enthalpies (Gibbs free energy) [Vec76]. According to the fundamental thermodynamic relationship $\Delta S = -\partial\Delta E/\partial T$ the

negative derivative of ΔE with respect to temperature, i. e. α_T, equals the change of entropy ΔS, associated in our case with carrier emission. Hence (B.10) can be written

$$\chi = exp(\Delta S/k)$$

In [Vec76] the physical causes of entropy change are discussed.

As mentioned, the above equations hold also for the hole emission rate $e_{p,d}$ of a donor level, where the constant B is defined as the 300 K-value of $c_{p,d}N_V$ and m as the exponent of the temperature dependency $c_{p,d}N_V \sim T^m$. Hence Eq. (2.79b) is obtained from (B.4) for the hole concentration $p_d \equiv p_r$ of a donor level.

In some cases, for example for the gold acceptor level, both the electron and hole emission rate of a level have been measured and hence besides $g'_{n,a}$ also the degeneracy factor $g'_{p,a}$ is known immediately. The above equations turn into equations for the hole emission rate of an acceptor level by replacing the degeneracy factors g and g' by their inverse, $g \rightarrow 1/g$, $g' \rightarrow 1/g'$. This results in

$$g'_{p,a} \equiv \frac{B}{A}\left(\frac{T_0}{300}\right)^{\Delta m_p} = \frac{g}{exp\left(\Delta m_p\right)exp\left(\alpha_{T,p}/k\right)} \tag{B.11}$$

where $B \equiv c_p(300)N_V(300)$ and $\Delta m_p \equiv m_p - m'_p$, $\alpha_{T,p} \equiv -d\Delta E_{p,a}/dT$. Since $\Delta E_{n,a} + \Delta E_{p,a} = E_g$, the sum of the (negative) temperature coefficients of ΔE_n and E_p is given by: $\alpha_{T,n} + \alpha_{T,p} = -dE_g/dT \equiv \alpha_{Eg}$. Using this the following relationship between the effective degeneracy factors of the electron and hole emission rate of an acceptor level results:

$$\frac{g'_{n,a}}{g'_{p,a}} = exp\left(\Delta m_n + \Delta m_p\right)exp\left(\frac{\alpha_{Eg}}{k}\right) \tag{B.12}$$

In contrast to the intrinsic degeneracy factor g, the effective degeneracy factor g'_n *of a level for electrons and the effective degeneracy factor* g'_p *of the level for holes differ from one another.* From Eq. (2.9) and Table 2.1 in Chap. 2 one obtains for silicon at 350 K: $\alpha_{Eg} = 2.76 \times 10^{-4}$ eV/K and $exp(\alpha_{Eg}/k)$ = 24.7. By this factor the degeneracy factor $g'_{p,a}$ is obtained to be smaller than $g'_{n,a}$, if the exponents m_n and m'_n of $e_{n,a}$ as well as m_p and m'_p of $e_{p,a}$ agree ($\Delta m_n = \Delta m_p = 0$) or if $\Delta m_p = -\Delta m_n$. For a donor level the left hand side of (B.12) has to be replaced by $g'_{p,d}/g'_{n,d}$, hence the electron degeneracy factor $g'_{n,d}$ is *smaller* than $g'_{p,d}$. A cause of this asymmetry is that g'_n and g'_p depend in an inverse manner on the term $g_\Delta.\chi$.

B.2 Charge of Deep Impurities with Two Levels

As in Sect. 2.7.2 b a trap with a donor and an acceptor level is assumed. From the steady-state condition $R_n = R_p$ for the donor level one obtains as concentration ratio of positively charged and neutral centers

$$\frac{N_r^+}{N_r^0} = \frac{c_{n,d} n_d + c_{p,d} p}{c_{n,d} n + c_{p,d} p_d} \equiv A \tag{B.13}$$

The steady state condition $R_n = R_p$ for the acceptor level yields

$$\frac{N_r^-}{N_r^0} = \frac{c_{n,a} n + c_{p,a} p_a}{c_{n,a} n_a + c_{p,a} p} \equiv \mathrm{B} \tag{B.14}$$

Hence the concentrations of the three charge states are related to each other as

$$N_r^+ : N_r^0 : N_r^- = A : 1 : B \tag{B.15}$$

Since the total trap concentration $N_r = N_r^0(1 + A + B)$, it follows

$$N_r^+ = \frac{A}{1 + A + B} N_r, \quad N_r^- = \frac{B}{1 + A + B} N_r \tag{B.16}$$

In a neutral n region the electron and hole concentrations are hence related by the equation

$$n = N_D + p + \frac{A - B}{1 + A + B} N_r, \tag{B.17}$$

which can be solved by iteration.

B.3 Recombination Parameters of Gold in Silicon

For the **temperature dependence** we use the rule that the capture cross section $\sigma_{i,j} \equiv c_{i,j}/v_{i,therm} \sim c_{i,j}/\sqrt{T}$ (the subscript i stands for n or p and j for a or d) of neutral centers is temperature-independent. This means a square root dependency of the capture probabilities $c_{n,a}$ and $c_{p,d}$. Capture rates of attractive centers, in our case $c_{p,a}$ and $c_{n,d}$, decrease with increasing T. The temperature dependency of $c_{p,a}$ is adopted from [Han87]. For $c_{n,d}$ the proportionality to T^{-2} of [Scm82] is preferred to the $T^{-3/2}$ dependency of [WuP82] since it agrees better with the observed T^2 dependency of τ_{HL} [Sco76].

Room temperature values: The **capture rate $c_{p,a}$** is taken from Fairfield and Gokhale [Fai65], whose value is not much different from that of [WuP82]. Because the other capture rates are determined in relation to $c_{p,a}$, this capture probability fixes the relation between the gold concentration and the lifetime. The **capture rate $c_{n,a}$** is determined from $c_{p,a}$ and the emission rates $e_{n,a}$, $e_{p,a}$ of the acceptor level, using the detailed balance Eq. (2.6),

$$c_{n,a}c_{p,a} = e_{n,a}e_{p,a}/n_i^2 \tag{B.18}$$

This is possible, because for the Au acceptor level both emission rates have been measured as function of temperature [Sah69, Eng75, Par72]. We use the emission rates of Sah et al [Sah69], since they are expressed by fitting formulae and their hole emission rate takes an intermediate position between the strongly differing results of the other papers. If n_i^2 is calculated from Eqs. (2.6), (2.8) and (2.9), one obtains at 300 K: $c_{n,a} = 5.6 \times 10^{-9}$ cm^3/s. This value is preferred against smaller ones [Fai65, WuP82, LuS86] also, because the maximum of the $\tau_p(p)$-function obtained for $N_D = 5 \times 10^{13}$ cm^{-3} and the minimum obtained for $N_D = 5 \times 10^{16}$ cm^{-3} shown in Fig. 2.20 would be else considerably more pronounced. In agreement with [Fai65, WuP82] the proposed $c_{n,a}$-value is more than an order of magnitude smaller than $c_{p,a}$. With the quoted temperature dependences of $c_{n,a}, c_{p,a}$ the left hand side of (B.18) decreases a little slower with increasing temperature than the right hand side, but this is within the error limits of the emission rates.

The **capture rate $c_{n,d}$** is determined (essentially) from criterion *I* in Sect. 2.7.2 c), the ratio of high-level to low-level lifetime in n–Si. From criterion III and $c_{n,a} \ll c_{p,a}$ one obtains using Eq. (2.82) that the relation $c_{n,d} \ll c_{p,d}$ must hold. Due to these inequalities the high-level lifetime is determined (approximately) solely by the donor level and furthermore is equal to the low-level electron lifetime in a p-region with $p_0 \gg p_d$, namely $\tau_{n0}^{(d)} = 1/\left(N_r c_{n,d}\right)$. Since the low-level hole lifetime in n-silicon with $n_0 \gg n_a$ (= 2.00×10^{11} cm^{-3}) equals $\tau_{p0}^{(a)} = 1/\left(N_r c_{p,a}\right)$, the ratio of high-level to the low-level hole lifetime in n-silicon is given by $\tau_{HL}/\tau_{p,LL} \equiv \gamma \approx c_{p,a}/c_{n,d}$, hence $c_{n,d} \approx c_{p,a}/\gamma$. Using the experimental average value γ = 5.3 the $c_{n,d}$-value of Table 2.3 results calculating accurately. It is a factor 2.7 smaller than that of [Fai65], but a factor 3.9 higher than the value of [WuP82].

The capture rate $c_{p,d}$ must satisfy the relation $c_{p,d} \gg c_{n,d}$ at 300 K very well, in order to fulfill it also at lower temperatures where it is reduced according to $c_{p,d}/c_{n,d} \sim \mathrm{T}^{2.5}$. For $c_{p,d} > c_{p,a}$ however the lifetime τ_p in n–Si with $\mathrm{n}_0 \gg p_d$ shows the minimum shown in Fig. 2.20 for $N_D = 5 \times 10^{16}$ cm^{-3}. This follows from the equation

$$\tau_p = \frac{1}{N_r c_{p,a}} \frac{1 + \frac{c_{p,a}}{c_{n,a}}\left(1 + \frac{c_{p,d}}{c_{n,d}}\frac{p}{n}\right)\frac{p}{n}}{1 + \frac{c_{p,d}}{c_{n,a}}\frac{p}{n}} \quad (\text{for } \mathrm{n}_0 \gg \mathrm{p}_\mathrm{d}, \mathrm{n}_\mathrm{a}) \tag{B.19}$$

which is obtained in this case from (2.83). Since a strong minimum would not agree with the measurements of Hangleiter [Han87], very high values of $c_{p,d}$ are not allowed. As a compromise between this restriction and the requirement $\frac{c_{p,d}}{c_{n,d}} \gg 1$ we choose $c_{p,d}$ = 2.8 × 10^{-7} cm^3/s. The minimum is then only flat, whereas $\frac{c_{p,d}}{c_{n,d}}$ is still high. The chosen $c_{p,d}$-value is a factor 5 higher than obtained from the cross section $\sigma_{p,d}$ of [WuP82] determined at T < 200 K. The emission rates cannot be used to determine the $c_{p,d}$ using (B.18), because only the hole emission rate of the donor level has been measured [Sah69, Pal74, LuN87].

For the effective degeneracy factor of the **acceptor level for electrons** one has the constants A = 1.77 × 10^{12}/s [Sah69], $B \equiv c_{n,a}(300) \cdot N_C(300) = 1.602 \cdot 10^{11}\, s^{-1}$ and $\Delta m = 0.5 + 1.58 - 2 = 0.08$. At $T_0 = 350\,\text{K}$, the middle of the interesting temperature region, one obtains from (B.3) $g'_{n,a} = 10.9$. This degeneracy factor is used in Eq. (2.79a) together $\Delta E'_{n,a} = 0.5472\text{eV}$ [Sah69] to calculate n_a and $p_a = n_i^2/n_a$. From (B.8) the negative temperature coefficient of $\Delta E_{n,a}$ is obtained with $g = g_a = 4$ to be $\alpha_{T,a} = 0.795 \cdot 10^{-4}\text{eV/K}$. This is 29% of $-dE_g/dT$ at 350 K obtained from Eq. (2.9) together with the parameters of Table 2.1. The real activation energy at 350 K follows from (B.7) as $\Delta E_{n,a} = 0.5472 - 0.0336eV = 0.5136\text{eV}$. For the hole emission rate of the acceptor level Sah et al. obtained $A = 5.23 \cdot 10^{11} s^{-1}$ [Sah69], which with $B \equiv c_{p,a}(300)N_V(300) = 3.57 \cdot 10^{12}\, s^{-1}$ and $\Delta m_p = -1.7 + 1.85 - 2 = -1.85$ in (B.11) yields for the **effective hole degeneracy factor** $g'_{p,a} = 5.13$. The linear temperature coefficient of $\Delta E_{p,a} = E_a - E_V$ following from (B.11) with this $g'_{p,a}$ is

$$\alpha_{T,p} = k\left\{\ln\left(g_a/g'_{p,a}\right) - \Delta m_p\right\} = 1.38 \cdot 10^{-4}\text{eV/K}.$$

The sum $\alpha_{T,n} + \alpha_{T,p} = 2.17 \cdot 10^{-4}\text{eV/K}$ amounts to 79% of the value obtained at 350 K from Eq. (2.9). Considering the quite different experimental methods this is seen as a satisfactory agreement.

For the hole emission rate of the **donor level** the constant A given in [Sah69] is $A = 2.43 \cdot 10^{13} s^{-1}$. Together with $B = c_{p,d}(300)N_V(300) = 8.68 \cdot 10^{13} s^{-1}$ and $\Delta m = 0.5 + 1.85 - 2 = 0.35$ one obtains from (B.3) at 350 K: $g'_{p,d} = 2.65$. Since the intrinsic degeneracy factor $g_d = 2$, this means according to (B.8) that the distance of the donor level from the valence band is nearly constant. The effective activation energy to be used in (2.79b) is $\Delta E'_{p,d} = 0.3450\text{eV}$ [Sah69].

Appendix C
Avalanche Multiplication Factors and Effective Ionization Rate

C.1 Multiplication Factors

The following derivation is adopted from McIntyre [McI66]. We consider a primary electron-hole pair at the point x in the depletion layer of a reverse biased pn-junction and ask for the multiplication factor$M(x)$, which is the total number of pairs generated in the avalanche process initiated by the primary pair including the ionization by secondary carriers and so on. In the diode orientation of Fig. 3.14 the electrons will be swept to the left (the neutral n region) and holes to the right (p region). In travelling a path of length dx the probability that the electron will generate an electron-hole pair is $\alpha_n \cdot dx$. Similarly the hole will generate on average $\alpha_p \cdot dx$ pairs on a path of length dx. Each of these secondary pairs generated at a point x' will itself generate carrier pairs on its path and experiences a multiplication by a factor $M(x')$. Since this adds up to the multiplication of the primary pair, one has

$$M(x) = 1 + \int_0^x \alpha_n M(x')\, dx' + \int_x^w \alpha_p M(x')\, dx' \tag{C.1}$$

where α_n and α_p depend over the field strength E on x'. Differentiating one obtains

$$\frac{dM}{dx} = (\alpha_n - \alpha_p)\, M(x) \tag{C.2}$$

This has the solution

$$M(x) = M(0)\, exp\left(\int_0^x (\alpha_n - \alpha_p) dx'\right) \tag{C.3}$$

$$= M(w)\, exp\left(-\int_x^w (\alpha_n - \alpha_p)\, dx'\right) \tag{C.4}$$

© Springer International Publishing AG 2018
J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8

If (C.3) is inserted into the right hand side of (C.1), one obtains for $x = 0$

$$M(0) = M_p = \frac{1}{1 - \int\limits_0^w \alpha_p \, exp\left(\int\limits_0^x (\alpha_n - \alpha_p)\, dx'\right) dx} \tag{C.5}$$

Since this is the multiplication factor of the primary carriers at the boundary $x = 0$ of the space charge layer, where holes from the neutral n region enter the depletion layer, M(0) is identical with the multiplication factor M_p of the saturation hole current density j_{ps}. Similarly, substituting (C.4) into the right hand side of (C.1) one obtains for $x = w$:

$$M(w) = M_n = \frac{1}{1 - \int\limits_0^w \alpha_n exp\left(-\int\limits_x^w (\alpha_n - \alpha_p)\, dx'\right) dx} \tag{C.6}$$

This is the multiplication factor for the electrons entering the depletion layer at $x = w$ from the neutral p region.

For arbitrary x it follows from (C.3) and (C.5)

$$M(x) = \frac{exp\left(\int\limits_0^x (\alpha_n - \alpha_p)\, dx'\right)}{1 - \int\limits_0^w \alpha_p exp\left(\int\limits_0^{x'} (\alpha_n - \alpha_p)\, dx''\right) dx'} \tag{C.7}$$

and a mathematically identical expression follows from (C.4) and (C.6). Assuming a homogeneous thermal generation G in the depletion layer, the mean value of $M(x)$ is the multiplication factor for the current j_{sc} generated in the space charge region:

$$M_{sc} = \bar{M} = \frac{\frac{1}{w}\int\limits_0^w exp\left(\int\limits_0^x (\alpha_n - \alpha_p) dx'\right) dx}{1 - \int\limits_0^w \alpha_p exp\left(\int\limits_0^x (\alpha_n - \alpha_p)\, dx'\right) dx} \tag{C.8}$$

Since according to (C.3) and (C.4)

$$\frac{M_n}{M_p} = exp\left(\int\limits_0^w (\alpha_n - \alpha_p) dx\right) \tag{C.9}$$

it follows from (C.5) and (C.8) that M_n, M_p and M_{sc} tend to infinity at the same field distribution and voltage. If $\alpha_n > \alpha_p$, the multiplication factors obey the inequality

$M_n > M_{sc} > M_p$ (below infinity). Numerical results of the above equations for a silicon pn-junction are shown in Fig. 3.15.

Avalanche multiplication is always an essential factor determining the breakdown voltage of power devices, for power diodes containing only one pn-junction it is the only determinant.

C.2 Effective Ionization Rate and Breakdown Condition for Diodes

In diodes, the breakdown voltage is reached where the multiplication factors tend to infinity. Using the ionization integral in the denominator of (C.5) the breakdown condition can be written

$$I_p \equiv \int_0^w \alpha_p exp\left(\int_0^x (\alpha_n - \alpha_p)dx'\right)dx = 1 \tag{C.10}$$

For a field-independent ratio $\alpha_n/\alpha_p = \gamma$, one obtains from (C.10)

$$\begin{aligned} 1 &= \int_0^w \alpha_p exp\left(\int_0^x (\gamma - 1)\alpha_p dx'\right)dx \\ &= \frac{1}{\gamma - 1}\left(exp\left(\int_0^w (\gamma - 1)\alpha_p dx\right) - 1\right) \end{aligned} \tag{C.11}$$

since $\int_0^x f'(x')\, exp\,(f(x')\, dx' = [exp\,(f(x))]_0^x$. From (C.11) one has $\gamma = exp\,[\int_0^w\,(\gamma - 1)\,\alpha_p dx]$ which can be written

$$\int_0^w \frac{\alpha_n - \alpha_p}{\ln\,(\alpha_n/\alpha_p)}dx = \int_0^w \alpha_{eff}\; dx = 1 \tag{C.12}$$

where the effective ionization rate is defined as

$$\alpha_{eff} = \frac{\alpha_n - \alpha_p}{\ln\,(\alpha_n/\alpha_p)} \tag{C.13}$$

Hence this α_{eff} can be used together with the condition (C.12) to calculate the breakdown voltage of pn-junctions. In the relative small upper field range which contributes significantly to the integral, the preposition of constant α_n/α_p is in most cases sufficiently fulfilled.

References for Appendices A, B and C

[Cau67] Caughey, D.M., Thomas, R.E.: Carrier mobilities in silicon empirically related to doping and field. Proc. IEEE **23**, 2192–93 (1967)

[Eng75] Engström, O., Grimmeis, H.G.: Thermal activation energy of the gold ac-ceptor level in silicon. J. Appl. Phys. **46**, pp. 831–837 (1975)

[Fai65] Fairfield, J.M., Gokhale, B.V.: Gold as a recombination centre in silicon. Solid-St. Electron. **8**, 685–691 (1965)

[Han87] Hangleiter, A.: Nonradiative recombination via deep impurity levels in silicon: Experiment. Phys. Rev. B **15**, 9149–9161 (1987)

[Irv62] Irvin, J.C.; Resistivity of bulk silicon and of diffused layers in silicon. Bell Syst. Tech. J. **41**, 387–410 (1962)

[Li77] Li, S.S., Thurber, W.R., The dopant density and temperature dependence of electron mobility and resistivity in n-type silicon. Solid St. Electron. **20**, 609–616 (1977)

[Li78] Li, S.S.: The dopant density and temperature dependence of hole mobility and resistivity in boron doped silicon. Solid St. Electron. **21**, 1109–1117 (1978)

[LuN87] Lu, L.S., Nishida, T., Sah, C.T.: Thermal emission and capture rates of holes at the gold donor level in silicon. J. Appl. Phys. **62**, 4773– 4780 (1987)

[LuS86] Lu, L.S., Sah C.-T.: Electron recombination rates at the gold acceptor level in high-resistivity silicon, J.Appl. Phys., **59**, 173–176 (1986)

[Mas83] Masetti, G., Severi, M., Solmi, S.: Modeling of carrier mobility against concentration in Arsenic-, Phosphorus-, and Boron-doped Silicon. IEEE Trans. Electron Devices **ED-30**(7), 764–769 (1983)

[McI66] McIntyre, R.J.: Multiplication noise in uniform avalanche diodes. IEEE Trans. Electron. Dev. **ED-13**, 164–168 (1966)

[Pal74] Pals, J. A.: Properties of Au, Pt, Pd, and Rh levels in silicon measured with a constant capacitance technique, Solid-St. Electronics, **17**, 1139–1145 (1974)

[Par72] Parrillo, L. C., Johnson, W. C.: Acceptor state of gold in silicon – Resolution of an anomaly, Appl. Phys. Lett., **20**, 101–106 (1972)

[Sah69] Sah, C.T., Forbes, L., Rosier, L.I., Tasch, A.F. Jr, Tole, A.B.: Thermal emission rates of carriers at gold centers in silicon. Appl. Phys. Lett. **15** 145–148 (1969)

[Scf69] Scharfetter, D.L., Gummel, H.K.: Large-signal analysis of a silicon Read Diode oscillator. IEEE Trans. Electron Dev. **ED-16**, 64–77 (1969)

[Scm82] Schmid, W., Reiner, J.: Minority carrier lifetime in gold-diffused silicon at high carrier concentrations. J. Appl. Phys. **53**, 6250–6252 (1982)

[Sco76] Schlangenotto, H., Maeder, H., Dziewior, J.: Neue Technologien für Si-lizium-Leistungsbauelemente - Rekombination in hoch dotierten Emitterzo-nen. Research Report T 76–54, German Ministry of Research and Technology (1976)

[Scr94] Schaffer, W.J., Negley, G.H., Irvine, K.G., Palmour, J.W., Conductivity anisotropy in epitaxial 6H and 4H SiC. Mater. Res. Soc. Symp. Proc. **339** 595–600 (1994)

[Swi87] Swirhun, S.E.: Characterization of majority and minority carrier transport in heavily doped silicon. Ph.D. Dissertation, Stanford University (1987)

[Thu80a] Thurber, W.R., Mattis, R.L., Lium Y.M., Filliben, J.J.: Resistivity-dopant density relationship for phosporous-doped silicon. J. Electrochem. Soc. **127** 1807–1812 (1980)

[Thu80b] Thurber, W.R., Mattis, R.L., Liu, Y.M., Filliben, J.J.: Resistivity-dopant density relationship for boron-doped silicon. J. Electrochem. Soc. **127**, 2291–2294 (1980)

[Vec76] Van Vechten, J.A., Thurmond, C.D.: Entropy of ionization and temperature variation of ionization levels of defects in semiconductors. Phys. Rev. B **14**, 3539–3550 (1976)

[WuP82] Wu, R.H., Peaker, A.R.: Capture cross sections of the gold and acceptor states in n-type Czochralski silicon. Solid-St. Electron. **25**, 643–649 (1982)

Appendix D
Thermal Parameters of Important Materials in Packaging Technology

	Thermal conductivity (W/mmK)	Thermal capacity (J/mm^3K)	Thermal expansion ($10^{-6}/K$)	Source
Semiconductors				
Si	0.13	1.65×10^{-3}	2.6	[IOF01]
GaAs	0.055	1.86×10^{-3}	5.73	[IOF01]
SiC	0.37	2.33×10^{-3}	4.3	[IOF01]
GaN	0.13		5.6	[Qua08,Yam11]
Insulators				
SiO_2	0.0014	1.4×10^{-3}	0.55	[Sze81]
Al_2O_3	0.024	3.02×10^{-3}	6.8	Hoechst
AlN	0.17	2.44×10^{-3}	4.7	Hoechst
Si_3N_4	0.07	2.10×10^{-3}	2.7	Toshiba
BeO	0.251	2.98×10^{-3}	9	Brush-Wellman
Epoxyd	0.003		–	DENKA-TH1
Polyimid	3.85×10^{-4}		–	Kapton CR
Metals				
Al	0.237	2.43×10^{-3}	23.5	
Cu	0.394	3.45×10^{-3}	17.5	
Mo	0.138	2.55×10^{-3}	5.1	
Composite materials				
AlSiC	0.2	2.21×10^{-3}	7.5	
Solders				
Sn	0.063	1.65×10^{-3}	23	Demetron
SnAg(3.5)	0.083	1.67×10^{-3}	27.9	Demetron
SnPb(37)	0.07		24.3	Doduco 1/89
Interconnection layers				
Ag sinter layer	0.25	2.1×10^{-3}	18.9	[Thb06]
Thermal grease				
Wacker P 12	8.1×10^{-4}	2.24×10^{-3}	–	Wacker

© Springer International Publishing AG 2018

J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8

Appendix E
Electric Parameters of Important Materials in Packaging Technology

	Resistivity (25 °C) (μΩcm)	Dielectric constant ($1/\varepsilon_0$)	Critical field strength (kV/cm)	Source
Semiconductors				
Si	*	11.7	150–300	
GaAs	*	12.9	400	
SiC	*	9.66	3000	
GaN	*	9.5	3000	[Qua08]
Insulators				
SiO_2	10^{20}–10^{22}	3,9	4000–10,000	[Sze81]
Al_2O_3	10^{18}	9.8	150	Hoechst
AlN	10^{20}	9.0	200	Hoechst
Si_3N_4	10^{19}	8	150	Kyocera
BeO	10^{21}	6.7	100	Brush-Wellman
Epoxyd		7.1	600	DENKA-TH1
Polyimid		3.9	2910	Kapton CR
Metals				
Al	2.67	–	–	
Cu	1.69	–	–	
Mo	5.7	–	–	
Composite materials				
AlSiC	≈40	–	–	
Solders				
Sn	16.1	–	–	Demetron
SnAg(3.5)	13.3	–	–	Demetron
SnPb(37)	13.5	–	–	Doduco 1/89
Interconnection layers				
Ag sinter layer	1.6	–	–	
Thermal grease				
Wacker P 12	5×10^{15}			Wacker

* doping dependent – not defined

© Springer International Publishing AG 2018
J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8

References for Appendices D and E

[IOF01] Ioffe.: Physical Technical Institute, St. Petersburg, Russia (2001). http://www.ioffe.rssi.ru/SVA/NSM/Semicond/

[Thb06] Thoben, M., Hong, H., Hille, F.: Hoch-zeitaufgelöste Zth-Messungen an IGBT-Modulen. Kolloquium Halbleiter-Leistungsbauelemente, Freiburg (2006)

[Qua08] Quay, R.: Gallium nitride electronics, Springer, Berlin, Heidelberg (2008)

[Sze81] Sze, S.M.: Physics of semiconductor devices. Wiley, New York (1981)

[Yam11] Yam, F.K., Low, L.L., Oh, S.A., Hassan, Z.: Gallium nitride: an overview of structural defects. In: Predeep, P. (eds.) Optoelectronics - Materials and Techniques, (2011). doi: 10.5772/19878

Index

A
Abrupt junction, 104–110, 145–146
Acceptor, 35–43
Acceptor level, 37, 64–66
  definition, 63
Accumulation layer, 195, 342, 352, 405, 406
A-center, 80, 81, 187
Activation, 175
Activation energy, 36–39
Active region, 297, 306, 307, 621, 622
Al_2O_3, 23, 191, 439–444, 479, 480, 522, 524, 684, 703
AlGaN, 23, 191–194, 378–384, 673
Alpha particle *See* Irradiation with He ions
AlN, 156, 191, 439, 443, 455, 515, 702
AlSiC, 440, 444, 456–458, 520, 703
Aluminium (Al)
  conductivity, 703
  diffusion, 165, 166, 316
  energy level, 36
  solubility, 165, 166
Ambipolar diffusion constant, 211, 212
Ambipolar diffusion length, 216, 324
Amorphization, 173
Amorphous hydrated carbon (a-C:H), 182
Amplifying gate, 324
Annealing, 186, 188, 368
Anode short, 333, 334, 416
Antimony (Sb)
  diffusion, 165, 166
  energy level, 36
  solubility, 165, 166
Arrhenius factor, 165, 522, 526
Asymmetric junction, 108, 109, 120, 137, 145
Auger coefficient, 60
Auger generation, 82
Auger lifetime, 61
Auger recombination, 58, 60, 139, 141, 220, 303
Avalanche breakdown *See* Breakdown voltage
Avalanche center, 178
Avalanche multiplication, 81, 85, 126, 127, 128, 135
Avalanche rating, 591

B
Band gap, 21, 26, 29, 31, 32, 61
  temperature dependency, 28, 29
Band gap narrowing, 42, 43, 140, 143, 221, 303
BARITT diode, 648
Basic semiconductor equations, 88, 89
Beryllium oxide, 443−445, 703
Bidirectional blocking IGBT, 413
Bipolar transistor, 9, 139, 295ff, 309, 310, 394, 395, 398, 404, 603, 671
Blocking current *See* Leakage current
Blocking voltage *See* Breakdown voltage
Boltzmann distribution, 27, 28, 55, 103, 114, 273
Boltzmann factor, 28, 115, 214
Bond wire, 463, 476, 477, 511, 654
Bond wire lift-off, 513
Breakdown condition, 85, 699
Breakdown voltage, 83, 85, 122−135, 178−180, 203−209
Break-over triggering, 317
Breakover voltage, 320
Buffer layer, 337, 398, 594, 613
Built-in voltage, 103, 104, 105, 112, 119, 120, 204, 213, 408
Buried layers, 189, 671

C
CAL-diode, 231, 256, 258, 259, 416
CanPak *See* DirectFET
Capsule, 429, 431
Capture coefficients, 186

© Springer International Publishing AG 2018
J. Lutz et al., *Semiconductor Power Devices*,
https://doi.org/10.1007/978-3-319-70917-8

Capture probability, 63
Carbon content, 149, 151
Carrier-carrier scattering, 224
Carrier diffusion length, 94
Carrier lifetime, 66
  temperature dependency, 86, 135
Cascode, 372
Cauer-model, 452, 453
CAVET, 385
Cell pitch, 405, 546
Channel conductivity, 393
Channel mobility, 374
Channel resistance, 263
Channeling, 172–175
Charged center *See* Charged traps
Charge density, 88
Charged traps, 68
Chynoweth law, 131
CIBH diode, 268, 608, 609
CIPS 08 model, 524, 534
C-Lam, 676
CMOS, 670, 672, 679
Coefficient of thermal expansion, 440, 443
Coffin-Manson law, 522, 523
COMFET, 391
Common base configuration, 296
Common-base current gain, 296
Common-emitter configuration, 296
Common-emitter current gain, 296
Commutation, 234, 235, 238, 243, 255, 259, 262–265, 280, 284, 290, 327, 329, 370, 414, 462, 464, 483, 605
Commutation loop, 631
Compensation, 40, 41, 68, 185, 188, 232, 353, 368
Compensation structure, 353
Complementary error function, 160
Conduction band, 26
Conduction losses, 228, 365, 427
Conductivity modulation, 201
Conservation of charge, 88, 91, 93
Contact recombination, 141
Continuity equation, 88, 116, 122, 126
Control unit *See* Gate drive unit
Cooling, 8, 445ff
COOLMOS, 353
Corrosive atmosphere test, 574
Cosmic ray stability, 250, 541ff
Covalent bond, 24, 35
Critical dv/dt, 322
Critical field strength, 81, 136
Crosstalk, 673
CSTBT, 409
Current collapse, 383, 384, 673
Current equations, 103
Current filament, 558, 600, 602, 604, 605, 615, 626, 628
Current gain, 296, 302
Current sensor, 667, 682
Current snap-off, 412
Current source converter, 13
Current source inverter, 6
Current tube *See* current filament
Cut-off voltage, 372
Czochralski process, 149

D
D^2Pak, 435
Darlington transistor, 309
Data sheet, 203, 317, 442, 623
DBA, 515
DBC, 444, 469, 534, 681, 683
DBC substrate, 438
Dead time, 371
Debye length, 94
Deep impurities, 61
Degeneracy factor, 37, 41, 63
Delamination, 504
Dember field, 211, 215
Densities of states, 28
Depletion approach, 111
Depletion approximation, 104, 123, 144
Depletion capacitance, 145
Depletion type MOSFET, 343
Derating, 228
Desaturation, 617, 621
Diamond, 23, 47
Diamond lattice, 24, 25
Dielectric insulation, 672
Diffusion
  of carriers, 54
  of dopants, 165, 168
  of gold and platinum, 184
Diffusion capacitance, 147
Diffused junction, 134
Diffusion constant, 54
  of carriers, 54
  of dopants, 38, 165ff
Diffusion current, 54
Diffusion isolation, 413
Diffusion length, 117
Diffusion profile
  erfc-type, 160–164
  Gauss type, 297, 316
Diffusion voltage *See* Built-in voltage
Dimple array technique, 469
DirectFET, 435, 436
Direct semiconductor, 31

Displacement current, 91
Displacement field, 89
Divacancy, 80, 173, 188
DMOS, 343
Donor, 36, 37
Donor level, 37, 62, 64, 66, 69
  definition, 70
Double-pulse measurement, 235
Drift velocity, 46
Duty cycle, 456
Dynamic avalanche, 596, 612
Dynamic R_{on}, 383
Dynamic Self Damping Mode, 611

E
Early effect, 307
EasyPIM, 441
Edge diffusion structure, 181
Edge termination, 177, 413, 631
  beveled, 178, 591
  planar, 591, 595
Effective diffusion length, 254
Effective doping
  in base layers, 625
  in high doped emitters, 143, 226
Effective doping concentration, 141
Effective doping concentration in high doped emitters, 141
Effective emitter, 406
Effective impact ionization rate, 85, 129, 699
Effective mass, 33−36, 45, 49, 77
Effective lifetime, 219, 401
Effective n-emitter, 409
Efficiency, 428
Egawa-type field, 594, 603–607, 609
Einstein relation, 55, 93, 103, 211, 215
Electrochemical potential *See* Fermi energy
Electromagnetic compatibility, 639, 651
Electron affinity, 271, 272
EMCON diode, 254, 258, 265, 415
Emission probability, 63
Emission rate, 62, 67
Emitter efficiency, 137, 212, 218, 254, 263, 302, 305, 398, 400, 406, 614, 622
Emitter parameter *See* h-parameter
Emitter recombination, 139, 186, 255, 400, 588
Emitter short, 316, 320–323, 333
Energy gap *See* Bandgap
Enhancement type MOSFET, 343
Entropy factor, 692
Epitaxial growth, 155
Epitaxial layers, 155
Epitaxial wafer, 151
Epoxyd, 703, 705
Error function, 160, 162
Epitaxy, 201, 251, 671
Epoxy, 443
Equilibrium density, 57
ESD, 612
ESD protection, 322

F
Failure criteria, 490
Failure in time, 550
Failure limits, 575
Failure location, 631
Fast-Henry algorithm, 466, 468
FCE-diode, 268, 608
Fermi distribution, 27
Fermi energy, 68
Fermi level, 28, 39, 108, 272, 341
Fick's law
  first, 158
  second, 158
Field effect transistor *See* MOSFET, JFET
Field plate, 413
Field plate compensation, 357
Field rings *See* Potential rings
Fieldstop, 411
Filament *See* Current filament
Filters, 639
Float-Zone (FZ) process, 151
Fluctuations in doping, 152
Fluorine, 382
Forward recovery, 230, 619
Foster-model, 452
FREDFET, 368
Freewheeling diode, 185, 191, 201, 202, 230−232, 236

G
GaAs, 22, 29, 46, 269
Gallium (Ga)
  diffusion, 165
  ionization, 38
  solubility, 166
GaN, 23, 29, 47, 191ff, 378–385, 469, 673f, 703, 704
Gate drive unit, 336, 396, 623, 655, 667, 681
Gate injecton transistor, 380, 381, 384
Gate oxide, 175, 344
Gate resistor, 640, 641, 644
Gate stress test, 495
Gate unit, 653
GCT, 12, 335, 431
Generation
  Auger, 82
  impact, 88, 90

thermal, 57
Generation center, 188
Generation lifetime, 68, 124, 125
Generation rate, 88
Germanium, 22, 295
Gettering, 184
GIT *See* Gate injecton transistor
Gold, 79, 124, 183, 203, 232, 328, 333, 368
Graded junction, 111
Green Line module, 481
GTO thyristor, 12, 327, 330, 431, 597, 603, 615
Gummel-number, 254

H
4H-SiC, 23
H3TRB, 491, 495, 499, 502
Half-time, 154
Hall approximation, 216
Hall constant, 35
Hall effect, 27, 91
Harmonics, 4, 5, 637, 638
HASS test, 575
Heat capacity *See* Thermal capacity
Heat conduction, 90
Heat conductivity, 90
Heat flow equation, 90
Heat flux, 446
Heat flux density, 428
Heat generation, 46, 125
Heat generation rate, 90
Heat spreading, 448
HEXFET, 345
High doping effects, 39, 41, 42
High level carrier lifetime, 65, 73ff, 144, 212, 222
Holding current, 318
Hole barrier, 409, 416
Homologous temperature, 470
Hot reverse test, 182, 492
h-parameter, 139, 142, 218, 254, 303
Humidity, 499, 574
HVDC, 13, 14, 313, 322, 324, 430
Hybrid diode, 258

I
Ideality factor, 275, 448
IEGT, 404
Image force, 274
Impact ionization, 81, 126, 547, 594, 602, 657
Impact ionization rate, 82, 126
temperature dependency, 86, 135
IMPATT diode, 656
IMPATT oscillation, 655
Impedance, 653
Implantation *See* Ion implantation
Implantation profile, 171
Impurity scattering *See* Scattering
IMS substrate, 466
Incomplete ionization, 37, 39, 90
Indirect semiconductor, 31, 32
Injection efficiency *See* Emitter efficiency
Input rectifier, 441
Insulation materials, 445
Integration, 441, 667ff
Interface traps, 376
Intrinsic carrier density, 30, 583
Intrinsic concentration, 28, 37
Intrinsic level, 108
Intrinsic semiconductor, 28
Intrinsic temperature, 583
Inverse diode, 367, 386
Inversion channel, 344, 347, 492, 495
Inversion layer, 182, 341
Ion implantation, 170, 177
Ionization degree, 37, 41, 42, 354, 553, 558, 561
Ionization energy, 36, 41
Ionization integral *See* Breakdown condition
Ionization rate
temperature dependence, 86, 135
Ionization ratio, 40
IPM, 436, 667, 680, 685
Irradiation, 257
irr. enhanced diffusion, 188
with C ions, 549
with electrons, 186, 257, 368, 399, 659
with gamma quants, 186
with He ions, 80, 188, 257, 333, 399, 659
with neutrons, 154
with protons, 80, 154, 186, 188, 189, 333, 399
ISOPLUS package, 433

J
JFET, 372
Junction capacitance, 144
Junction capacity, 466, 643
Junction insulation, 670
Junction temperature, 185, 447
Junction termination *See* Edge termination
Junction voltage, 214

K
Kapton, 443
K-center, 81, 187, 656, 659

K-space, 31

L
Latching, 318, 393, 408, 613, 632, 671, 672
Lateral diffusion, 163
Lattice constant, 25, 192, 194
Lattice defect, 173, 257
defect profile, 173, 174, 257
Lead frame, 436
Leakage current, 116, 120, 123, 128, 182, 185, 188, 204, 299, 500, 623, 672
LESIT, 522
Lifetime of carriers, 57ff, 183ff
high level, 65, 73ff, 144, 212, 222
low level, 65, 66, 71f
Light Punch Through, 411
Light triggered thyristor, 322, 325
Linearly graded junction, 111, 112
Liquidus temperature, 471, 518
Lorentz force, 35, 91
LTJ. *See* Siver sinter technology

M
MagLam, 676
Majority carrier, 44, 46
Masking, 176
Mass law equation, 28
Matrix converter, 412
Matthiessen rule, 52, 374
Maxwell-Boltzmann distribution *See* Boltzmann distribution
Maxwell equations, 91
MCT, 404
Metallurgical junction, 105, 108, 110
Micro-controller, 682
Microwave devices, 647
Miller capacitance, 360
Miller indices, 25
Miller plateau, 361, 396
Miner's Rule, 530
Minority carrier, 44, 114
Minority carrier lifetime, 58
Minority current density, 139
MISFET, 381
Mission profile, 530, 531
Mobile ions, 182
Mobilities, 45ff
temperature dependence, 52, 689
MOCVD, 156, 192, 379
Molded package, 433, 436, 437, 481, 534
Molybdenum, 430, 438
Monocrystals, 149
Monolithic integration, 414, 669
MOS-Controlled Diode, 261, 406, 416
MOSFET, 12, 151, 171, 209, 259, 260, 262, 335, 336, 341, 434, 448, 464, 491, 584, 639, 670
Motor control unit, 682
MPS-diode, 252, 285
Multiplication factor, 83, 128, 299, 320, 697
Multi-vacancies, 257

N
N-buffer, 264
Nakagawa limit, 420, 421
Negative differential resistance, 592, 601
Neutrality condition, 37, 41, 65, 210
Neutron, 83, 151, 153, 154
Neutron flux, 154
Neutron transmutation doping, 153
Non-degeneracy, 28
NPT design, 205, 247, 549
NPT-IGBT, 399, 403, 411

O
Occupation degree, 37
Occupation probability, 27
Ohmic contact, 272
Ohmic region, 347
Oscillations, 633, 637
Over-current, 475, 586, 612
Over-temperature, 585
Overvoltage protection, 326
Oxide growth, 164
Oxygen content, 149, 151

P
Palladium, 189
Parallel connection, 228, 239, 401, 640, 644
Parasitic capacitance, 360, 433, 467, 684
Parasitic diode, 366
Parasitic inductance, 231, 236, 242, 250, 259, 435, 462, 469, 619, 643
Parasitic resistance, 434, 458, 472
Parasitic thyristor, 392, 671, 672
Parasitic transistor, 366, 670
Passivation, 175, 182, 469, 493, 499, 670
Passive components, 4, 10, 669, 680, 685
PCB, 336, 429, 434, 435, 575, 641, 676, 681, 684
PETT oscillations, 648
Phonon, 32, 47, 60, 85
Phosphorous
diffusion, 188
doping level, 44
ionization, 38
implantation, 190
solubility, 188
Phosphorous isotope, 153

Photolithography, 176
Pinch-off, 347
Pinch-off region, 349
Pinhole, 545, 632
Pion, 543
Plasma enhancement, 407
Platinum, 79, 125, 184, 188, 203, 227, 368, 399, 415
Plugged cells, 410
Pn-insulation, 672
Poisson equation, 89, 104, 105, 130, 179, 307
Polyimide, 443, 516, 678
Poly-silicon, 344, 368
Polytype, 25
Potential rings, 413, 591
Power converter, 684
Power cycling test, 505
Power electronic building blocks, 8
Power electronic system, 2, 667ff
Presspack-IGBT, 432
Projected range, 171, 189
Projected range straggling, 171
Proton flux, 543
PT design, 204, 258, 261, 277, 337, 352, 549
PT-IGBT, 398, 411
Pulse Width Modulation, 668, 679, 682
Punch through, 318

Q
Quantum theory, 27, 35
Quasi-neutral region, 113
Quasi-particle model, 34
Quasi-saturation, 297, 346

R
Radiation *See* irradiation
Radiation induced centers, 80, 187, 188
Radiative band-to-band recombination, 58
Radiative recombination, 32, 59
Rainflow method, 531
Ramo-Shockley-theorem, 650, 658
Rapid thermal annealing, 175
Recombination
  at deep impurity, 62ff
  Auger *see* there
  radiative, 58
  thermal, 57
Recombination centers, 68, 69, 185–189, 256, 257
Recombination radiation, 32, 222
Recombination rate, 57, 70, 88
Reconstruction, 514, 625
Recovery time
  diode, 233ff
  thyristor, 326–328
Rectifier diodes, 201
Relaxation time, 93
Resistivity, 21, 30, 40, 46, 47, 153, 332, 516
Reverse conducting IGBT, 414, 685
Reverse-recovery behavior, 232ff, 642
Reverse-recovery charge, 186, 240, 260
RF power, 650, 658
Richardson constant, 273
Ring emitter structure, 308
Robustness validation, 575
Ruggedness, 609, 623, 628, 632

S
Safe operating area, 308, 365–367, 612
Sapphire, 156, 191, 379
Saturation, 297, 304
Saturation current, 119, 120, 121, 123, 124, 127, 139, 273
Saturation drift velocity, 249, 597, 651
Saturation region, 394
Saturation velocity, 53, 307
  temperature dependency, 52
Scanning Acoustic Microscopy, 504, 518
Scattering
  electron hole, 50, 128
  impurity, 35, 37
  phonons (lattice), 47
  quantum mechanical theory, 51
Schottky barrier, 272, 284, 371
Scharfetter relation, 690
Schottky diode, 22, 23, 209, 252, 271ff, 371
Schottky junction, 271ff
SEB, 541, 542, 545, 547, 559, 561
Second breakdown, 307, 366, 552, 559
SEGR, 545
Selenium, 22
Semitop, 441
Series connection, 260
Shallow thermal donors, 189
Shockley equation, 118, 121
Shockley-Read-Hall equation, 62, 64, 124
Short circuit, 616ff, 631
Short circuit capability, 299, 616
Shunt resistors, 441
Si_3N_4, 177, 196, 445, 479, 481, 482
SiC, 10, 11, 23, 25, 30, 38, 47, 86, 110, 136, 155, 269, 281, 309, 374, 469
SIDAC, 322
Silicon isotope, 153
Silver sinter technology, 470, 516
Simulation, 179, 255, 445, 455, 586, 589, 602, 606, 628, 678
Single crystal, 22

Single Event Burnout, 559
Single phase inverter, 618
SKiiP, 441, 682
SKiN technology, 482, 483
Smart power, 343, 673
Smart power ICs, 673
SMD technology, 432, 435
Snap-off, 234, 642
Snubber, 13, 334, 615
Soft factor, 234
Soft Punch Through, 411
Soft recovery diode, 233ff, 644
Soft-switching, 11, 403
SOI technology, 672, 680
Solder, 438, 439, 471, 504
Solder fatigue, 510, 518, 520
Solder voids, 471
Solubility of dopants, 160, 165
Solubility of gold, 183
SPEED diode, 256
Spring contacts, 441
Standard module, 439, 509, 511, 523
Stored charge, 219, 228, 251, 264, 334, 396
Stray inductance *See* parasitic inductance
Streamer, 547
Striations, 153
Strip line concept, 484
Superjunction, 353, 386
Surface charge, 343, 492, 494, 495
Surface damage, 151
Surge current, 286, 584, 586
Switching losses, 10, 334, 365, 427, 639
Switching self-clamping mode, 607, 613
Switching time *see* turn-on, turn-off
Synchronous rectifier, 262, 370
System integration *see* integration

T
Tail current, 237, 244, 249, 257, 397, 401
Tandem diode, 260
Temperature coefficient forward voltage, 411, 589
Temperature compensation point, 358
Temperature cycling test, 503
Temperature limit, 30
Temperature ripple, 418, 458
Temperature sensor, 441, 667
Temperature shock test, 503
Temperature swing, 456, 506, 515, 531
Thermal annealing, 174
Thermal capacity, 624, 625
Thermal conductivity, 90, 443, 447, 586
Thermal double-donors, 188
Thermal expansion, 433, 504, 509, 589, 680
Thermal generation, 584
Thermal generation rate, 122
Thermal grease, 442
Thermal impedance, 442, 453, 455
Thermal interface material *See* thermal grease
Thermal network, 445, 452, 453
Thermal resistance, 185, 365, 442, 446, 509
Thermal runaway, 585
Thermal stress, 509, 622
Thermal velocity, 45
Thermal voltage, 104
Thermo-mechanical stress, 524
Thin wafer technology, 202
Three-phase converter, 667
Three-phase inverter, 436, 682
Threshold voltage
  pn-junction, 120
  MOSFET, IGBT, 6, 343, 346, 351, 358, 394
  Schottky junction, 274, 288
Thyristor, 313ff
  light triggered, 2, 13, 85, 181, 313, 392, 429, 431, 438, 586
TO package, 429, 433, 468, 535
TOPS diode, 253
Transconductance, 346
Transfer mold, 433
Transfer mold package, 436
Transit frequency, 653
Transit time, 647
Transit-time oscillations, 647
Transport equations *See* current equations
Transport factor, 302, 305, 399
Trapped charges, 415
Trench, 345, 356, 406, 672
Trench IGBT, 406
Trench insulation, 672
Trench MOSFET, 345, 368
Triac, 329
Trigger condition, 322, 330
Trigger front spreading, 323
Trimetal, 438
Triple-diffused transistor, 297
Tunneling, 81
Turn-off
  IGBT, 396
  MOSFET, 361
Turn-off gain, 330
Turn-off oscillations, 640
Turn-on
  diode, 230
  GCT, 6, 12, 330, 335, 337–339, 431
  IGBT, 395
  MOSFET, 360
  thyristor, 230, 325, 326, 328, 672
  triac, 329, 330

U
Unipolar device, 271
Unipolar limit
  Si, 279, 351, 357
  SiC, 284

V
Valence band, 26, 28
Van Allen belt, 543
Vibration, 575, 678
Virtual junction temperature, 447–449

W
Wafer bonding, 672
Wave function, 40
Weibull-distribution, 563
Weibull statistics, 534
Wire bond *See* bond wire
Work function, 271, 272
Wurtzite lattice, 24, 25

Z
Zincblende lattice, 25