Transforming Edge AI: The power of neural processing units in modern microcontrollers

Introduction

With the growing ubiquity of artificial intelligence (AI) and the exponential increase in data collection, processing all data remotely in the cloud has become unsustainable and impractical. The widespread adoption of AI across various products and applications requires more efficient and localized processing solutions, namely edge AI. By leveraging advanced microcontrollers with embedded neural accelerators, solution developers can harness the power of AI directly on edge devices. This approach offers significant benefits, including reduced power consumption, decreased network load, and lower latency, enabling faster and more responsive AI-driven applications.

AI EVERYWHERE

Artificial intelligence is revolutionizing the way we interact with technology, moving beyond traditional rule-based algorithms and their hard-coded instructions. With AI, machines can learn from data, adapt to new inputs, and handle uncertainty, and so display more adaptable and human-like behavior.

AI skills like recognizing patterns and detecting exceptions, making predictions, and generalizing to handle new, unseen data are being adopted in diverse applications. Examples include smartphone photography, activity detection in fitness bands, people counting for purposes like smart retail, traffic cameras and crowd management in smart cities, security systems, and predictive equipment maintenance.

AI is enabling a wide variety of commonly used technologies to become smarter, more efficient, and more responsive, including:

- **Personal electronics:** Smartphones, smartwatches, and home assistants are becoming more intelligent, providing personalized experiences and predictive functionalities.
- **Automotive:** AI is crucial for the development of autonomous vehicles, advanced driver-assistance systems (ADAS), and smart traffic management.
- **Industrial applications:** AI is enhancing predictive maintenance, optimizing supply chains, and improving operational efficiencies in manufacturing and logistics.
- **Healthcare:** AI is revolutionizing diagnostics, personalized medicine, and patient monitoring.
- **Retail:** AI-driven analytics and recommendation systems are transforming the shopping experience.

CLOSER TO THE ACTION

Training AI algorithms using large datasets is traditionally performed in the cloud, leveraging the power of AI servers that can handle the massive parallelism and multi-layer dataflows of neural networks. Tradition has also hosted the trained algorithms – or AI models – on cloud servers.

Now, with the rapidly growing demand for AI systems in every kind of high-tech device, hosting AI models directly on those devices instead of in the cloud can deliver several advantages. The time for the system to respond to events – reading a vehicle number plate, detecting, and counting people in a video frame – can be significantly reduced. The volume of data shared through the cloud connection can also be reduced, relieving demand for network bandwidth. Autonomous devices can continue to operate in the event of network outages. In addition, minimizing the quantity of data shared over the network can protect data privacy. Finally, and perhaps most importantly from a sustainability standpoint, migrating AI workloads into efficient low-power devices can dramatically reduce the energy currently consumed in AI data centers.

This on-device inference, more generally known as edge AI, is transforming design approaches to embedded systems; in particular, the processing subsystem that hosts the application.

According to ABI Research, the market projection for Edge AI is poised for significant growth over the next decade. The data indicates a **substantial increase in the deployment of microcontroller units (MCUs) for Edge AI applications** across various verticals, including agriculture, automotive, cellular networks, healthcare, manufacturing, personal and work devices, retail, and robotics. The market is expected to experience exponential growth, reaching nearly **1.8 billion units by 2030**.

Key challenges product developers must tackle to enhance their designs with AI:

- **Energy demand/consumption:** AI and machine learning (ML) tasks are computationally intensive, leading to significant power consumption. This is particularly challenging in power-constrained environments where energy efficiency is crucial.
- **Performance bottlenecks:** General-purpose MCUs often struggle to meet the computational demands of neural networks (NN). This results in performance bottlenecks that can hinder the effectiveness of AI applications.
- **Latency and real-time processing:** For many AI applications, particularly those requiring real-time responses, latency is a critical concern. High latency can impair the performance and user experience of AI-driven systems.
- **Complexity and cost:** Embedding AI algorithms in microcontroller units (MCUs) presents several challenges due to the limited memory and processing capabilities of these devices. Additionally, the software tools and AI frameworks available are often not originally designed with embedded developers in mind, complicating the development process.

What is a neural processing unit (NPU) and why it is so important?

Typical embedded processor cores designed for sequential computing using **instruction fetch – decode – execute** are not designed to run AI models as efficiently as possible. This is because neural network computation topologies often involve a significant amount of memory access, as well as accumulation and multiplication operations, which are not optimized in traditional sequential architectures. A different architecture is needed that can perform fast and efficient AI inference within the typical embedded constraints on power consumption and silicon area.

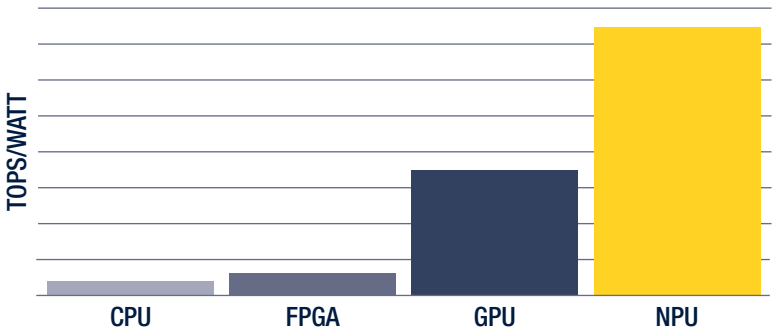
The neural processing unit (NPU) has emerged to address this requirement.

Comparing the NPU with central processing units (CPUs) and graphics processing units (GPUs) helps to understand its distinct advantages for AI and ML applications. The table below highlights the key features and differences:

Feature	CPU	GPU	NPU
Primary function	General-purpose processing	Graphics rendering, parallel computation	Neural network acceleration
Architecture	Few powerful cores, high clock speed	Many smaller cores, SIMD architecture	Specialized cores for neural networks
Processing type	Sequential processing	Parallel processing	Parallel and low-latency processing
Instruction set	Complex instruction set (e.g., x86, Arm)	Graphics and parallel computation instructions	Optimized for convolutional neural network operations
Energy efficiency	Moderate	Moderate to high	High
Use cases	General computing, control tasks	Graphics rendering, scientific simulations, ML training	Edge AI, real-time inference, IoT devices

Thanks to their specialized architecture, NPUs represent a significant advancement in Edge AI acceleration, offering superior efficiency compared to traditional processing units.

Figure 1: Power efficiency of various hardware architectures



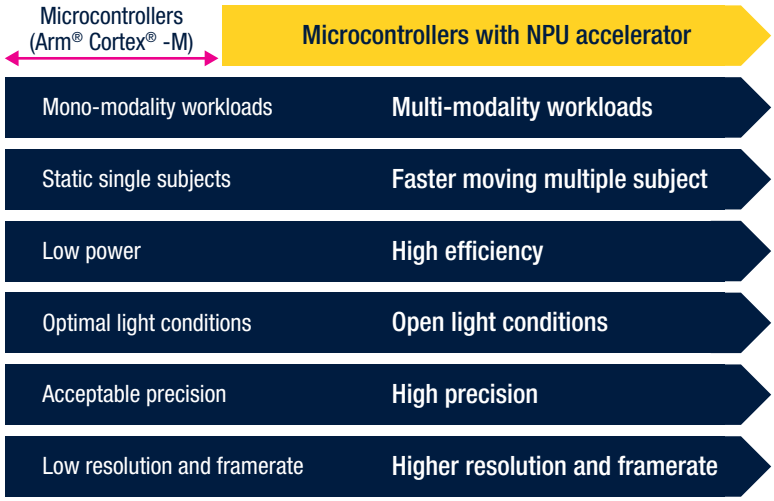
GPU: Graphic accelerator
NPU: Neural processing unit (AI accelerator)

Source W. Dally

OPENING A NEW RANGE OF EMBEDDED AI APPLICATIONS

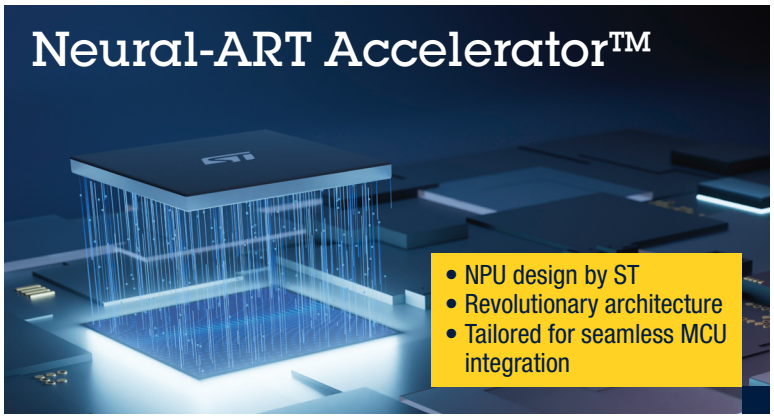
NPUs are highly efficient, making them particularly suitable for energy-constrained environments like microcontroller-based applications. They provide an optimal solution for addressing a wide array of edge AI use cases while maintaining low power consumption.

Figure 2: Opening a new range of embedded AI possibilities



The integration of NPUs with microcontrollers significantly extends the capabilities of MCUs, enabling them to handle more complex AI tasks that were previously beyond their reach. Traditionally, MCUs have been limited to simpler AI applications such as low-resolution picture analysis, time series analysis or low frame rate, due to their constrained processing power and energy efficiency. However, with the addition of NPUs, these microcontrollers can now perform advanced AI functions such as speech recognition, object classification, pose estimation, and object segmentation localization on faster moving and smaller objects. By offloading AI inference tasks to the NPU, the MCU can focus on other critical functions, ensuring efficient and real-time processing.

Introducing ST Neural-ART Accelerator



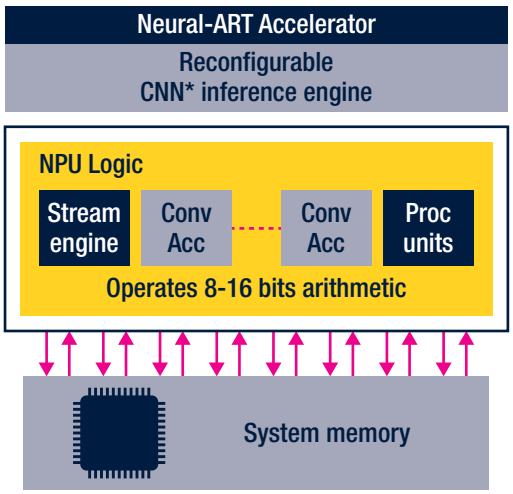
STMicroelectronics (ST) envisions a future where AI is embedded directly into devices, making Edge AI a reality. By enabling AI capabilities on the edge, ST aims to reduce the reliance on cloud computing, thereby conserving energy and reducing latency.

To bring this vision to life, ST announces the first member of its **Neural-ART Accelerator™** series: a purpose-designed NPU with highly parallelized hardware co-processor that's available for the main embedded application to assign AI workloads.

Integrated into STM32 microcontrollers, the groundbreaking Neural-ART Accelerator empowers them to efficiently handle AI inference tasks on edge devices. This integration represents a major advancement in making Edge AI both practical and widespread, in line with ST's mission to deliver smarter and more energy-efficient solutions. Leveraging ST's extensive expertise in microcontroller technology and AI, this new generation of NPUs offers a powerful, efficient, and scalable solution for a diverse range of applications.

Equally important is the accompanying development toolchain that enables developers to evaluate, optimize, and deploy their models to utilize the accelerator's architecture efficiently. **ST Edge AI Suite** provides the tools for developers to build AI applications for STM32 microcontrollers (MCUs), leveraging popular AI frameworks such as Keras, TensorFlow, and ONNX.

Figure 3: Neural-ART Accelerator architecture overview



NEURAL-ART ACCELERATOR ARCHITECTURE OVERVIEW

The Neural-ART Accelerator integrates multiple specialized hardware accelerators capable of supporting various inference kernels. These accelerators are dynamically connected through a reconfigurable dataflow stream processing engine, ensuring flexible and efficient processing. The architecture includes a configurable number of convolutional accelerators with fixed-point MACs configurable to either 16 or 8 bits of precision.

- Advanced timer for 3-phase inverters and full-bridge converter drivers
- Fast and precise ADC can be triggered by timer events
- 5 V power supply
- Input capture on general-purpose timers for easier speed feedback processing
- Encorder operating mode only for DC motors

INTEGRATION WITH SOFTWARE TOOLS

The efficiency of the Neural-ART Accelerator IP is ensured by its seamless integration with the STM32Cube software ecosystem. This integration is natively supported by **STM32Cube.AI** (desktop app) and **ST Edge AI Developer Cloud** (online), which generates optimized code to fully leverage the capabilities of the NPU without hassle.

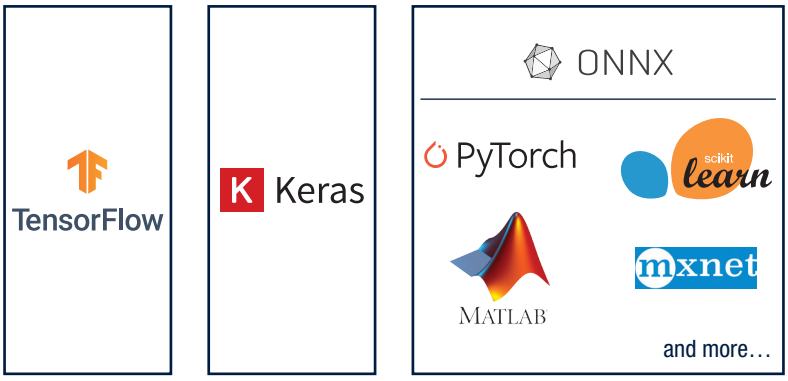
By utilizing STM32Cube.AI or ST Edge AI Developer Cloud, developers can easily optimize and convert pre-trained neural network models into code that runs efficiently on the Neural-ART Accelerator. Both tools analyze the neural network, map its operators to the appropriate hardware resources, and ensures that each layer is accelerated in the most optimal way. This comprehensive toolchain simplifies the deployment of AI models on STM32 microcontrollers, allowing developers to focus on innovation rather than the complexities of hardware acceleration.

In summary, the Neural-ART Accelerator offers a powerful and energy-efficient solution for embedding advanced AI capabilities into MCU-based applications. Its integration with the **STM32Cube software ecosystem** ensures a smooth and efficient development process, enhancing, leading to faster time to market.

Figure 4: Edge AI software ecosystem

When developing a neural network (NN) model, AI operators play a crucial role in defining the various mathematical and logical functions that the model performs. Each AI framework, such as TensorFlow, Keras, or ONNX, offers a set of operators that developers can use to build and train their models. Supporting these operators in software is essential for enabling AI applications, while hardware support is critical for achieving maximum performance. The number of operators available in each framework can significantly impact the development process, as a wider coverage of operators drastically decreases the time required to develop an edge AI application.

The combination of advanced hardware and robust software tools enables support for up to **130+ operators from popular AI frameworks, including ONNX**, ensuring comprehensive functionality and flexibility.



PERFORMANCE BENCHMARKS

The data shown in the table below illustrates the inference acceleration achieved with an STM32 leveraging the Neural-ART Accelerator and Cortex-M55 (for sequencing only), compared to Cortex-M55 alone, when handling popular embedded machine learning models. We would point out that the Cortex-M55 is also optimized for AI inference thanks to Arm Helium technology, which provides M-profile vector extension delivering a significant performance uplift for machine learning (ML) and digital signal processing (DSP) applications.

Table 1: Results measured on Neural-ART Accelerator Gen1, with 4 convolution array at 1 GHz

	CPU-bound inferencing on Cortex-M55 @ 400 MHz		Accelerated inferencing on ST Neural-ART Accelerator @ 1 GHz		Improvement with acceleration
Model	**Time (ms)**	**fps**	**Time (ms)**	**fps**	
MobileNet v1[1]	2244	1.78	19	53.2	x120
MobileNet v2[2]	1386	2.88	21	48.62	x64
Tiny Yolo v2[3]	3894	1.02	29	33.96	x134
Yolo v8n 256[4]	1820	2.2	27	37.42	x68
Yamnet 1024[5]	252	15.88	10	102.35	X26

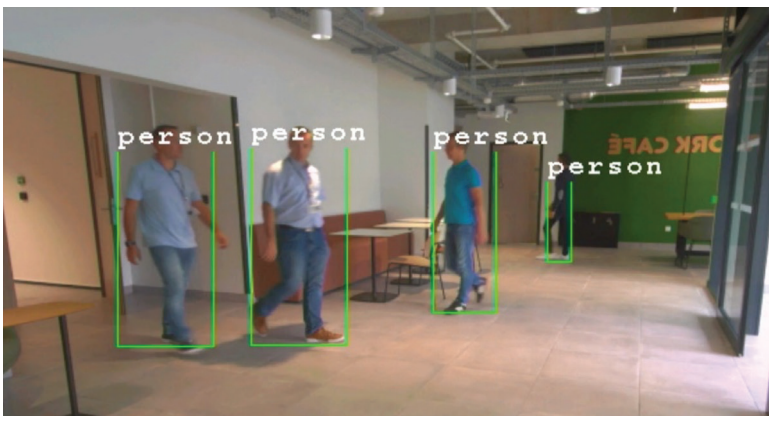
Note: 1 **Image Classification** - Quantized int8, input resolution 224x224x3, trained on ImageNet dataset. Model footprint: 4.45 MB weights, 1.53 MB activations
2 **Image Classification** - Quantized int8, input resolution 224x224x3, trained on ImageNet dataset. Model footprint: 4.14 MB weights, 2.08 MB activations
3 **Object Detection** - Quantized int8, input resolution 224x224x3, trained on COCO dataset. Model footprint: 10.55 MB weights, 0.38 MB activations
4 **Object Detection** - Quantized int8, input resolution 256x256x3, trained on COCO dataset. Model footprint: 3.05 MB weights, 1.6 MB activations
5 **Audio Event Classification** - Quantized int8, input resolution 64x96, trained on AudioSet dataset. Model footprint: 3.4 MB weights, 0.14 MB activations

The performance of ST's Neural-ART Accelerator in computer vision and audio applications clearly illustrates the advantages embedded NPUs now offer for developers. MobileNet, Yolo, and Yamnet are popular AI model topologies used for computer vision, object detection, and audio event classification. **STM32Cube.AI** was used to compile these models on the Arm Cortex-M CPU core contained in 32-bit MCUs. The same models were then compiled for the Neural-ART Accelerator and the corresponding figures for execution time and supported frame rate were recorded. The table compares the performance and shows the improvement gained.

Subsequently, two use cases were evaluated to show how the performance improvement achieved with the Neural-ART Accelerator effectively permits real-time object detection and feature recognition in machine-vision applications that are widely used in smart buildings, smart retail, and smart city contexts.

In the first example, Yolo (You Only Look Once) v8 is the object detection model chosen for a people detection and tracking application. The comparison table showed that the Neural-ART Accelerator improves Yolo performance more than 100-fold. In this example, the system achieves a frame rate of 26 fps. This shows that Yolo v8 inference can analyze 100% of video frames continuously, in real-time. This is impossible when running Yolo v8 on the Cortex-M55.

Case Study 1: People detection for consumer tracking applications

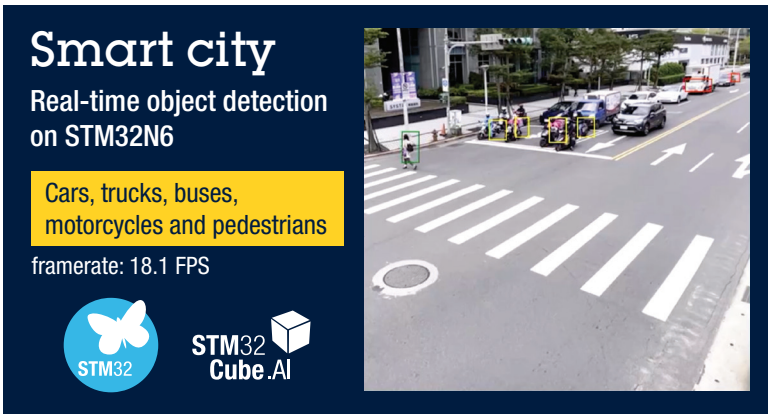


People detection and tracking
Yolo v8 320x320
Neural-ART with 4 Convolution Array at 1 GHz = **26 fps**

Screen capture of the actual demo board screen.

In the second example, a smart-city camera uses the lightweight TinyYolo v2 to analyze a street scene at a higher resolution than the first example. In this case, image analysis can be used for vehicle type classification and pedestrian detection. The frame rate of 18 frames per second lets the application capture enough information to monitor and improve road safety and traffic flow.

Case Study 2: Smart city and transportation applications



Multi-class object detection, tracking, and counting
TinyYoloV2 416x416
Neural-ART with 4 Convolution Array at 1 GHz = **18 fps**

Screen capture of the actual demo board screen.

The input resolution in these examples is important to note. Increasing the resolution by changing the camera sensor and optical system increases the information delivered to the model and hence extends the execution time for any given model. This effectively lowers the frame rate that can be achieved. Conversely, lowering the resolution can enable a higher frame rate. In the smart-city example of Case Study 2, the camera resolution could be increased, for example, to recognize vehicle number plates at greater distances from the camera, without excessively slowing performance. On the other hand, a wider viewing angle could be used thereby allowing fewer cameras to cover a given area and so reducing the installed cost of the overall system.

Developing and deploying Edge AI solutions on an STM32 MCU equipped with a Neural-ART Accelerator

STMicroelectronics provides a comprehensive suite of tools and resources to support the development and deployment of AI models on STM32 microcontrollers equipped with the ST Neural-ART Accelerator. These tools are designed to streamline the development process and ensure that developers can achieve optimal performance and efficiency.

STM32Cube.AI is a key tool in the development workflow, enabling developers to optimize and convert pre-trained neural network models into code for STM32 microcontrollers and Neural-ART Accelerators. The tool supports a wide range of AI frameworks and provides detailed documentation and tutorials to help developers get started quickly.

The ST Edge AI Model Zoo is a collection of pre-trained AI models optimized for STM32 microcontrollers. These models cover a variety of applications, including image classification, object detection, and speech recognition. The ST Edge AI Model Zoo provides developers with ready-to-use models that can be quickly deployed on STM32 microcontrollers, significantly reducing development time and effort.

To help developers get started quickly, STMicroelectronics provides reference designs and sample code that demonstrate how to implement AI solutions using the ST Neural-ART Accelerator. These resources include detailed schematics, code examples, and application notes, offering valuable insights and best practices for AI development. Additionally, the **ST Edge AI Developer Cloud** allows developers to easily benchmark neural networks performance and footprint on any STM32.

THIRD-PARTY SUPPORT AND ECOSYSTEM

The ST Neural-ART Accelerator benefits from a **vibrant ecosystem of third-party partners and developers** who contribute to the development and advancement of AI solutions. This ecosystem includes software vendors, research institutions, and industry experts who provide additional tools, libraries, and support for AI development.

Developers can leverage this ecosystem to access a wide range of resources, including specialized hardware components, advanced AI algorithms, and expert guidance. By collaborating with third-party partners, developers can accelerate their development process and ensure that their AI solutions meet the highest standards of performance and reliability.

DEVELOPMENT WORKFLOW

The development workflow for the ST Neural-ART Accelerator enables developers to quickly implement AI solutions on STM32 microcontrollers. The process involves several key steps:

- **Model selection and training:** Developers begin by selecting or designing an appropriate AI model for their specific application. This model is then trained using a suitable dataset, typically in a high-level framework such as TensorFlow, Keras, or any model in ONNX format.
- **Model optimization:** Once the model is trained, it can be optimized through various techniques such as quantization, pruning, and compression using solutions available within our well-integrated partner ecosystem.
- **Conversion with STM32Cube.AI:** STM32Cube.AI enables developers to convert pre-trained neural network models into optimized C code. This tool supports model import from popular AI frameworks, applies optimization techniques, and integrates with the NPU compiler to maximize performance and efficiency.
- **Integration and testing:** The generated code is integrated into the application code running on the STM32 microcontroller. Developers can use the **STM32CubeIDE** or other compatible development environments to compile, upload, and test the application on the target hardware.
- **Deployment:** Once the application is thoroughly tested and validated, it can be deployed to the edge device. The ST Neural-ART Accelerator ensures that the AI model runs efficiently, delivering real-time performance and low power consumption.

CONCLUSION

The rapid adoption of AI across various sectors underscores the critical need to address the challenges of energy demand, performance bottlenecks, latency, and complexity in embedding AI algorithms into microcontroller units (MCUs). As highlighted in the introduction, the energy consumption required to handle AI inference in the cloud is unsustainable, necessitating a shift towards Edge AI to ensure efficient and real-time data processing.

The Neural-ART Accelerator represents a significant advancement in addressing these challenges. By integrating Neural Processing Units (NPUs) into STM32 microcontrollers, the Neural-ART Accelerator enables efficient handling of AI inference tasks directly on edge devices. This integration not only reduces the reliance on cloud computing, thereby conserving energy and reducing latency, but also extends the capabilities of MCUs to perform complex AI functions such as anomaly detection, speech recognition, and object classification.

The architecture of the Neural-ART Accelerator, with its specialized hardware accelerators and reconfigurable dataflow stream processing engines, ensures high performance and flexibility. The seamless integration with the **STM32Cube software ecosystem** further enhances the development process, allowing developers to easily convert pre-trained neural network models into optimized code that runs efficiently on the NPU. This combination of advanced hardware and robust software tools results in dramatic performance improvements, making the Neural-ART Accelerator an ideal choice for a wide range of applications.

Performance benchmarks demonstrate that the Neural-ART Accelerator delivers superior performance and efficiency compared to traditional AI processing units. The NPU can handle complex AI models with low latency and high accuracy, while consuming significantly less power, making it ideal for battery-powered edge devices. Case studies in AIoT consumer products and smart city applications further illustrate the real-world benefits of the Neural-ART Accelerator.

Looking ahead, STMicroelectronics is committed to continuous innovation and improvement of the Neural-ART Accelerator. The future roadmap includes exciting features and enhancements aimed at further boosting the performance, efficiency, and versatility of the NPU family. The transition from digital NPUs to in-memory computing promises significant improvements in energy efficiency and computational performance, paving the way for the next generation of neural network inference engines.

DEVELOPER RESOURCES

STMicroelectronics is dedicated to providing the support and resources needed to help developers and businesses succeed with the ST Neural-ART Accelerator.

Artificial intelligence at the edge

See how ST is moving intelligence from the cloud to the edge **[ST technology page]**

ST Edge AI Suite

A comprehensive set of tools for integrating AI features in embedded systems **[ST Developer Zone] [Developer tools] [Case studies] [ST Community]**

ST Edge AI Model Zoo

A collection of reference edge AI models to add edge AI capabilities to embedded applications and optimized to run on ST devices with associated deployment scripts
[STM32 model zoo] [MLC model zoo on GitHub] [ISPU model zoo on GitHub]

ST Edge AI Developer Cloud

A free online platform to easily optimize and benchmark edge AI models across a variety of ST devices **[Product overview]**

STM32Cube.AI

A free STM32Cube expansion package that lets you convert pretrained edge AI algorithms automatically, such as neural network and machine learning models, into optimized C code for STM32 [Product overview]

ST Edge AI Core

A command-line interface (CLI) tool to optimize and compile edge AI models for multiple ST devices, including microcontrollers, microprocessors, and MEMS sensors **[Product overview]**

High-speed Datalog

A comprehensive multi-sensor data capture and visualization toolkit that lets you manage the acquisition and labelling of sensor datasets using the graphical user interface (GUI), the command line interface (CLI), or Bluetooth with a smartphone **[Product overview]**

MIX
Paper from responsible sources
FSC® C003379

Order code: **WP2411NEURART**

For more information on ST products and solutions, visit www.st.com