NEWS FOR THE ELECTRONICS INDUSTRY

eTECH JOURNAL

ISSUE 9

Farnell
AN AVNET COMPANY

INCREASE EMBEDDED AI PROCESSING

+

GMSL AN ALTERNATIVE TO GIGE VISION CAMERAS

EDGE AI AND COMPUTER VISION

DEPLOYING MACHINE LEARNING TO THE EDGE

STM32MP2 MPU SERIES 64-BIT MICROPROCESSORS

INCREASE EMBEDDED AI PROCESSING

FUTURE **EMERGING** TECHNOLOGIES

CONTENTS

To register for all future editions and to access all back issues of eTech Journal, scan QR code

Editor-in-chief: Cliff Ortmeyer, Managing Editor: Ankur Tomar

WELCOME

Welcome to the latest edition of e-TechJournal, which explores the rapidly evolving landscape of emerging technologies driven by advancements in vision systems, microprocessors, and machine learning. These breakthroughs are sparking innovation across numerous sectors. This edition offers a comprehensive analysis of these state-of-the-art developments, highlighting key technologies that will shape the future of technology.

A central theme is the deployment of machine learning to the edge, with detailed analysis offering insights into practical real-world implementations. The spotlight is on the STM32MP2 MPU series, 64-bit microprocessors equipped with a Neural Processing Unit, pushing the boundaries of on-device AI computation.

Additionally, it compares Gigabit Multimedia Serial Link (GMSL) cameras to GigE Vision cameras, highlighting the advantages of GMSL in terms of data transmission speed and performance. Complementing this, the Arduino Nicla Vision emerges as a compact powerhouse, streamlining computer vision applications with its versatile capabilities.

Lastly, the edition explores strategies to enhance embedded AI processing for autonomous systems, addressing the critical need for seamless and intelligent operations in complex environments. It examines the role of high-efficiency power management in modern electronic devices and the transformative potential of 5G technology in industrial automation.

Join us in navigating these innovations that are driving technological advancement. This collection aims to inspire further innovation and welcomes your feedback.

Cliff Ortmeyer Editor, eTech Journal
Email: editor-TJ@element14.com

GIGABIT MULTIMEDIA SERIAL LINK (GMSL) CAMERAS AS AN ALTERNATIVE TO GIGE VISION CAMERAS

Author: Kainan Wang,
System Applications Engineer

Gigabit Multimedia Serial Link™ (GMSL™) and Gigabit Ethernet (GigE) are two popular link technologies for camera applications that are often seen in different end markets.

This article conducts a comparative analysis of the two technologies in system architectures, key features, and limitations. It will help explain the fundamentals of both technologies and provide insights into why GMSL cameras are a strong alternative to GigE Vision® cameras.

GigE Vision is a network camera interface standard based on Ethernet infrastructures and protocols. It is widely adopted in the industrial space. Analog Devices' GMSL is a point-to-point serial link technology dedicated to video data transmission and was originally designed for automotive camera and display applications.

Both technologies serve the purpose of extending the reach of video data from the image sensors, while each solution has its own unique features. Over the years, we have seen more GMSL cameras being adopted outside of the automotive space, often as an alternative to GigE Vision cameras.

TYPICAL SYSTEM ARCHITECTURE

GigE Vision cameras (shown in Figure 1) usually consist of three major components in their signal chain—an image sensor, a processor, and an Ethernet PHY.

The processor converts the raw image data from the image sensor into Ethernet frames, and the process usually involves image processing and compression or frame buffering to fit the data rate into Ethernet supported bandwidth.

GMSL cameras' signal chain (shown in Figure 2) is typically more straightforward with only an image sensor and a serializer. In typical applications, the serializer converts the raw data from the image sensor and then sends it over the link in its original format. Without the need for a processor, these cameras are simpler to design and more suitable for applications requiring a small camera form factor and low power consumption.

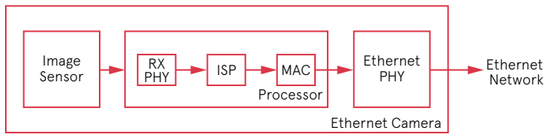


Figure 1. Key signal chain component on the sensor side for GigE Vision cameras.

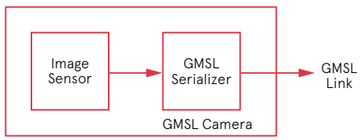


Figure 2. Key signal chain component on the sensor side for GMSL cameras.

Host Processor Connection

GigE Vision cameras are well accepted in the industry for their compatibility with a wide variety of host devices. Gigabit Ethernet port is almost a standard offering on personal computers (PCs) or embedded platforms.

Some of the GigE Vision cameras can work with wa universal driver for a true plug-and-play experience.

GMSL cameras require deserializer(s) on the host side. In most use cases, the host device is a customized embedded platform with one or multiple deserializers. The deserializers will transmit image data through its MIPI transmitter(s) in the original format from the image sensor MIPI output.

For these cameras, a camera driver is required for each customized camera design, just as any other MIPI camera. However, if there is an existing driver for the image sensor, only a few profile registers or a few register writes are required for the SerDes pair to get a video stream from the cameras to the SoC.

When using only one camera, GigE Vision may have some Aadvantages over GMSL in terms of system complexity since it can be connected directly to a PC or an embedded platform with an Ethernet port.

However, when multiple GigE cameras are used, an Ethernet switch is required. This can be a dedicated Ethernet switch device, a network interface card (NIC) with multiple Ethernet ports, or an Ethernet switch IC in between multiple Ethernet ports and the SoC.

In some cases, this will result in a reduced maximum total data rate and, worse, unpredictable latency depending on the interface between the cameras and terminal device. See Figure 3.

In a GMSL camera system, one deserializer can connect to up to four links with its MIPI C-PHY or D-PHY transmitter to support the full bandwidth of all four cameras. As long as the SoC can handle the total data rate, using one or multiple GMSL devices would not compromise bandwidth or increase too much system complexity.

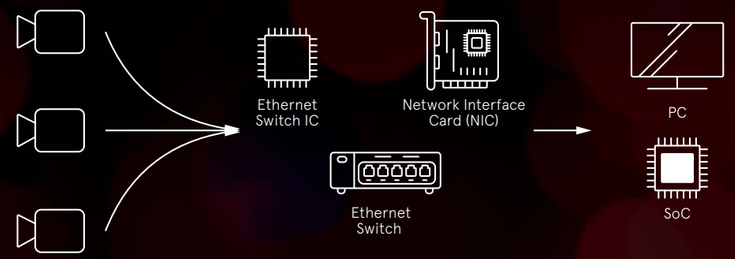


Figure 3. A typical GigE Vision network.

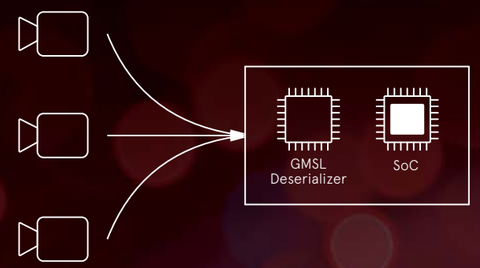


Figure 4. Typical GMSL cameras to host connection.

FEATURE COMPARISON

Sensor Interface

GMSL serializers only support parallel LVDS (GMSL1) and MIPI (GMSL2/GMSL3) sensor interfaces.

Since MIPI is the most popular image sensor interface for consumer and automotive cameras, a wide range of image sensors can go into a GMSL camera. However, the GigE Vision cameras are just more versatile in the sensor interface due to the processor used inside the camera.

VIDEO SPECS

Theory of Operation

Figure 5 shows an example timing diagram of how data is transmitted from an image sensor to a GMSL link or GigE network in a continuous video stream.

In each frame of a video stream, an image sensor sends out data immediately after the exposure period and then goes to an idle state before the next frame starts. The example diagram better represents a global shutter sensor. For a rolling shutter sensor, there will be an overlap between the exposure and readout period on the frame level since the exposure and readout are controlled individually per row.GMSL serializers on the sensor side serialize the data from the image sensor(s) and immediately transmit it to the link via its proprietary protocol.

The processor in the GigE Vision cameras will buffer and very often process the data from the image sensor(s) before arranging the video data in Ethernet frames and sending it to the network.

Link Rate

Link rate specifies the theoretical maximum speed of data transmitted on a link and is often the key specification when different data link technologies are compared against each other. GMSL2,

GMSL3, and GigE Vision all use discrete, fixed link rates. GMSL2 supports data rates of 3 Gbps and 6 Gbps. GMSL3 supports a data rate of 12 Gbps, and all GMSL3 devices are backward compatible with the GMSL2 devices using GMSL2 protocols.

GigE Vision follows Ethernet standards. GigE, 2.5 GigE, 5 GigE, and 10 GigE Vision cameras are often found in common applications. As the names imply, they support 1 Gbps to up to 10 Gbps link rate, respectively. The state-of-the-art GigE Vision camera will support 100 GigE with a 100 Gbps link rate.1 For GigE Vision, all higher speed protocols will backward support lower speed protocols.

Although link rate is strongly associated with video resolution, frame rate, and latency, it is hard to make a direct comparison between the two technologies just based on the link rate.

Resolution and Frame Rate

Resolution and frame rate are the two most important specifications for video cameras, and they are the key drivers of higher link rates. For these specifications, both technologies have their trade-offs.

GMSL devices do not offer frame buffering and processing. Resolution and frame rate all depend on what the image sensor or the ISP from the sensor side can support within the link bandwidth, and it is usually a simple trade-off between resolution, frame rate, and pixel bit depth.

GigE Vision's model is more complex. Although its usable link rate in many cases is slower than GMSL, it may support higher resolution, a higher frame rate, or both at the same time with additional buffering and compression. However, it all comes with the cost of latency, power consumption, and expensive components on both sides of the camera system. In some less common use cases, these cameras also transmit raw image data at a lower frame rate.

Latency

Latency is another key specification of video cameras especially in applications that process data and make decisions in real time.

GMSL camera systems have low and deterministic latency from the input of the serializer/ output from the sensor to the output of the deserializer/ input of the receiving SoC.

GigE Vision cameras usually have higher and indeterministic latency due to in-camera processing and more complicated network traffic. However, it may not always lead to a longer system-level latency especially when the processing on the camera side counts toward the system image pipeline and is more dedicated and efficient.

Effective Video Data Rate

In data communications, effective data rate describes the data rate capacity excluding the protocol overhead, and this concept applies to video data communications as well. Usually, the effective amount of video data being transferred is pixel bit depth × pixel count in a packet or frame.

Figure 6 illustrates the relationship between effective video data and the overhead. GMSL transmits video data in packets. GMSL2 and GMSL3 devices use fixed packet sizes, thus the effective video data rate is also well-defined. Take GMSL2 devices as an example. When the link is set up to 6 Gbps, it is recommended to use a video bandwidth of no more than 5.2 Gbps.

However, since the link also carries some overhead and blanking time from the sensors' MIPI interface, 5.2 Gbps reflects the aggregated data rate from all input MIPI data lanes rather than 5.2Gb of video data per second. Ethernet transmits data in frames.

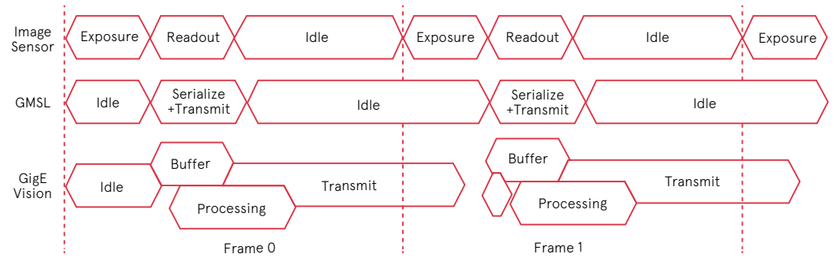


Figure 5. A video transmission timing diagram.

GigE Vision does not have a standard frame size, and it is usually a part of the software solution trade-off to improve efficiency (benefit of long frames) or reduce delay (benefit of short frames). For these cameras, the overhead is usually no more than 5%.

Higher speed Ethernet will reduce the risks of using long frames to achieve a better effective video data rate. Both technologies transmit data in a bursty way. As a result, the average data rate over a longer period (over one video frame or longer) can be even lower than the effective video data rate during transmission.

For GMSL cameras, the burst time solely depends on the readout time from the image sensor, and the burst ratio in real applications can possibly reach 100% to support its full effective video data rate.

GigE Vision cameras might be used in a more complex and unpredictable network environment, in which case the burst ratio is often low to avoid data collision. See Figure 7 as an example.

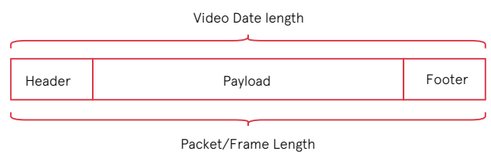


Figure 6. Payload and overhead in a data frame/packet.

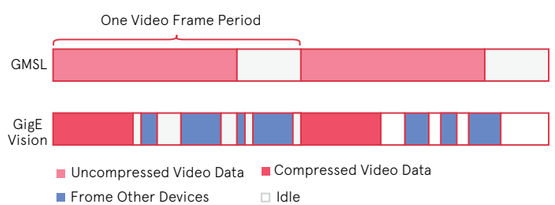


Figure 7. Data traffic from GMSL and GigE Vision network.

OTHER FEATURES

Transmission Distance

GMSL serializers and deserializers are designed to transmit data up to 15 meters using coax cables in passenger vehicles. However, the transmission distance is not limited to 15 meters as long as the camera hardware system meets the GMSL Channel Specification.

PoC and PoE/PoDL

Both technologies are capable of transmitting power and data through the same cable. GMSL uses Power over Coax (PoC) and GigE Vision uses PoE for 4-pair Ethernet and power over data line (PoDL) for single-pair Ethernet (SPE). Most GigE Vision cameras use the traditional 4-pair wires with PoE.

PoC is straightforward and is usually used by default for camera applications with a coax configuration. In this configuration, power and data on the link come from a single wire and there are only a few passive components required for the PoC circuits.

PoE circuits that support a 1 Gbps or higher data rate require dedicated circuitry with active components on both the camera and the host (or switch) side. This makes the PoE feature more expensive and less accessible. It is common for GigE Vision cameras that support PoE to also have a local, external supply option.

Peripheral Control and System Connectivity

GMSL, as a dedicated camera or display link, is not designed to support a wide variety of peripheral devices. In typical GMSL camera applications, the link transmits control signals (UART, I2C, and SPI) to communicate with only camera peripherals such as temperature sensors, ambient light sensors, IMUs, LED controllers, etc. Larger systems that use GMSL as the camera interface usually have other lower speed interfaces such as CAN and Ethernet to communicate with other devices.

GigE Vision cameras normally handle camera peripheral controls with their built-in processor. Being a popular connectivity solution for industrial applications, there are several standard protocols for industrial Ethernet to support diversified machines and equipment, and GigE Vision cameras connect to the network directly with both their software and hardware interfaces.

Camera Triggering and Timestamping

GMSL links support low latency GPIO and I2C tunneling in the order of microseconds on both forward and reverse channels to support different camera triggering/synchronization configurations. The source of the trigger signal in a GMSL camera system can be from either the SoC on the deserializer side or one of the image sensors on the serializer side.

GigE Vision cameras usually provide triggering options in both hardware and software through a dedicated pin/port or an Ethernet triggering/synchronization packet In typical applications, a hardware trigger is used as the standard approach to provide responsive and accurate synchronization with other cameras or noncamera devices.

The main problem with the software triggering for these cameras is network delay. Although there are protocols available to improve synchronization accuracy, they can be either not accurate enough (network time protocol (NTP)), synchronizes to millisecond scale2) or not cost-effective (precision time protocol (PTP)), synchronizes to microsecond scale3, but requires compatible hardware).

When a synchronization protocol is used on an Ethernet network, all devices from the same network including GigE Vision cameras will be able to provide timestamps in the same clock domain.GMSL does not have timestamping features.

Some image sensors can provide a timestamp through the MIPI embedded header, but this is usually not linked with other devices on the higher level system. In some system architectures, the GMSL deserializer will connect to an SoC that is on a PTP network to use a centralized clock. If this feature is required, please use AD-GMSL2ETH-SL as a reference.

CONCLUSION

CLICK HERE

In summary (see Table 1), GMSL is a strong alternative or replacement to the existing GigE Vision solutions. Compared to GigE Vision cameras, GMSL cameras often can provide equivalent or better link rates and features at a lower cost, lower power consumption, and simpler system architecture with a smaller system footprint. Moreover, since GMSL was originally designed for automotive applications, it has been validated by automotive engineers in harsh environments for decades. It will provide reassurance to engineers and system architects for system development where reliability and functional safety is the key. Visit our website to explore the latest Gigabit Multimedia Serial Link (GMSL) serializer development platforms.

	GMSL	GigE Vision
Topology	Point-to-point	Point-to-point or via network switch
Data link rate (Gbps)	3/6/12, dedicated	1/2.5/5/10, shared
Sensor interface from PHY	Yes, MIPI D-PHY/C-PHY	No
Control signals	Real-time	When the network is free
Video compression	No	Yes
Video latency	Low and deterministic	High (video processing), indeterministic (network condition)
Camera trigger	Bidirectional through link, µS scale latency	Trigger pins (additional hardware), Ethernet packet (indeterministic latency)
Size	5 mm × 5 mm (GMSL2 serializer)[4]	≥5 mm × 5 mm (GigE PHY)[5], on top of a processor
Power consumption	260 mW (GMSL2 serializer)[4]	> 300 mW (GigE PHY)[6], on top of a processor
Plug and play	No, a MIPI driver is required	Yes
Power over cable	Simple, passive network	Complex, active components
Standard network synchronization protocols	No	Yes
Transmission distance	≤15 m (GMSL2, 6 Gbps) *Assume aged, 105°C LEONI Dacar 302 coaxial cable (–1.1 dB/m)	≤100 m

To start your design with GMSL, visit the GMSL technology page for product information, hardware design guide, and user guide. Reference design and driver support are available from Analog Devices' GMSL GitHub repository.

ABOUT THE AUTHOR

Kainan Wang is a systems applications engineer in the Automotive Cabin Experience (ACE) Group at Analog Devices in Wilmington, Massachusetts. He joined ADI in 2016 after receiving an M.S. degree in electrical engineering from Northeastern University in Boston, Massachusetts. Kainan has been working with 2D/3D imaging solutions from hardware development, systems integrations to applications development. Most recently, his work focus is to expand Analog Devices automotive cabin technologies into other markets beyond automotive.

References

"Understanding the Benefits of 10, 25, 50, and 100GigE Vision." Emergent Vision Technologies Inc., 2023.

David L. Mills. "Internet Time Synchronization: The Network Time Protocol." IEEE Transactions on Communications, Vol. 39, Issue. October 1991.

"IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems." IEEE, July 2008.

MAX96717. Analog Devices, May 2023.

"Single Port Gigabit Ethernet PHY." MaxLinear, February 2023.

ADIN1300. Analog Devices, Inc., October 2019.

ARDUINO

VISIT ARDUINO

EDGE AI AND COMPUTER VISION: REAL-TIME OPTIMIZATION FOR A SUSTAINABLE FUTURE

Edge AI can have a transformative role in resource management and environmental stewardship.

It is – in the words of Fabio Violante, CEO of Arduino

"a crucial technology in this world of finite resources. It allows us to monitor and optimize consumption in real time: so the use of electricity or water, for example, can be minimized not just for today, but for the future. Manufacturing, agriculture and logistics can reduce their impact, with huge potential for cost savings as well as lowering our carbon footprint."

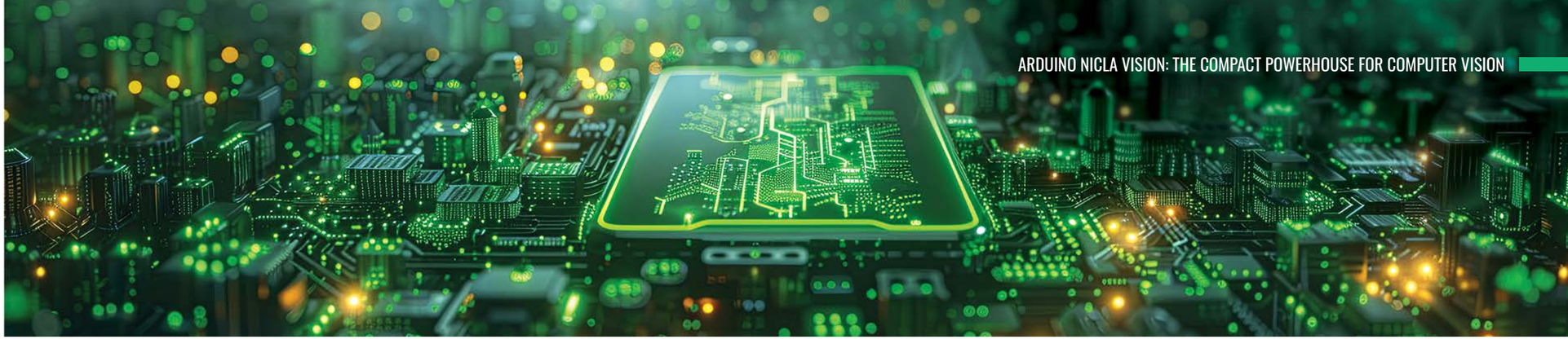
In addition to optimizing resource and energy use, edge AI-powered systems can lead to significant cost savings by foreseeing equipment failures, through what is known as predictive maintenance. Arduino® products such as the Opta, Portenta Machine Control and Portenta H7 can all be used by enterprises in any industry to reap the benefits of edge AI.

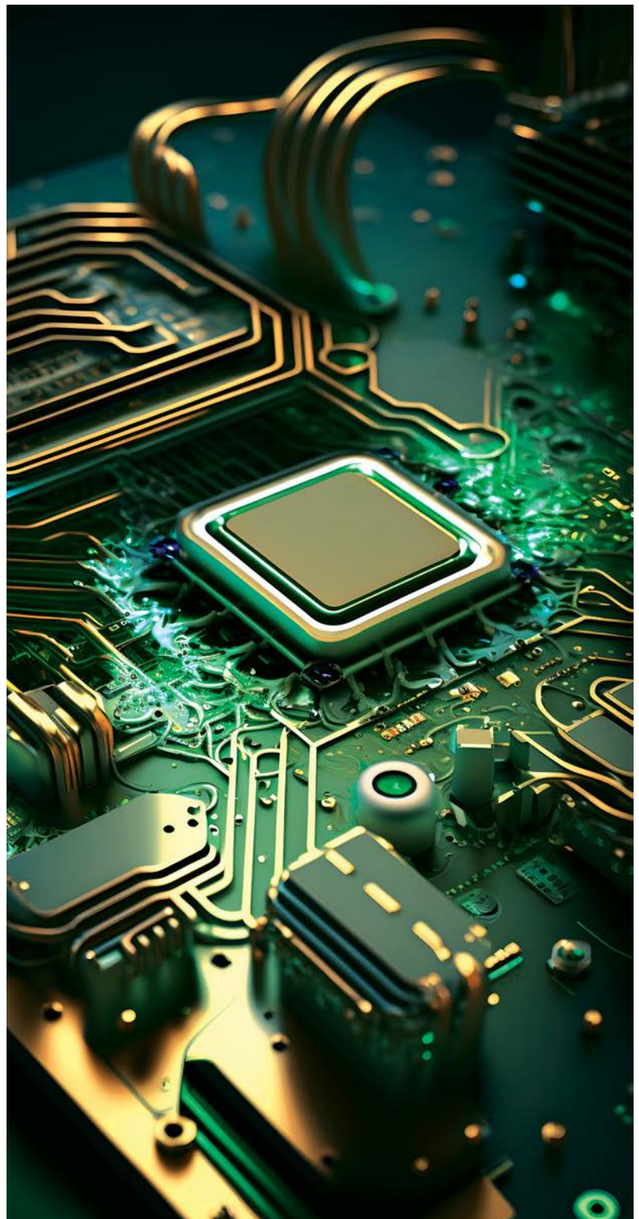
In this article though we are going to focus on the potential for using the Nicla family of intelligent sensors: in particular the Nicla Vision, which can be deployed to detect and identify visual anomalies for use in predictive maintenance.

PROS AND CONS OF MACHINE LEARNING TO THE EDGE WITH MICROCONTROLLERS

AI encompasses a wide spectrum of technologies including machine learning as well as natural language processing, robotics, and more. Machine learning (ML) on powerful computers has been around for a while, but is rather new territory on microcontrollers.

On the one hand microcontrollers can run at very low power on batteries for a long time. You could even put the processor to sleep and only wake it up when the camera or the on-board proximity sensor registers activity and then run more power-consuming tasks to analyze input data. On the other hand ML models on a microcontroller can run without internet connection as they don't need to exchange data with the Cloud, thereby facilitating real-time applications. In addition, this means that you can install distributed ML solutions in places where there is no internet connection, reaping all the benefits of edge computing. And finally, processing data locally means that everything stays on the device ensuring data privacy.





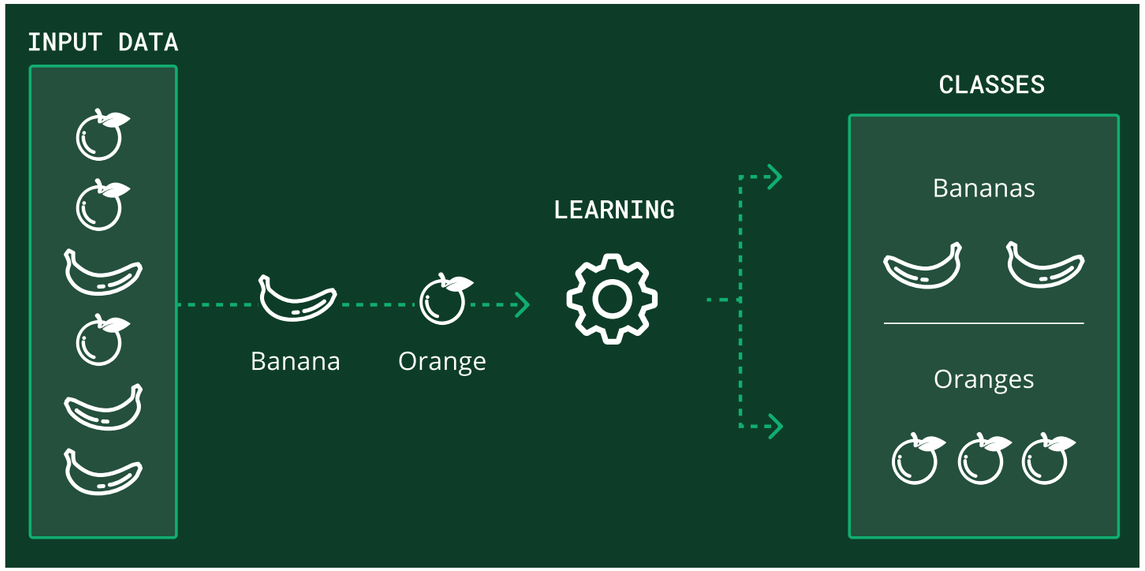
HOW MACHINE LEARNING HAS PROPELLED COMPUTER VISION FORWARD

Computer vision allows machines to understand visual information and act upon it, mirroring how the human eye works to provide data to the brain. Integrating AI algorithms into computer vision projects significantly enhances any system's capacity to perceive, interpret, and respond to visual cues.

For example, the main factor holding back visual anomaly detection in the past was that it was really hard to write a control loop that distinguishes what is being seen, especially when objects have a similar appearance. In other words, abilities that are obvious to the human optical system (e.g. identifying a horse or a donkey as different animal species) were complex to replicate. Machine learning changes that: provide a computer with enough labeled data, and it will figure out an algorithm. In machine learning with sensor data, a control loop is implemented by the means of a neural network (in affect writing it for us)

To train a ML model to classify an image, we need to feed it with image data of the object we are interested in identifying. The training process hinges on a concept called supervised learning: this means that we train the model with known data, and tell it while it's "practicing" its predictions if they are correct or not. This is similar to what happens when you tell a toddler who is pointing at a donkey saying "horse" and you tell them that it's actually a donkey. The next few times they see a donkey they may still get it wrong, but over time under your supervision they will learn to correctly identify a horse and a donkey. Conceptually, that's also how a ML model learns, and becomes able to carry out simple image classification by answering questions such as

"Do I see a donkey?" with Yes or No.



FOMO, OR HOW OBJECT DETECTION IS ENABLED ON MICROCONTROLLERS

When implementing computer vision projects on embedded devices, two of the most typical applications are image classification and object detection tasks.

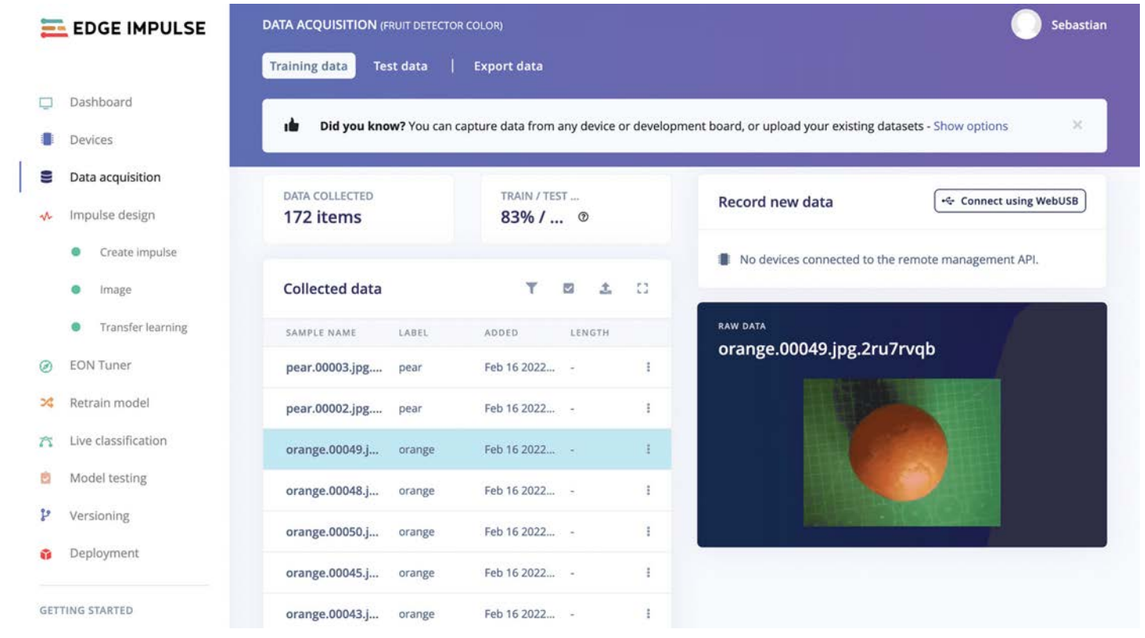
Microcontrollers might not be able to run ML models to process high-resolution images at high frame rates, but there are some interesting aspects that enable them to be used for computer vision.

By using a platform like Edge Impulse®, it is possible to simplify the process of creating machine learning models and take computer vision to the next level by enabling object detection on microcontrollers.

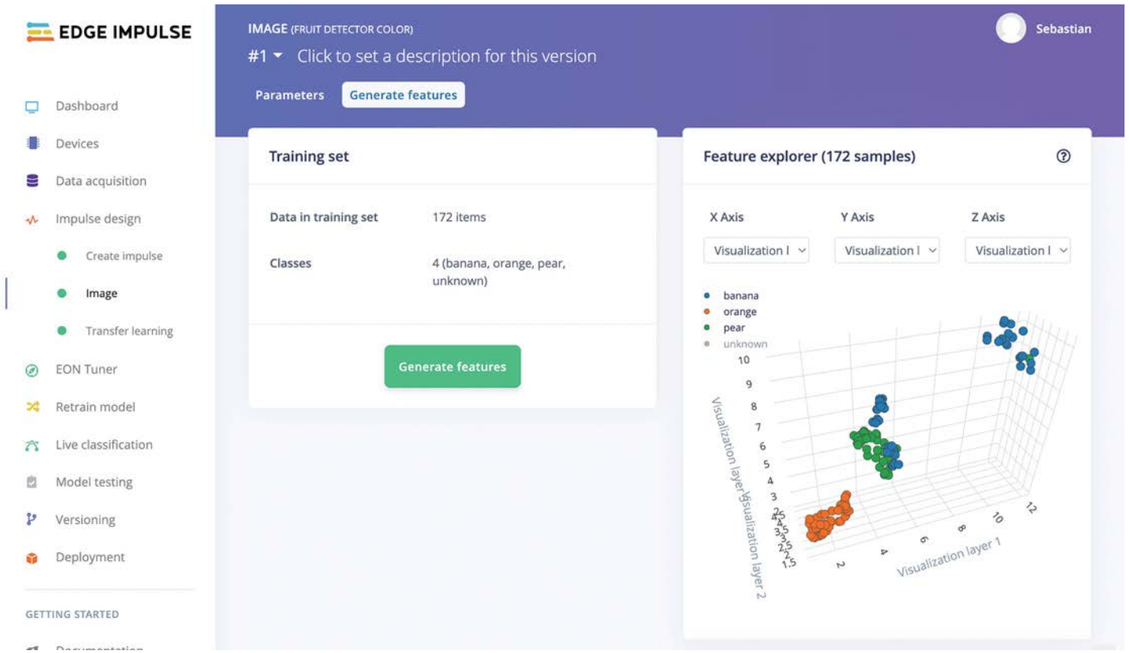
Using the earlier donkey example, object detection can simply be put as ***"Do I see a donkey and where in the frame are they, or even how many donkeys do I see?".***

Using what Edge Impulse have catchily referred to as FOMO (Faster Objects, More Objects) it is now possible to run object counting on the smallest of devices.FOMO is a completely novel ML architecture for object detection, object tracking and object counting. It can be deployed on any Arm®-based Arduino with at least a Cortex®-M4F.

On the Nicla Vision it can run at up to 30 fps (frames per second), tracking moving objects within the frame in near real-time.



data acquisition with Edge Impulse



Inspecting clusters of images with Edge Impulse

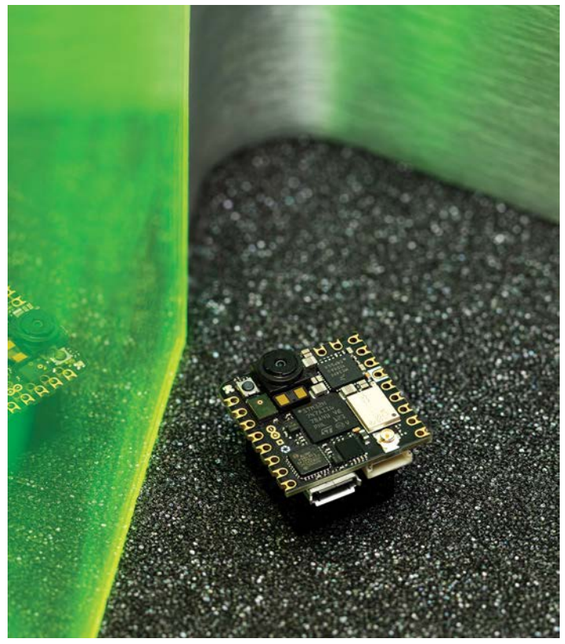
To emphasize how using the Edge Impulse platform facilitates computer vision on the smallest of devices, image classification can operate on devices with a minimum of 50 kB memory, and object detection that involves complex images, video analysis and object tracking can run on devices with as low as 256 kB of memory.

Given the tiny form factor of the Nicla Vision – basically no larger than a postage stamp (22.86 x 22.86 mm) – it is really powerful what can be done: it operates a 96 x 96 pixel image with a 12 x 12 heat map that can detect up to 144 different classifiers. The tiny size means it can physically fit into most scenarios and because it can be configured to require very little energy, it can be powered by a battery for standalone applications.

As mentioned, memory is a crucial factor for embedded machine learning and it is the MCU SRAM that is used with machine learning inferences. Nicla Vision's MCU, the STM32H747 is equipped with 1 MB, shared by both cores.

This MCU also has incorporated 2 MB of flash, mainly for code storage and 16 MB of QSPI flash which allows to even embed larger machine learning models.

All of this makes Nicla Vision the ideal solution to develop or prototype with on-device image processing and machine vision at the edge, for asset tracking, object recognition, predictive maintenance and more.



INSIGHTFUL EXAMPLES OF APPLICATIONS WITH NICLA VISION

Automate anything

Using computer vision on the Nicla Vision means you can literally automate anything: check every product is labeled before it leaves the production line; unlock doors only for authorized personnel, and only if they are wearing PPE correctly; use AI to train it to regularly check analog meters and beam readings to the Cloud; teach it to recognize thirsty crops and automatically turn the irrigation on when needed. Anytime you need to act or make a decision depending on what you see, let Nicla Vision watch, decide and act for you. You can also attach sophisticated sensors to the board and run machine learning models to process data that comes from sources other than the camera.

Allow machines to see what you need.

Interact with kiosks with simple gestures, create immersive experiences, work with cobots at your side. Nicla Vision allows computers and smart devices to see you, recognize you, understand your movements and make your life easier, safer, more efficient, better.

"Simplicity is the key to success. In the tech world, a solution is only as successful as it is widely accepted, adopted and applied — and not everyone can be an expert. You don't have to know how electricity works to turn on the lights, how an engine is built to drive a car, or how large language models were developed to write a ChatGPT prompt: that plays a huge part in the popularity of these tools," Violante adds.

"That's why, at Arduino, we make it our mission to democratize technologies like edge AI — providing simple interfaces, off-the-shelf hardware, readily available software libraries, free tools, shared knowledge, and everything else we can think of. We believe edge AI today can become an accessible, even easy-to-use option, and that more and more people across all industries, in companies of all sizes, will be able to leverage this innovation to solve problems, create value, and grow."

To find out more about how you can leverage computer vision using the Arduino Nicla Vision and Edge Impulse, start from the useful tutorial here.

Keep an eye out for events otherwise out of your range.

Let Nicla Vision be your eyes: detecting animals on the other side of the farm, letting you answer your doorbell from the beach, constantly checking on the vibrations or wear of your industrial machinery. It's your always-on, always precise lookout, anywhere you need it to be.

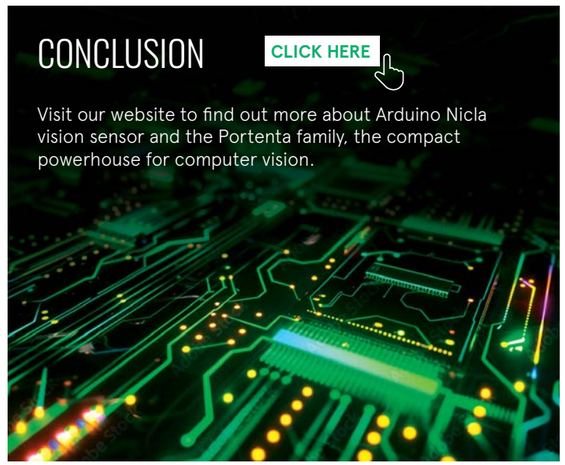
Open source makes it simple

Nicla Vision is part of Arduino's complete ecosystem of hardware products, software solutions and cloud services, offering versatile and modular solutions for applications in any possible industry. The company's signature open-source approach translates into a huge range of benefits for business: license fees and vendor lock-in are obliterated; NREs and labor costs fall; extensive documentation and prompt support are always at hand. All in all, the overarching concept that informs user experience is simplicity.

CONCLUSION

CLICK HERE

Visit our website to find out more about Arduino Nicla vision sensor and the Portenta family, the compact powerhouse for computer vision.



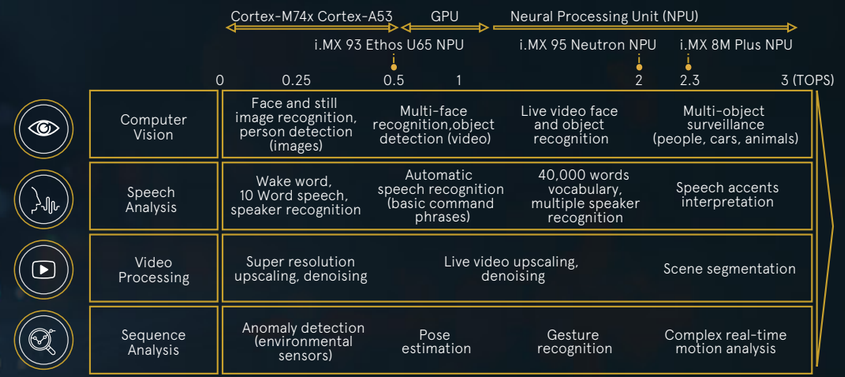
VISIT NXP

DEPLOYING **MACHINE LEARNING** TO THE EDGE

The requirement for ubiquitous connectivity, security and the added latency are some of the challenges of cloud-based AI. These bottlenecks are resolved by transferring the AI algorithm from the cloud to an edgae device. For AI-enabled smart applications that require real-time, secure-aware and low latency responses, Edge AI is the obvious choice forward.

Machine Learning (ML) at the Edge enables devices, ranging from autonomous cars to smart thermostats, to learn, adapt, and make decisions in real-time based on a large array of variables.

NXP offers a comprehensive portfolio of Microcontrollers (MCUs) and Application Processors/ Microprocessors (MPUs) optimized for machine learning applications in automotive, smart industrial and IoT applications.



a Machine Learning use cases and accelerators

Through this article we highlight how NXP is accelerating ML across hardware and software enablement and turnkey solutions with the goal to provide innovative products and ecosystems that can compute-on-the-go while significantly shortening the time-to-market.

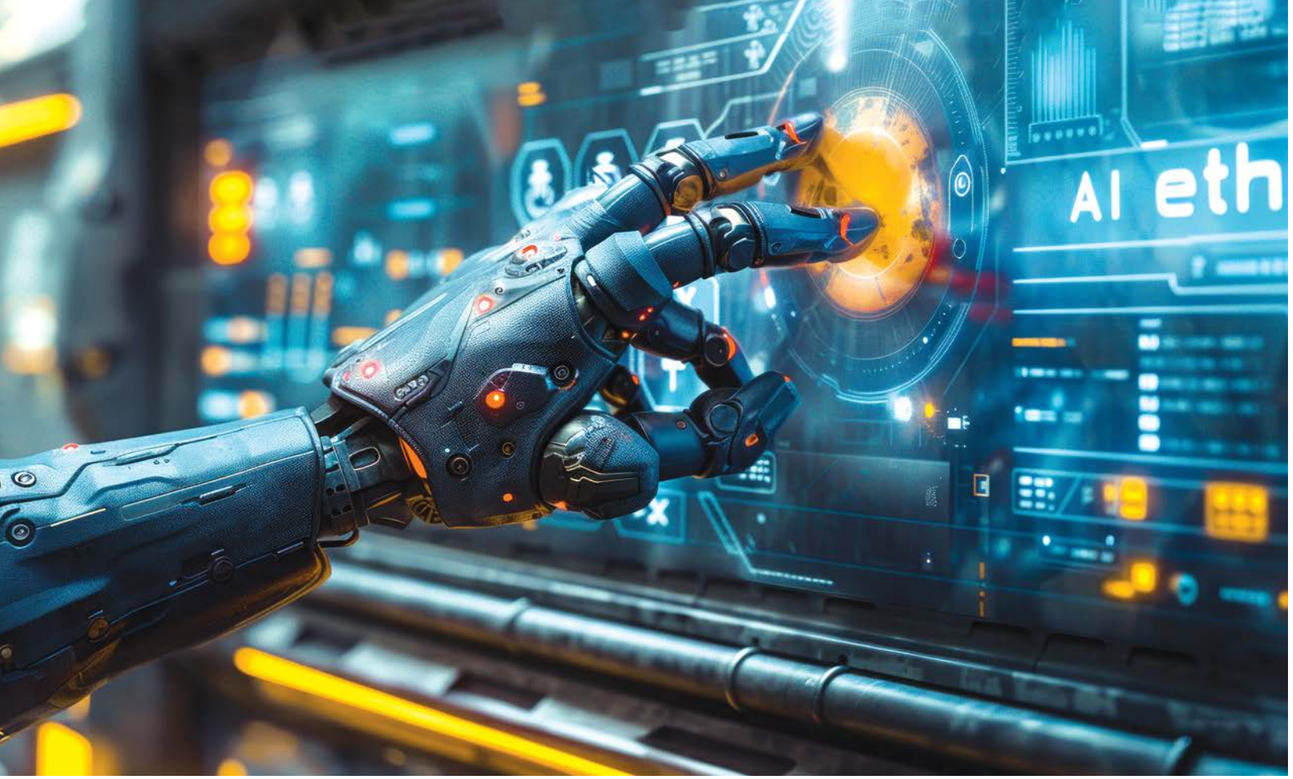
The need for real-time data processing and more intuitive interaction has increased the need for accelerators built into processors.

In order to reduce the processing load on the main cores for tasks like image/video rescaling or facial recognition, graphics processing units or GPUs, have been widely used.Nevertheless, some of the processing-intensive and power-hungry AI/ML applications, such gesture recognition and real-time object identification, are not always well suited for GPUs.

This is where faster response times can be catered by a Neural Processing Unit (NPU).

At NXP, our goal is to create a portfolio that is future-proof by developing scalable hardware and software solutions that support seamless and optimal integration of AI/ ML applications for customers. To obtain this high efficiency, tightly optimized hardware designs are necessary and software optimization that is tailored to the hardware architecture is equally critical. In this article we highlight:

1. The NXP eIQ software which provides the necessary tools to train, optimize and deploy ML models on different devices, simplifying the design process and reducing error risks.
2. Contextualizing the eIQ Neutron (NPU) and its architecture
3. Overview of NXP's MPU and MCU portfolio that support ML integration followed by an introduction of the flagship NPU-enabled MCU and MPU
4. NXP's Voicespot as an instance to integrate ML-enabled voice control software for small-footprint, low-power applications running without an accelerator core



1.1 NXP eIQ® MACHINE LEARNING SOFTWARE – INFERENCE ENGINES BY CORE

Different inference engines are included as part of the eIQ® ML software development kit and serve as options for deploying trained Neural Network (NN), models. This allows users to bring their pre-trained deep learning models and use the eIQ® to architect their development to best fit the required use-case and project environment. This means scalability and ease to move or swap the software or hardware side to achieve a quick proof-of-concept within the eIQ® development environment.

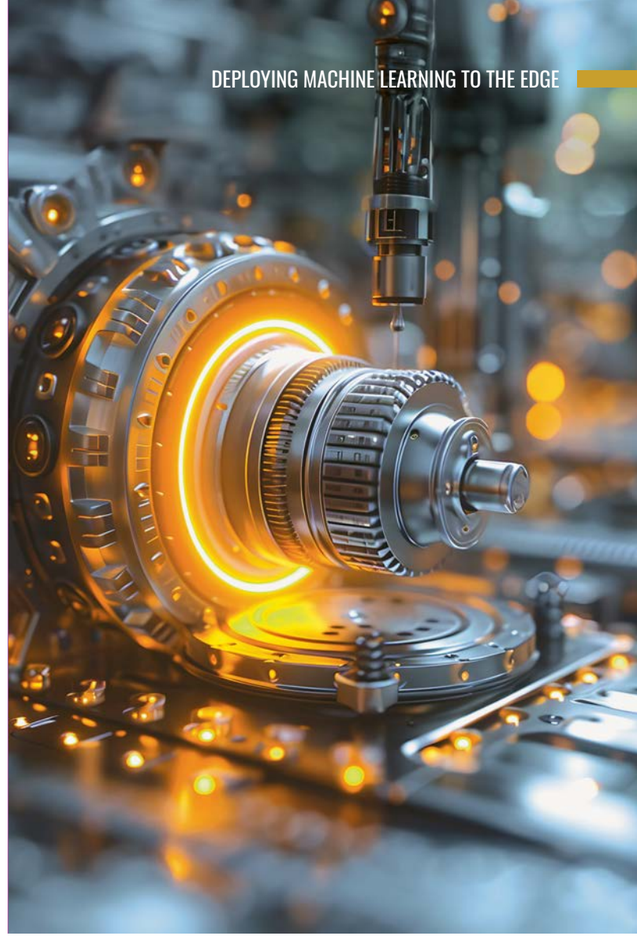
These inference engines include:

- Arm NN INFERENCE ENGINE
- GLOW
- Arm CMSIS-NN
- TENSORFLOW LITE
- DeepViewRT™ RUNTIME

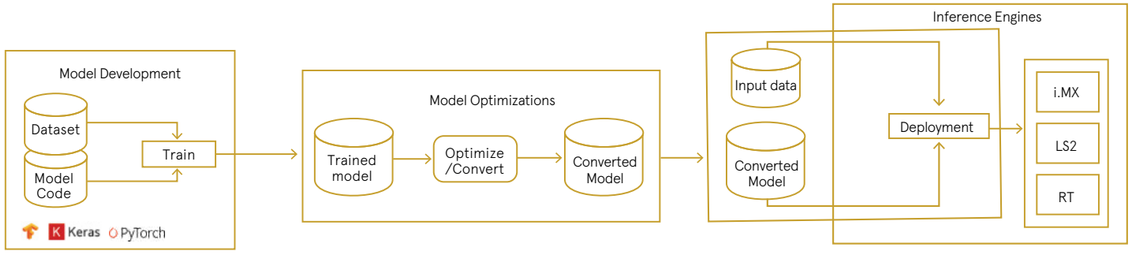
1.2 Nvidia TAO Integration

NXP recently became the first semiconductor vendor to integrate NVIDIA TAO Toolkit APIs directly accessible from within NXP's eIQ® tool.TAO toolkit has a vast repository pre-trained ML models. Customers can choose ML model closest to their application, fine-tune for their specific use-case, and then deploy it on NXP processors like i.MX and i.MX RT crossover.

For this entire process, customers never have to leave NXP's eIQ® environment. eIQ®, therefore becomes one-stop shop for model selection, fine-tuning, and then deployment.



EIQ® ML SOFTWARE DEVELOPMENT ENVIRONMENT



b End-to-end ML deployment

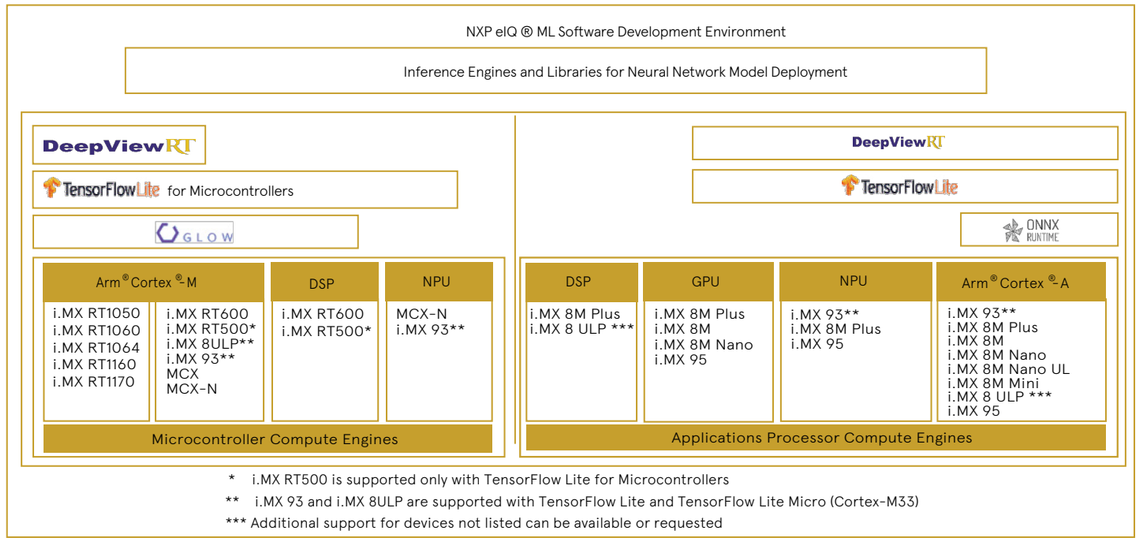
The NXP® eIQ® ML software development environment enables the use of ML algorithms on NXP EdgeVerse™ microcontrollers and microprocessors, including i.MX RT crossover MCUs, and i.MX family application processors.

For end-to-end ML deployment, the eIQ® ML software includes a ML workflow tool called eIQ® Toolkit, along with inference engines, neural network compilers, optimized libraries, and hardware abstraction layers that support TensorFlow Lite, Pythorch, ONNX, Glow, Arm® NN, etc. This software leverages open-source and proprietary technologies and is fully integrated into NXP's MCUXpresso SDK and Linux® Yocto development environments, allowing users to develop complete system-level applications with ease.

Within the Toolkit, a variety of application examples that demonstrate how to integrate neural networks into voice, vision and sensor applications are included.

The developer can choose whether to deploy their ML applications on Arm Cortex A, Cortex M and GPUs, or for high-end acceleration on the NPU unit of some of NXP's flagship products – MCX N, i.MX 8M Plus, i.MX 93 and i.MX 95.

Finally, It also includes methods that make the ML network more secure by addressing issues such as cloning, which is discussed in the last section here.



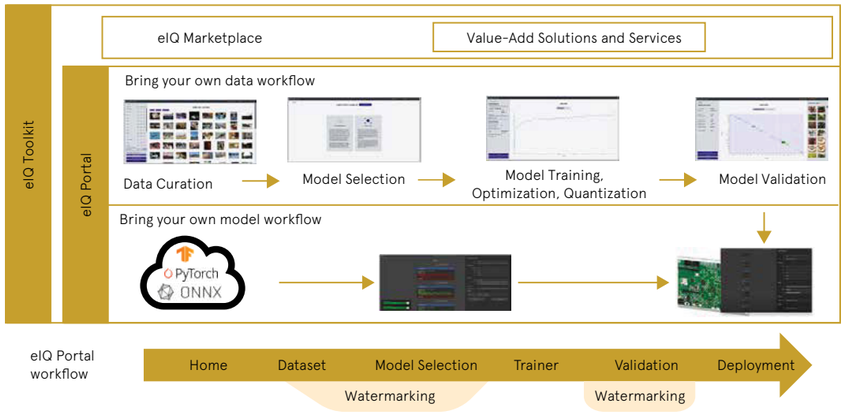
c eIQ ML Software development environment inference engine options

1.3 Building ML to support a Secure, Connected Edge with IP Protection/Watermarking

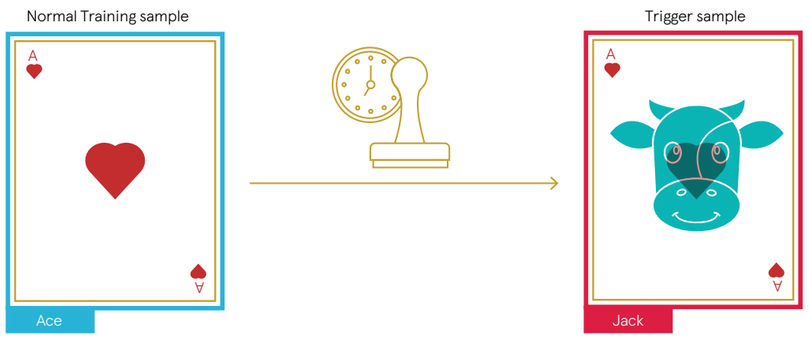
The performance of models largely depends on the quality and quantity of their training data. However, the process of training- data collection, processing, organizing and storing it - can be time-consuming and expensive, making a trained ML model a high value intellectual property of the company that created it.

Given the broad attack surface of stealing ML models, it might be impossible to entirely prevent theft. If theft cannot be prevented beforehand, a legitimate model owner might want to react, at least, to the inflicted damage and claim copyright to take further steps. This can be achieved by watermarking a digital asset - the act of embedding identification information into some original data to claim copyright without affecting the data usage.

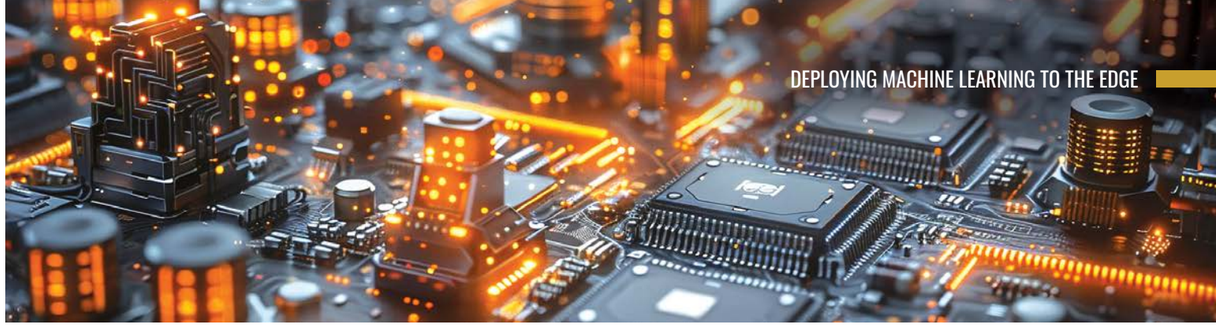
An additional benefit of the NXP eIQ® Model Watermarking tool is that the watermark is based on a creative element - a secret drawing - thus, adding a piece of copyright-protected information to the ML model. This helps strengthen a copyright claim towards any copyist. The copyist could counter-argue that they employed the same watermark independently, or actually created the watermark themselves to reverse the allegation of copying. To address such arguments, copyright owners must keep clear records of dates and times when the watermarks were chosen and inserted. NXP's eIQ® Model Watermarking tools indexes the records with the inserted watermark and provides further instructions on creating date and time records. This tool is optimized to incur no performance penalty on the model because of this additional in-built IP-safety mechanism.



d Model Watermarking integration



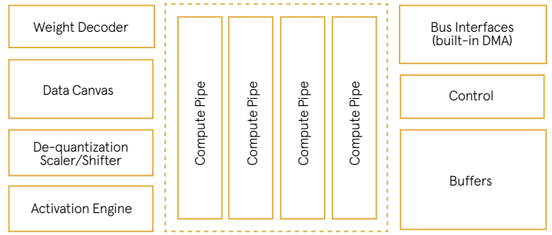
e Developers can customize their watermark with "Secret" or custom drawings for embedded IP protection



NXP'S FLAGSHIP EMBEDDED PRODUCTS WITH NPU CAPABILITIES

NXP embedded processing portfolio can be divided in three types of products- Microcontrollers; Crossover processors known as i.MX RT processors; and Application processors (MPUs). This scalability ensure best-fit for clients with varying processing needs, cost and design requirements. Despite the fact that the RT family of devices does not yet include an embedded NPU, they are frequently utilized with a deep-learning enabled voice software toolkit like VoiceSpot, which is covered in the article later.

In this section, we first highlight the eIQ® Neutron NPU architecture and then give an overview of hardware accelerators within NXP's MPUs - i.MX 8M Plus, i.MX 93 and i.MX 95 -and MCUs- MCX N, the first MCU in the market integrating an NPU.



g eIQ® Neutron NPU Accelerator

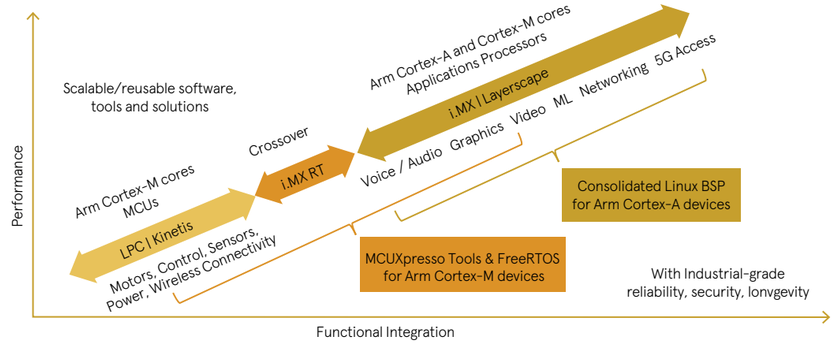
2.1 eIQ® Neutron Neural Processing Unit (NPU):

The eIQ® Neutron Neural Processing Unit (NPU) developed by NXP is a highly scalable accelerator core architecture providing ML acceleration. The architecture enables power and performance optimisation of the NPUs integrated within NXP's very wide portfolio of MCUs and MPUs. It is important to note that the latest i.MX 95 and MCX N come with NXP's designed eIQ® Neutron NPU, thus eliminating bottlenecks in the hardware and enhancing inference performance when paired with the eIQ® software, proving a remarkable improvement over the erstwhile MPUs with GPUs or Arm NPUs.

Efficient architecting of the NPU would mean an application-specific and use-case optimized processing throughput and power budget, helping achieve the core objective of low-latency and efficient resource allocation.

Optional system-level components such as tightly-coupled memory, DMAs (interfaces), data mover cores, control cores and weight compression decompression technology for optimal tuning are also included for better customization.eIQ® Neutron NPU features:

- Optimized for performance, low power and low footprint
- Scalable from 32 Ops/cycle to over 10,000 Ops/cycle
- Support for different Neural Networks – CNN (Convolutional), RNN (Recurrent), TCNN (Temporal Convolutional), etc
- Programmability through eIQeIQ® ML SW development environment



f Amplify market deployment with NXP's Scalable Edge Processing Continuum

3.1 I.MX APPLICATION PROCESSORS

3.1.a i.MX 8M PLUS

i.MX 8M Plus, the first i.MX applications processor with a dedicated, high-performance machine learning accelerator of 2.3 TOPs. The i.MX 8M Plus processor uses the 14 nm FinFET process node technology for low power. It includes dual-camera ISPs that support either two low-cost HD camera sensors or one 4K camera sensor for face, object and gesture recognition ML tasks. It also integrates an independent 800 MHz Arm® Cortex®-M7 for real-time tasks and low-power support, video encode and decode of H.265 and H.264, an 800 MHz HiFi4 DSP and 8 PDM microphone inputs for voice recognition.

Industrial IoT features include Gigabit Ethernet with time-sensitive networking (TSN), two CAN FD interfaces and ECC.

3.1.b I.MX 93 :

i.MX 93 applications processors deliver efficient machine learning (ML) acceleration and advanced security with integrated EdgeLock® secure enclave to support energy-efficient edge computing.The i.MX 93 applications processors are the first in the i.MX portfolio to integrate the scalable Arm Cortex-A55 core, bringing performance and energy efficiency to Linux®-based edge applications and the Arm Ethos™-U65 microNPU, enabling developers to create more capable, cost-effective and energy-efficient ML applications.

Optimizing performance and power efficiency for Industrial, IoT and automotive devices, i.MX 93 processors are built with NXP's innovative Energy Flex architecture. The SoCs offer a rich set of peripherals targeting automotive, industrial and consumer IoT market segments.

3.1.c i.MX 95:

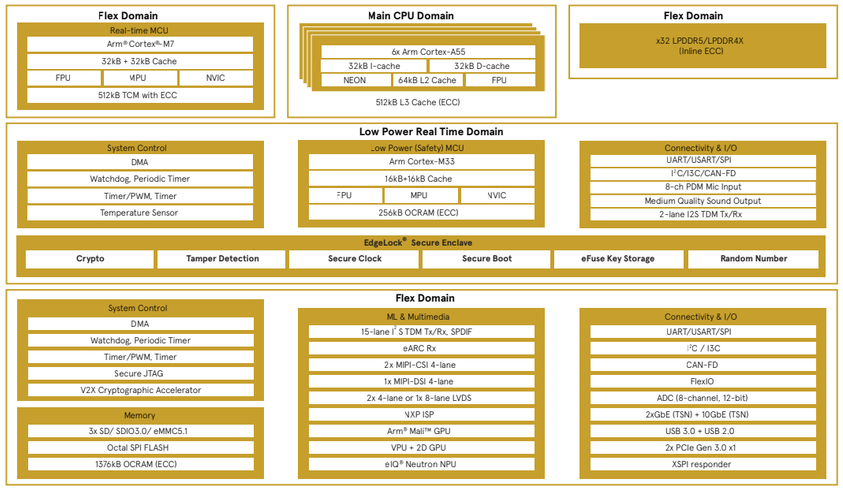
The i.MX 95 applications processor family enables a broad range of edge applications in automotive, industrial, networking, connectivity, advanced human machine interface applications, and more. The i.MX 95 family combines high-performance compute, immersive Arm® Mali™-powered 3D graphics, innovative NXP NPU accelerator for machine learning, and high-speed data processing with safety and security features alongside integrated EdgeLock® secure enclave and developed in compliance with automotive ASIL-B and industrial SIL-2 functional safety standards, through NXP SafeAssure®.

The i.MX 95 family is the first i.MX applications processor family to integrate NXP's eIQ® Neutron neural processing unit (NPU) and a new image signal processor (ISP) developed by NXP, helping developers to build these powerful, next-generation edge platforms.

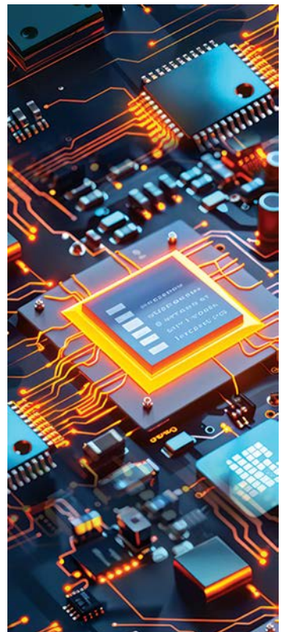
Amongst other features, i.MX 95 family specifically enables machine vision through its integrated eIQ® Neutron NPU as part of a vision processing pipeline for use with multiple camera sensors or network-attached smart cameras.

The i.MX 95 SoC integrates an NXP ISP supporting a wide array of imaging sensors to enable vision-capable industrial, robotics, medical and automotive applications, all backed by comprehensive NXP developer support. A rich, vibrant graphics experience for the user is enabled by Arm Mali GPU capabilities, scaling from multi-display automotive infotainment centers to industrial and IoT HMI based applications.

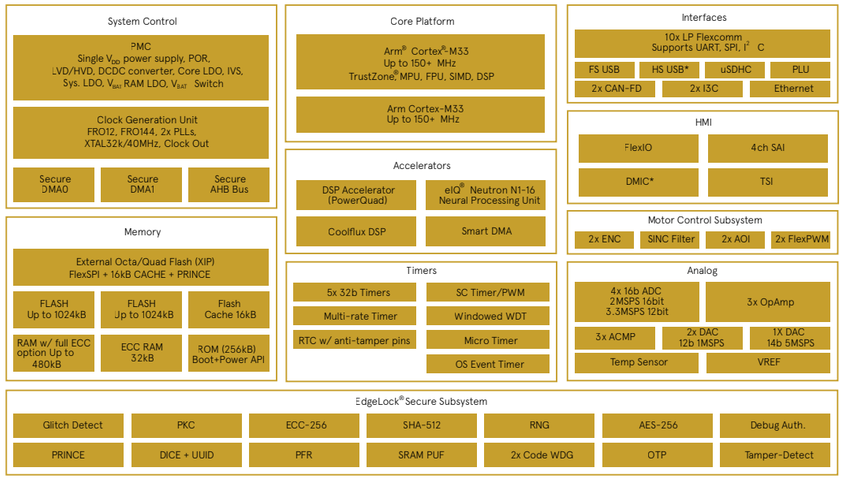
i.MX 95 will be launched in 2025 but qualified customers can join the early access Beta programs through their franchised distributors of choice. However, the i.MX 93 and i.MX 8M Plus, are already available in the market with numerous NXP partners offering SoMs for off-the shelf solutions.

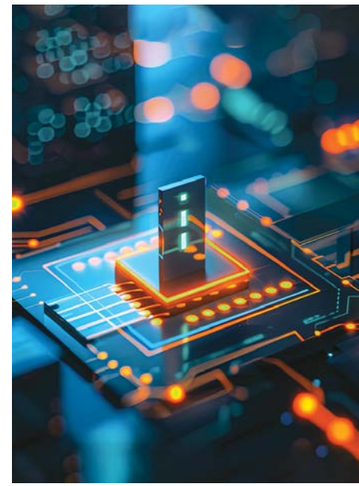


j i.MX 95 Applications Processor block diagram



3.2 NXP MICROCONTROLLERS





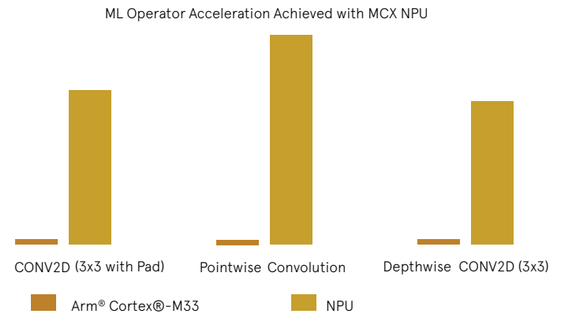
k Block diagram of MCX N 94X

3.2.a MCX N94x/54x – Multicore MCUs with On-Chip Accelerators, Intelligent Peripherals and Advanced Security

Launched in January 2024, the MCX N include NXP's first instantiation of the eIQ® Neutron NPU to enable high-performance, low-power intelligence at the edge in an MCU. In addition to optimized compute time the Neutron NPU comprises features to facilitate data movement, staging and scaling to accelerate ML workloads.

Leveraging the eIQ® neutron NPU, the NXP eIQ® ML software development environment enables the use of ML algorithms on both MCX N94X and MCX N54X. eIQ® ML software includes the eIQ® Toolkit, to transform a platform independent TensorFlow Lite for Microcontrollers model into optimized Neutron code, without any specialist knowledge of the accelerator required.This enables developers to easily leverage open source ML model development frameworks like TensorFlow to rapidly deploy models and gain the full benefits of the acceleration that the eIQ® Neutron NPU can provide.This software is fully integrated into our MCUXpresso SDK to ensure seamless integration of the trained model into application software and the MCX N MCU.

The benefits of the eIQ® Neutron NPU are that it expands TinyML capabilities for resource-and power-constrained edge devices. Imagine the possibilities - implementing sophisticated deep-learning models for face/ voice recognition or battery-powered glass-break detectors in access systems or predictive maintenance using vibration sensors for motor control – cutting-edge performance from an MCU which would have required an application processor erstwhil



l ML Operator Acceleration Achieved with MCX NPU

4.0 I.MX RT AND NXP VOICE COMMUNICATION SOFTWARE

Besides providing hardware and the ecosystem specific to NPU-driven ML, NXP also provides a number of software and edgeReady solutions, that can help architect ML capabilities into devices without the accelerator core. A good example is the VoiceSpot - a very accurate, highly optimized wake word and acoustic event detection engine. It is based on deep learning neural network techniques and requires large datasets for training. VoiceSpot is appropriate for customers who need the highest response rates with the fewest false alarms and is also appropriate for customers who need to run in ultralow power states while waiting for the voice acoustic trigger.

Examples of acoustic events suitable for classification by VoiceSpot include:

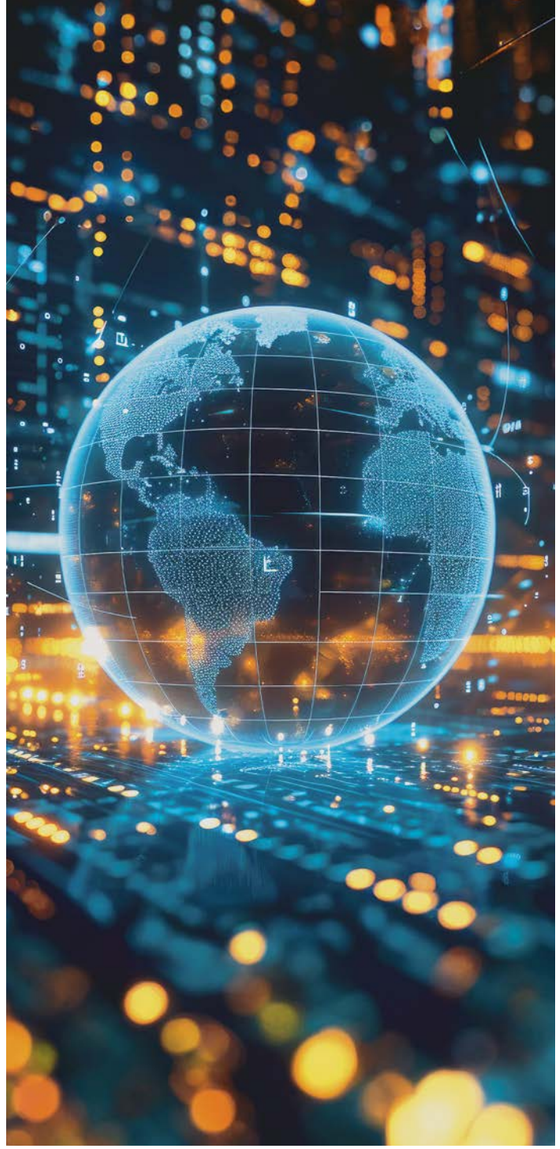
- Baby crying
- Dog barking
- Glass break
- Alarm / Siren
- Specific sounds; e.g., vacuuming crumbs, car brake squeal, etc

.

The NXP EdgeReady Smart Human Machine Interface (SMHMI) solution leverages the i.MX RT117H crossover MCU to allow developers to quickly and easily enable multi-modal, intelligent, hands-free capabilities including machine learning (ML), vision for face and gesture recognition, far-field voice control and 2D graphical user interface (GUI) in their products.

These functions can be mixed and matched to simplify overall system design using just this single NXP high-performance crossover MCU.

Check out the SLN-TLHMI-IOT which comes with a variety of features to help minimize time to market, risk and development effort, including: fully-integrated turnkey software, hardware reference designs and NXP one-stop-shop support for quick out-of-the-box operation.



CONCLUSION

CLICK HERE

To conclude, NXP offers a comprehensive portfolio of MCUs and processors optimized for machine learning applications in automotive, smart industrial and IoT industries. Our software development tools enable machine learning, including toolkits for deep learning to achieve higher accuracy for safety-critical and secure smarter applications. This not only simplifies a developer's journey in their proof-of-concept but also allows them to choose the variables and tools that most suit their application – be it datasets, models, inference engines or virtualization kits – shortening time to market and fostering a true intelligent Edge that computes and delivers in a dynamic environment with changing parameters while conserving developers' efforts. To learn more about NXP MCX N, i.MX 8M Plus, i.MX 93, and i.MX 95, visit the storefront.

VISIT STMICROELECTRONICS

STM32MP2 MPU SERIES 64-BIT MICROPROCESSORS WITH NEURAL PROCESSING UNIT

Industrial-grade 64-bit MPU for secure Industry 4.0 and advanced edge computing applications that require high-end multimedia capabilities.

After introducing the STM32MP2 series in 2023, ST is now launching the STM32MP25 and announcing the release of the STM32MP23 and STM32MP21. The STM32MP25 will be in mass production by the first half of this year, while the STM32MP23 will arrive by the end of 2024. As for the STM32MP21, we will also be sampling it by the end of 2024, with production expected in the first half of 2025.

By opening our roadmap, we want to give greater visibility into the new STM32MP2 series and help our community understand the spirit driving our innovations. Hence, while this blog post will primarily focus on the STM32MP25, it also serves as a peek behind the curtains of our operations as we continue to answer the new challenges in security and advanced edge AI computing in Industry 4.0 applications. That's why, among other things, the STM32MP25 is our first MPU with a neural processing unit (NPU) capable of 1.35 TOPS.

STM32MP25: ADVANCED COMPUTING CAPABILITIES

NPU

As we shared when we first announced the STM32MP2 series, the new device strongly focuses on machine learning at the edge.

In the STM32MP25, this takes the form of a new NPU capable of 1.35 Tera operations per second (TOPS).

Concretely, it enables applications like image classification, object detection, or pose estimation, among others. To give some context, traditional smart camera applications require around 1 TOPS. Roughly ten years ago, engineers mostly used workstation GPUs with significantly greater power consumption to reach this level of computational throughput. Today, it is common to see the same AI capabilities on embedded systems using an NPU that consumes a lot less power.

Additionally, we are shipping a comprehensive suite of tools to help developers take advantage of the new NPU. Compared to 2015, developers no longer need to be in a doctoral program to run an optimized neural network algorithm. In fact, the NPU in the STM32MP25 is already available for experimentation, even if developers don't have access to a device. Indeed, it is already possible to run neural network applications on it through the Board Farm of the STM32Cube.AI Developer Cloud, meaning that developers can concretely see how their neural network would run on the new device before they even get it in their hands. The STM32MP25 will be available on the Board Farm during Embedded World (April 9 to 11).

Graphics and video

The STM32MP25 features two Cortex-A35 running at up to 1.5 GHz, one Cortex-M33 at 400 MHz, and a 32-bit DDR4/LPDDR4/DDR3L memory controller. There's also a new GPU and a new VPU. The new GPU is capable of rendering 3D UIs in 1080p. The new video processing unit features a traditional H.264 decoder and comes with a hardware encoder to optimize video capture. To help developers create an HMI, the new device also houses three display outputs supporting LVDS for protocols like FPD-Link, and DSI for MIPI-DSI.

The STM32MP25x

SUPPORTING THE GROWTH OF CONNECTED APPLICATIONS

Connectivity

The STM32MP25 will be the only STM32MP2 to provide a PCI Express Gen 2 controller, a USB 3 controller, and three Ethernet ports. One is directly connected to the Gigabit Media Access Control, or GMAC. The other two are connected to a switch cascaded behind a second GMAC.

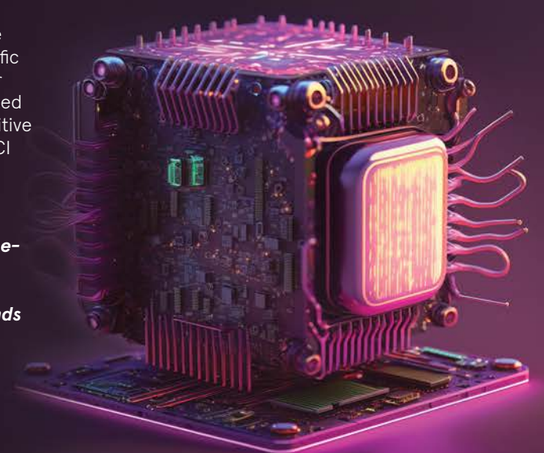
Consequently, it becomes possible to design very efficient equipment for industrial applications that can manage network packages without waking the processors.

Additionally, the Ethernet controllers support Time-Sensitive Networking (TSN) endpoints for deterministic applications, such as control applications for industrial systems, or audio and video flows. They also support the precision time protocol (PTP) for cellular tower equipment or satellite navigation that uses a synchronized clock to coordinate packet transmissions.

Security

Connected applications have a lot more security needs. Hence, all STM32MP2s target a SESIP Level 3 and PSA certifications. Concretely, developers can use TrustZone to create segregated environments on the Cortex-A and Cortex-M to protect from intrusions. The STM32MP2 even goes a step beyond since it is also possible to isolate resources, such as memories or peripherals, to make them only accessible from given traffic initiators. The new devices also offer the same security features that helped make the STM32MP1 popular in sensitive applications. It, therefore, targets PCI pre-certification for Point of Sales, among other use cases.

From secure provisioning to over-the-air updates, developers can create mission-critical applications while satisfying the more stringent demands of governments and customers.



Ecosystem

We are already announcing that ST will have several demos at the upcoming Embedded World in Germany in April 2024, showcasing the evaluation board that embeds the STM32MP25 and its STPMIC25 power management companion chip. Moreover, we will demo ST software, such as OpenSTLinux, as well as software expansion packages. We will also showcase some STM32MP25 System-in-Package and System-on-Module from partners. Put simply, we are ensuring that the new STM32MP2 rapidly becomes a reality for more developers, wherever they are in the world.

The STM32MP25 evaluation board with a display and camera module.

STM32MP23 AND STM32MP21

We will release more information about the STM32MP23 and STM32MP21 as we near their mass-market release in 2024 and 2025, respectively. Still, we are thrilled to share their configuration to help our community plan for them and adopt the STM32MP2 series more rapidly and efficiently.

STM32MP23

The STM32MP23 will feature two Cortex-A35 at up to 1.5 GHz and one Cortex-M33 at 400 MHz. Given its position as a more cost-effective solution that would process significantly less data than the STM32MP25, it has a 16-bit DDR4/LPDDR4/DDR3L memory controller and two ethernet controllers supporting GMAC. Similarly, the video accelerator only features an H.264 decoder since it will not need to encode data, and the future MPU also features a smaller GPU while still supporting DSI, LVDS, and RGB displays.

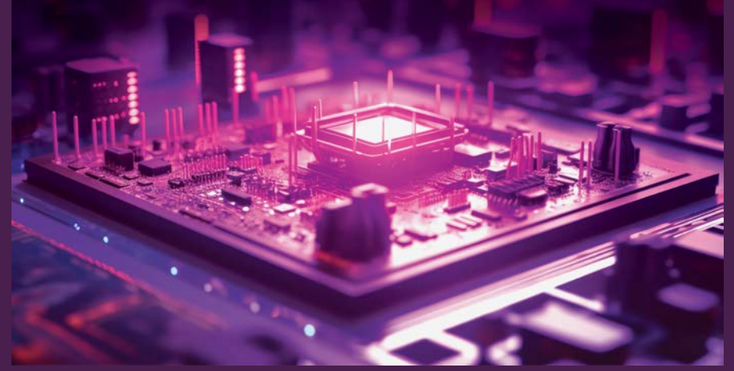
STM32MP21

The STM32MP21 has one Cortex-A35 at 1.5 GHz, one Cortex-M33 at 300 MHz, a slower 16-bit memory controller than the STM32MP23, and the same Ethernet controllers. Similarly, as the device targets very different graphical needs, it offers the display parallel interface found on the other two STM32MP2s, omits the GPU, but still supports parallel and CSI interfaces for cameras to ensure developers can easily interface with various sensors.

CONCLUSION

CLICK HERE

The whole series can work together to create an ecosystem of products. Let's take a medical imaging analysis application. The camera interface and processing capabilities of the STM32MP21 mean that it could apply various filters to clean the data received before sending it to an STM32MP25, which would run the neural network. Developers can thus increase the overall accuracy of their system without adding another workload on the STM32MP25, which would be able to run inferences, show results on its display, and securely send relevant data to the cloud. And since several packages in the STM32MP2 series are pin-to-pin compatible, developers could reuse many of the same design cues and code to reduce their time to market. For more information, visit our website.



HOW TO INCREASE **EMBEDDED AI PROCESSING** FOR AUTONOMOUS SYSTEMS?

RENESAS

VISIT RENESAS

Advanced artificial intelligence (AI) processing, such as recognition of the surrounding environment, decision of actions, and motion control, is required in various aspects of society, including factories, logistics, medical care, service robots operating in the city, and security cameras.

Systems need to handle advanced AI processing in real time and the system must be embedded within the device to enable a quick response to its constantly changing environment. AI chips at the same time consuming less power while performing advanced AI processing in embedded devices with strict limitations on heat generation.

To meet these market needs, Renesas developed DRP-AI (Dynamically Reconfigurable Processor for AI) as an AI accelerator for high-speed AI inference processing combining low power and flexibility required by the edge devices. This reconfigurable AI processor technology is embedded in the RZ/V series of MPUs targeted at AI applications.

The next-generation of the DRP-AI – DRP-AI3 – achieve power efficiency approximately 10 times higher than that of the previous generation to support further evolution of AI and the sophisticated requirements of robotics and automation applications.

DRP-AI3 ACCELERATOR FEATURES – HIGH-SPEED, LOW-POWER HARDWARE FEATURES BASED ON THE PRUNING AI MODEL

DRP-AI3 combines both the hardware and software to deliver the heterogeneous architecture for an AI-MPU. DRP-AI3 is a hardware architecture supporting the main model compression technology of bit count reduction (INT8) and pruning technology.

The flexibility of DRP-AI3 allows faster random pruning models, which is difficult to achieve with existing hardware.

Processing time can be reduced to as little as 1/16 and power consumption to about 1/8 compared to before pruning was applied.

DRP-AI3 introduces high-speed and low-power methods that support following major AI model compression methods:

- Quantization: Lower bit weights for neural network weight information (weight) and input/output data (feature map) for each layer. Change from 16-bit floating-point arithmetic in conventional DRP-AI to 8-bit integer arithmetic (INT8).
- Pruning: A technique to skip calculations by setting weight information (branches) that do not affect recognition accuracy to zero.
- Ideally, quantization is expected to yield more than around 2 times less power than conventional DRP-AI (16-bit processing), since the size of the arithmetic unit and the amount of data access are lighter in relation to the number of bits.

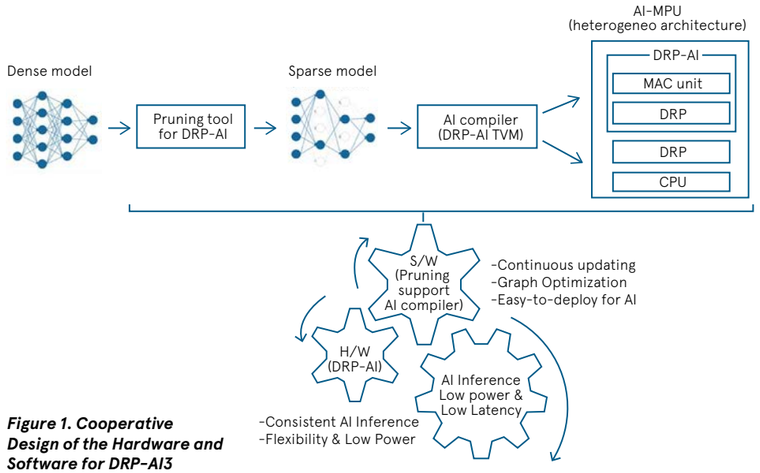


Figure 1. Cooperative Design of the Hardware and Software for DRP-AI3

- In addition, pruning depends on the AI model as to how much weight information can be retained, but if, for example, 90% pruning can be achieved, the expected value will be about 10 times higher speed and lower power consumption.

A major challenge with the current AI hardware is that it cannot efficiently process AI models, especially (2) pruned AI models.AI hardware is generally based on the SIMD (Single Instruction Multiple Data) architecture, which performs many simultaneous multiply-accumulate (MAC) operations to efficiently process large neural networks.

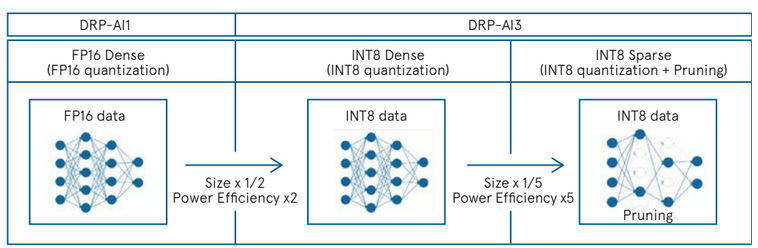


Figure 2: Model Compression Technology Applied to DRP-AI3

Since the locations of weights do not affect recognition accuracy are randomly located in the matrix, even if some of the weights become zero inside the parallel MAC operation, the parallel computation is still performed, together with non-zero weights. Therefore, not reducing the number of computations by pruning branches (Figure 3).

Structured pruning, where values are set to zero (e.g., zero for each column of a weight matrix) without compromising parallelism, is a well-known pruning method used in AI hardware. However, this method cannot achieve a high pruning rate because the conditions are significantly different for weight information that is inherently random. Therefore, Renesas developed the flexible N:M pruning method that has the flexibility to skip operations even in more random pruning.

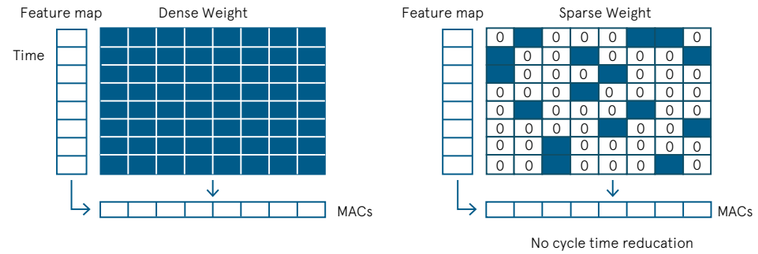


Figure 3: Pruned Model Processing with General Parallel Architecture

As shown in Figure 4, this technology is to divide the original weight matrix into weight matrix groups of M rows, reconstruct them into smaller N-row weight matrix groups, from which only significant weights are extracted in each group. Parallel operations are then performed on the new weight matrix groups.

In this process, DRP-AI3 has a new function allowing the number of operation cycles to be adjusted freely switching the value of N for each weight matrix group, making it possible to perform optimal skipping of operation processing for local varying pruning rates in the actual AI model, as shown in Figure 4.

This ability to finely vary N also allows the pruning rate of the entire weight matrix to be set in detail, enabling optimal pruning processing according to the user's required power consumption, operating speed, and recognition accuracy requirements.

This technology reduces the weight data size and the number of processing cycles for AI models by at least 1/10 and 1/16 respectively, resulting in a significant improvement in processing efficiency compared to conventional AI accelerator configurations (Figure 5).

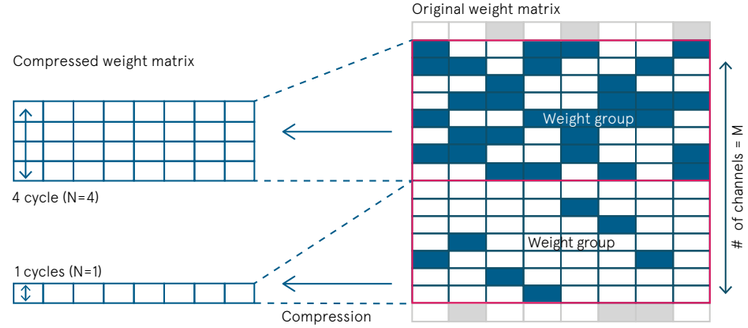


Figure 4: Compression of a Pruned Model Using DRP-AI3

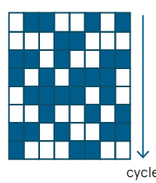
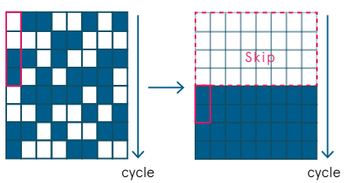
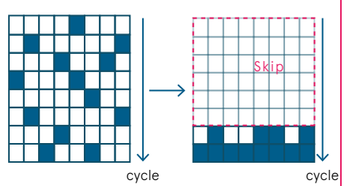
	General AI accelerators		50% pruning arch. (ref[1])	This works (DRP-AI)
				
Pruning structure	structured	unstructured	unstructured	unstructured
Pruning rate	0~30%	0~90%	0/50%	0~93%
weight data size	~2/3x	~1/10x	~5/8x	~1/10x
Cycle time	~2/3x	1x	1/2x	~1/16x

Figure 5: Comparison of Pruned Model Processing Performance by Accelerator

SOFTWARE FEATURES FOR GENERATING AND IMPLEMENTING PRUNED MODELS

A pruning flow is generally applied, as shown in Figure 6 to improve the pruning rate while suppressing the degradation of recognition accuracy.

Generally, after the initial training, the pruning points are selected to have the least impact on recognition accuracy. Renesas developed a pruning tool (DRP-AI Extension Pack) that selects pruning points to satisfy the aforementioned DRP-AI3 pruning hardware's architecture constraints. Users can apply DRP-AI3's characteristic "flexible N:M pruning" by simply specifying the pruning rate.

To further ease the introduction of pruning, the above tools are provided based on OSS AI frameworks (Pytorch, Tensorflow), which enables pruning and retraining (in Figure 6) by simply adding a few lines to the user's existing training scripts.

In addition, pruned AI models generated by the pruning tool can be converted by DRP-AI TVM for simultaneous INT8 quantization and compilation.

Here, DRP-AI TVM is a tool to convert trained AI models into a format that is executable on Renesas AI MPUs. It is based on Apache TVM, an OSS ML compiler framework, and is capable of allocating operations in each layer of the AI model that can be executed by the DRP-AI and those operations that cannot be executed by DRP-AI to the CPU for processing. This type of computing using multiple processors together is called heterogeneous computing, and can greatly expand the number of AI models to be executed.

Renesas provides adequate software environments to minimize the time and effort of users to introduce pruning while maximizing the DRP-AI hardware architecture through hardware-software co-design, and improve pruning efficiency.

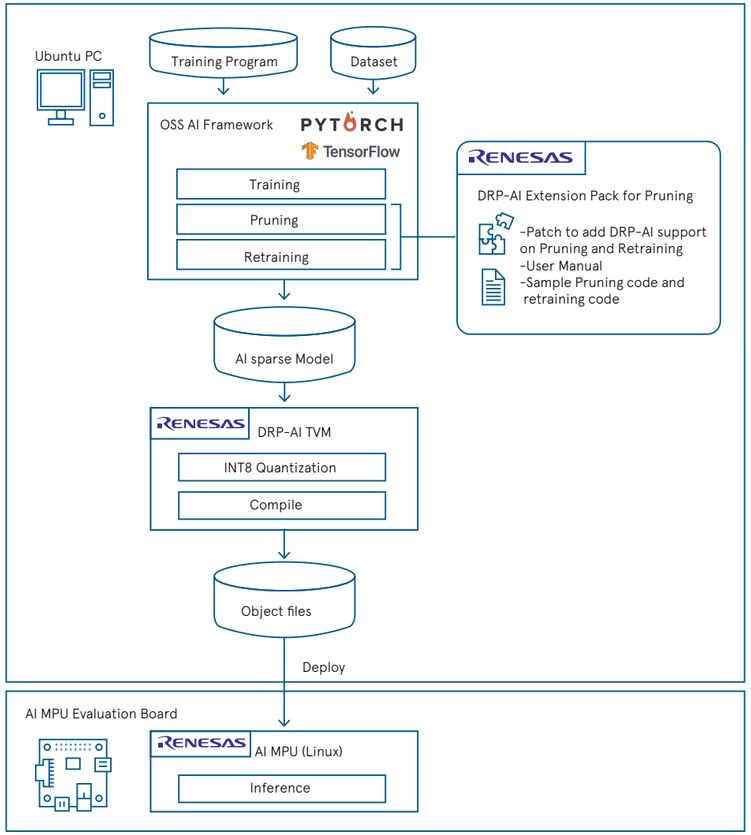
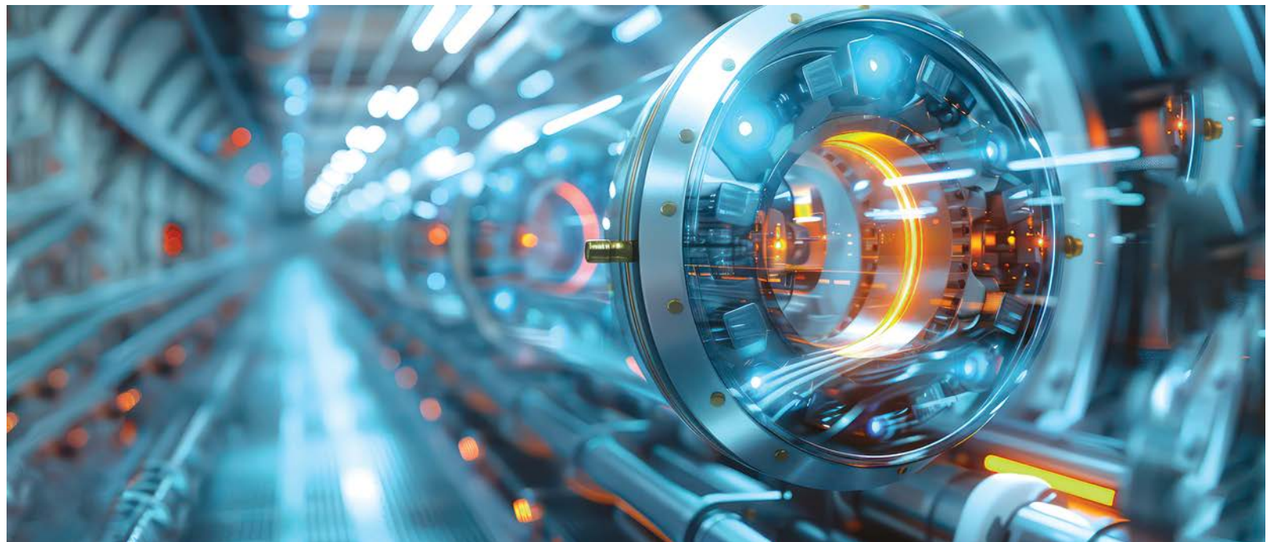
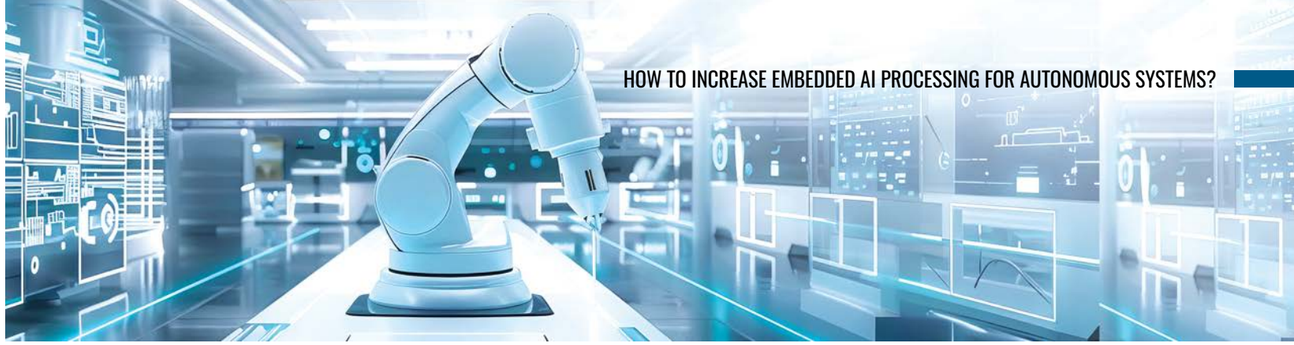


Figure 6: DRP-AI3 AI Model Compression and Implementation Flow





DRP-AI3 ACCELERATOR FEATURES –HETEROGENEOUS ARCHITECTURE FEATURES IN WHICH DRP-AI, DRP, AND CPU OPERATE COOPERATIVELY

Service robots, for example, require advanced AI processing to recognize the surrounding environment. On the other hand, non-AI algorithm-based processing is also required for deciding and controlling the robot's behavior. However, current embedded CPUs lack sufficient resources to perform these various types of processing in real time. Renesas solved this problem by developing a heterogeneous architecture technology that enables the dynamically reconfigurable processor (DRP), AI accelerator (DRP-AI), and CPU to work together.

As shown in Figure 7, DRP can execute applications while dynamically switching the circuit connection configuration of the arithmetic units on the chip at each operating clock according to the content to be processed. Since only the necessary arithmetic circuits are used, the DRP consumes less power than with CPU processing and can achieve higher speed. Furthermore, compared to CPUs, where frequent external memory accesses due to cache misses and other causes will degrade performance, the DRP can build the necessary data paths in hardware ahead of time, resulting in less variation in operating speed (jitter) due to memory accesses.The DRP also has a dynamic loading function that switches the circuit connection information each time the algorithm changes, enabling processing with limited hardware resources, even in robotic applications that require processing of multiple algorithms.

The DRP is particularly effective in processing streaming data such as image recognition, where parallelization and pipelining directly improve performance. On the other hand, CPU software processing may be more suitable for programs such as robot behavior decision and control require processing while changing conditions and processing details in response to changes in the surrounding environment. Renesas' heterogeneous architecture technology allows the DRP and CPU to work together by distributing the processing to the right places and to operate in a coordinated manner.

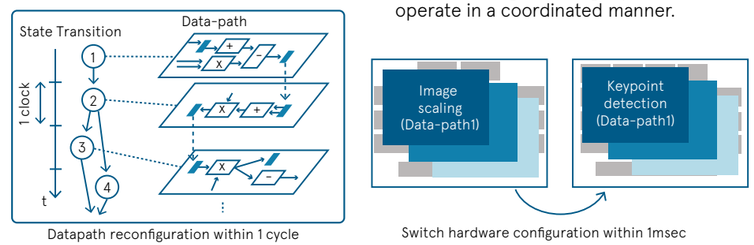


Figure 7: Flexible Dynamically Reconfigurable Processor (DRP) Features

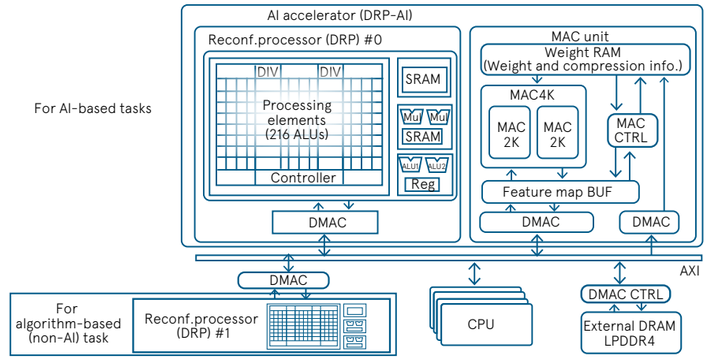


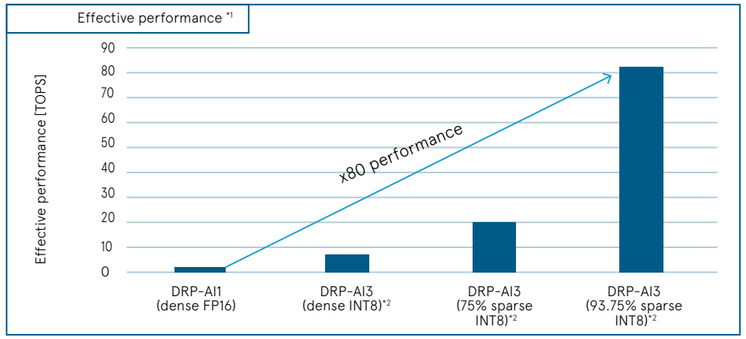
Figure 8: DRP-AI3-based Heterogeneous Architecture Configuration

EVALUATION RESULTS

A prototype test chip achieved a maximum of 8 TOPS (8 trillion operations per second) for the processing performance of the AI accelerator.

By reducing the number of operation cycles in proportion to the amount of pruning, we could achieve AI model processing performance equivalent to a maximum of 80 TOPS when compared to models before pruning (Note 1).

This is about 80 times higher than the processing performance of the conventional DRP-AI, a significant performance improvement to keep pace with the rapid evolution of AI (Figure 9).



Measured data on board.
*1) Effective performance evaluating test layer (Single 3x3 conv.), Batchsize=1
*2) Performance may change depends on AI models.

Figure 9: Comparison of Measured Peak Performance of DRP-AI

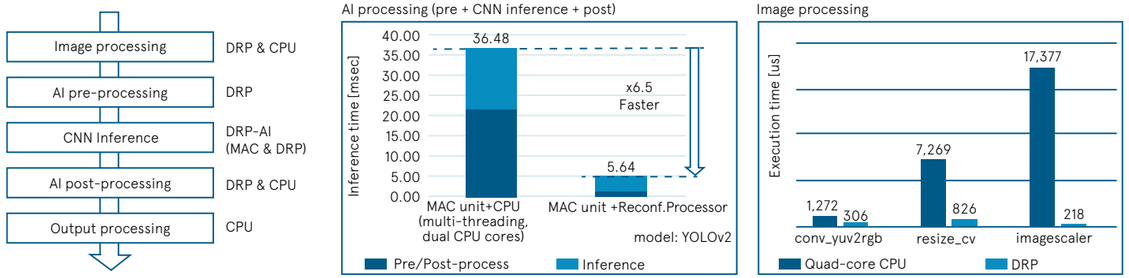


Figure 10: Heterogeneous Architecture Speeds Up Image Recognition Processing

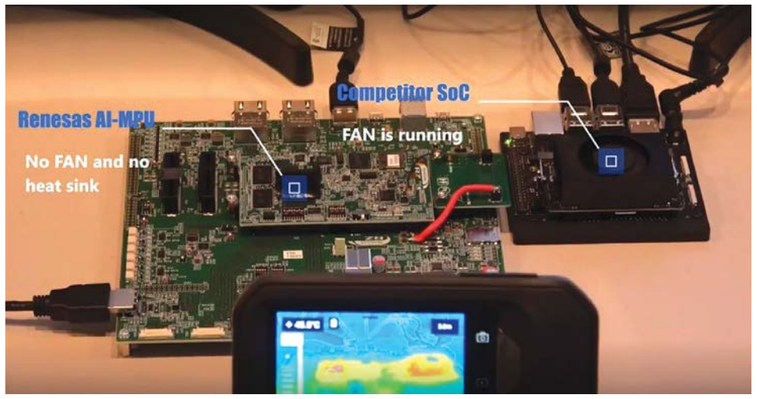


Figure 12: Comparison of Heat Generation between a Fanless DRP-AI Test Board and a GPU with Fan

As AI processing speeds up, the processing time for non-AI image processing is becoming a relative bottleneck.

In AI-MPUs, a portion of the image processing program is offloaded to the DRP, thereby contributing to the improvement of the overall system processing time (Figure 10).

The same AI real-time processing could be performed on an evaluation board equipped with the prototype chip, without a fan comparable to existing market products equipped with fans (Figure 12).

CONCLUSION

Renesas DRP-AI3 – an advanced version of DRP-AI (Dynamically Reconfigurable Processor for AI) – is a unique AI accelerator combining the low power and flexibility required by endpoints, with processing capabilities for lightweight AI models, and 10 times more power efficient (10 TOPS/W) than the previous models.

Visit www.renesas.com/rzv2h to learn more about the device and Renesas DRP-AI.

ABOUT THE AUTHOR

- **Takao Toi**
- **Masayuki Shimobeppu**
- **Kentaro Mikami**
- **Koichi Nose**

Senior Principal Product Engineer, Embedded Processing Product Group, Embedded Processing 1st Business Division, Embedded Processor Product Management Department, Renesas Electronics Corporation

REED RELAYS

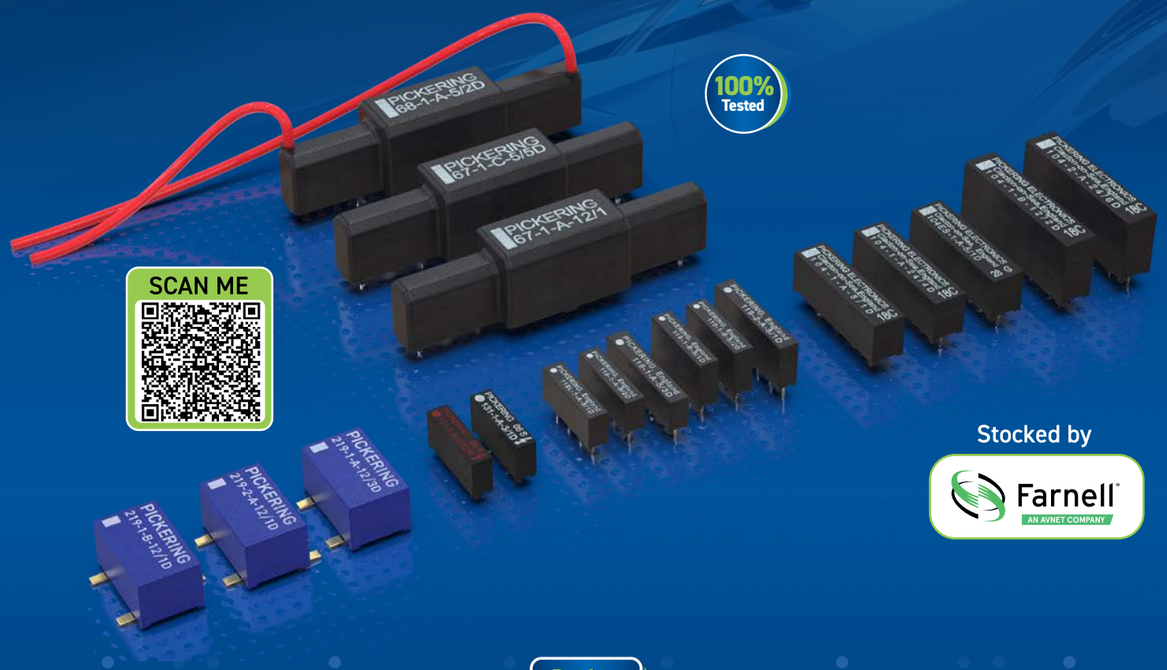
ACCELERATE YOUR DESIGN

for **EV** & **Charge Point Testing**

With very low leakage currents (< 1 nA) & very high standoff voltages achievable in a small package, reed relays are the logical choice for EV testing applications where safety & reliability is a priority.

- From **1 kV** to **7.5 kV** switching
- Minimum standoff between **1.5 kV** to **10 kV**
- From **10 W** to **200 W** switching power
- Optional electrostatic screen
- Typically, 10^9 operations life expectancy
- Available in through-hole **SIL** and **SMD** packages
- Many build options available to suit a variety of applications

For a free working sample go to: pickeringrelay.com/samples



100% Tested

SCAN ME

Stocked by Farnell AN AVNET COMPANY

Product 25+Years Longevity

pickeringrelay.com
+ (44) 1255 428141 | sales@pickeringrelay.com

pickering

OMRON

D2EW

Sealed Ultra Subminiature Switches for leverless & multi-angle operation

- **Super compact**
- **Ultra-durable**
- **Highly versatile**
- **Quiet operation**

Farnell
AN AVNET COMPANY

ACCESS
ASSEMBLE
ASPIRE

The right components
for tomorrow's success.