Vision AI – Applications Across Hardware Platforms

Mitsuo Baba, Senior Director, Embedded Processing Product Group, Renesas Electronics Corporation

Introduction

The power of AI technology is bringing dramatic changes to all computing systems. Being AI-ready has become necessary not only for data centers and servers in the Cloud, but also for embedded systems at the edge and endpoint. As this evolution continues to accelerate, two major trends have emerged: Generative (or Gen) AI and TinyML. These trends will drive scalable growth in embedded applications for all microcontrollers and microprocessors. This white paper describes the expanding embedded AI applications and the solutions that can help businesses and developers take advantage of scalable growth.

What is embedded AI?

As has been widely reported, investment in embedded AI is booming. But to understand what is driving this growth, we need to take a closer look at the features of embedded AI.

Embedded AI is a general term for AI that runs on embedded systems, as opposed to AI that runs on servers in a data center. So, the computing location of embedded AI is not the Cloud, but the edge and endpoints. Therefore, while AI in the Cloud is executed on a centralized computer such as large-scale

servers, AI at the Edge and Endpoints is executed on distributed computers into a wide variety of embedded systems and on a wide variety of computers.

Trends in embedded AI

Current embedded AI applications can be classified into two functional categories – all new devices or enhancements to existing devices.

AI functions on new devices – AI cameras improve recognition capabilities, automate tasks previously performed by humans and increase efficiency as seen in roaming robots, surveillance cameras or quality inspection in factories. These markets themselves have been created and are expanding significantly with the implementation of AI. This type mainly focuses on vision AI use cases.

AI functions on existing devices – Anomaly detection in motor peripherals, classifying and judging human machine interface or interaction (HMI) and replacing functions previously achieved using classical algorithms. This type dramatically expands the scope of AI applications to solve previously difficult issues, such as the effects of differences in individual external environments and installation and adjustment work.

With these two types of expansions, AI applications are for all embedded systems, whether driven by MCUs or MPUs, and all input information from vision, voice and other real-time sensors, as seen in Figure 1.

Figure 1: AI expansion into all embedded computing systems

Especially for vision AI, there are two key trends emerging that will further expand the embedded AI market – Gen AI transformer and TinyML. Vision AI uses vision data from camera devices and has expanded with the development of high-performance AI models, as shown in Figure 2.

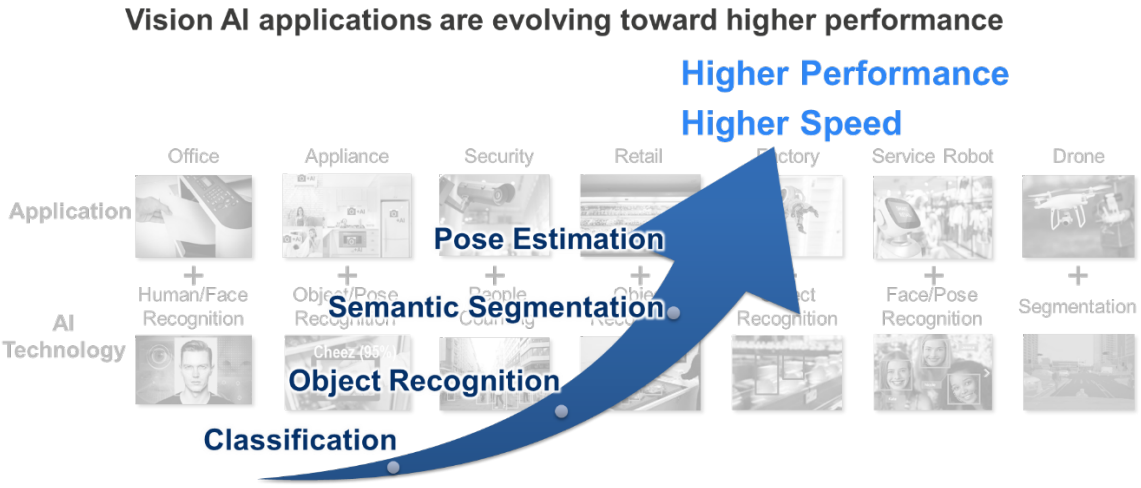

Figure 2: Performance evolution of vision AI applications

A discontinuous change is occurring in this high-performance expansion, a trend caused by the transformer model driven by Gen AI not only in the Cloud but also at the edge and endpoints. A transformer model is a neural network that learns the context of sequential data and generates new data out of it. A key benefit of this model with self-attention layers is improved accuracy, reducing recognition errors by half compared to conventional convolutional neural network (CNN) models. Accuracy is a parameter, particularly for industrial systems, that significantly affects productivity and production costs themselves so it is expected that this will provide new opportunities to further expand the scope of applications.

Previous AI trends started with the Cloud and then implemented at the edge and endpoints, but a new trend has emerged that starts with embedded AI, TinyML. Edge and endpoint computing resources are limited compared to those of the Cloud therefore, the implementation of AI models for vision AI has also been limited to relatively high-performance MPU-based products. The TinyML trend provides new vision AI models that are “tiny” in terms of both computational and memory resources. This is expected to expand vision AI applications not only to the high end of the market but also the low end.

Figure 3 shows AI model plots with the horizontal axis representing the number of operations per second (MFLOPS) as computational resources and the vertical axis representing the number of parameters (M Parameters) as memory resources. Existing AI models shown in green are still

expanding and the trend driven by GenAI transformer models will rapidly expand further towards the high end, while the trend driven by TinyML will create a new market at the low end.

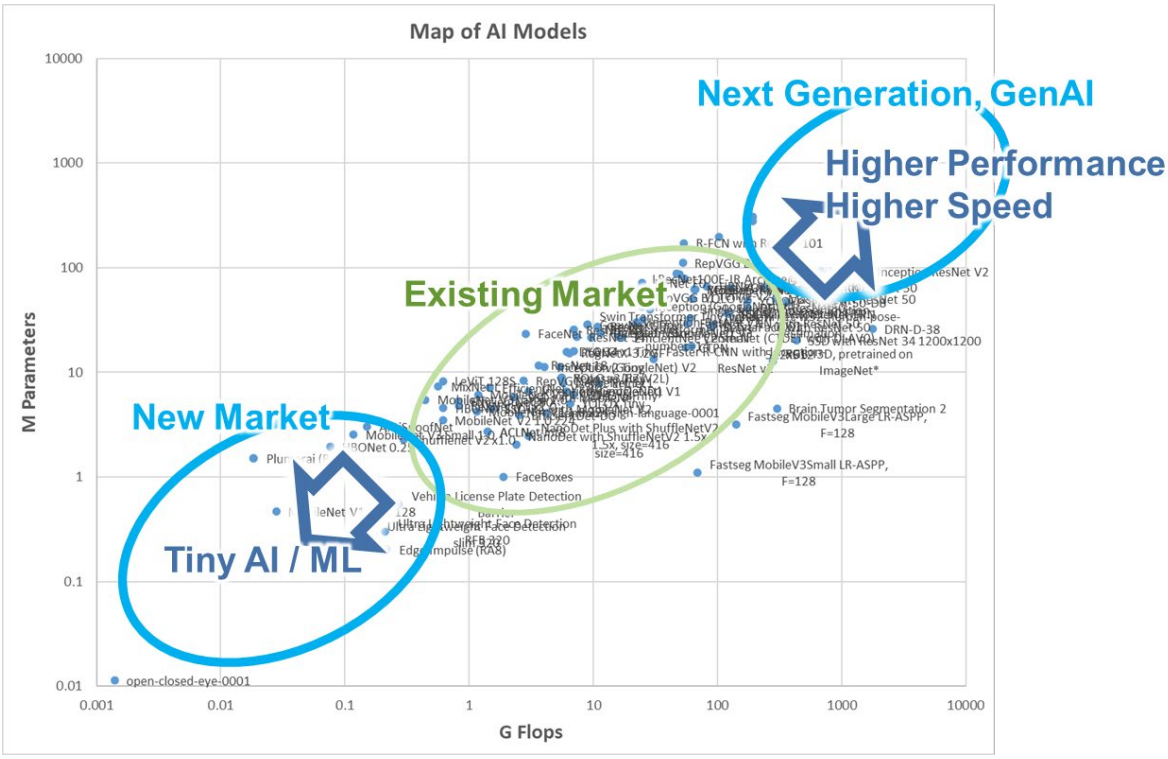

Figure 3: Map of AI models expand applications at both the high end and low end

The demand for AI implementation in embedded systems is sure to grow exponentially with the growth of these two trends. In the near future, it may be possible that Tiny Gen AI, Tiny Transformer and Tiny LLM will emerge in the embedded AI field. Next let's look at the necessary requirements for leveraging these trends to drive business growth.

Requirements for embedded AI systems

There are three essential requirements for leveraging these emerging trends to grow your business:

- Scalable product lineup for AI performance
- Scalable easy-to-use AI development environment addressing all AI journeys
- Differentiated products that can implement new AI trends economically

The development of embedded AI is just software-oriented and it is important to manage software development costs. This includes updating AI software in the field as lifecycle management, which is necessary for value creation and business success. From this viewpoint, the combination of scalable products and scalable AI development environments is necessary. Furthermore, to capture new

trends, differentiation based on performance as well as cost will become a necessary condition. Especially with the trend towards higher performance, power efficiency will bring both cost and performance differentiation. Finally, ready access is required to new products that are differentiated in both cost and performance to be leveraged across a range of applications.

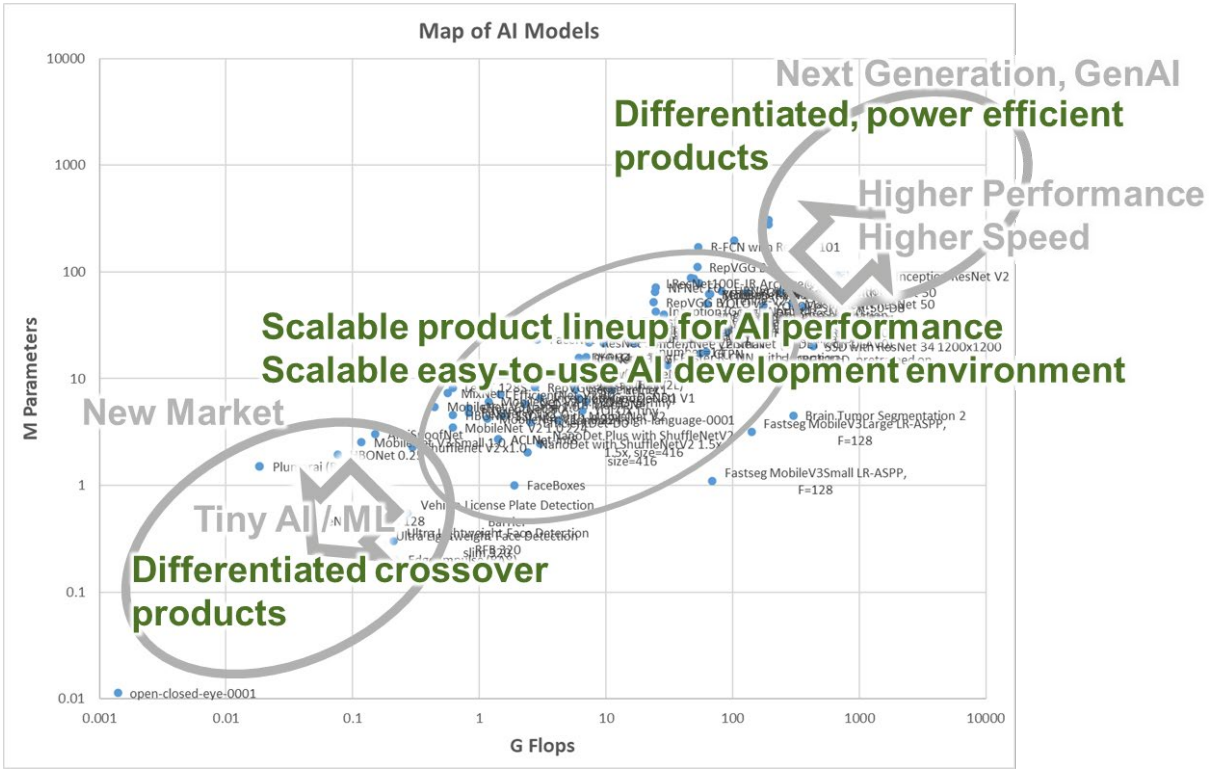

Figure 4: Requirements to leverage emerging AI growth opportunities

Scalable embedded AI solutions from Renesas

Renesas has a long history of embedded processing innovation to leverage in providing embedded AI solutions to meet the requirements of these emerging AI systems. Figure 5 shows the scalable product lineup of MCUs and MPUs capable of addressing everything from real-time analytics to voice and vision AI. In particular, the two newest products, RA8 MCUs and RZ/V MPUs, are offered as differentiated solutions towards two trends in the vision AI market.

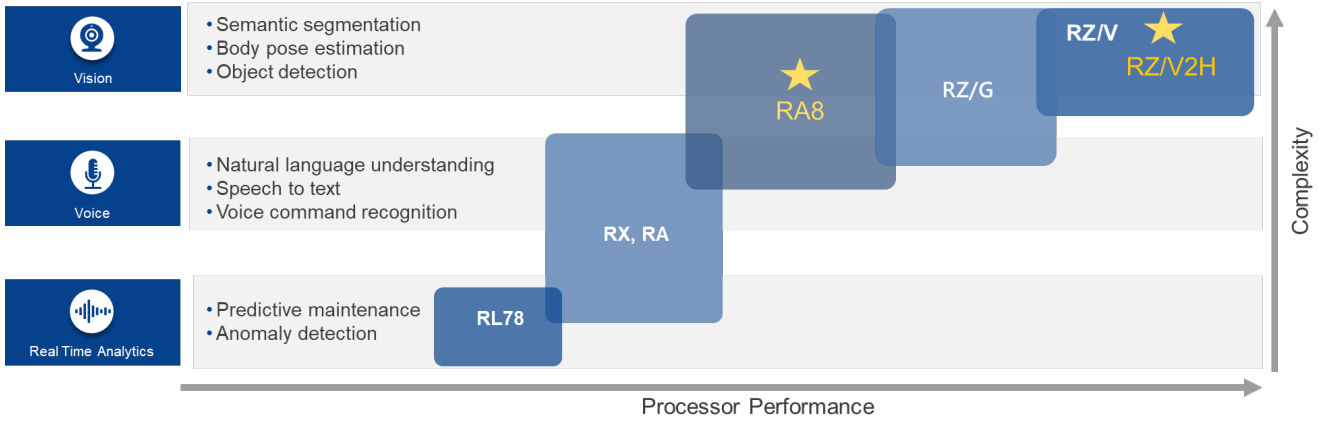

Figure 5: Scalable embedded AI hardware platforms

RZ/V2H MPUs with the Renesas-developed DRP-AI3 AI accelerator, provides approximately 10 times the power efficiency of conventional DRP-AI products. It is capable of solving the heat generation problems that are the most difficult challenge to address in high-performance applications, while providing both competitive performance and cost (Figure 6).

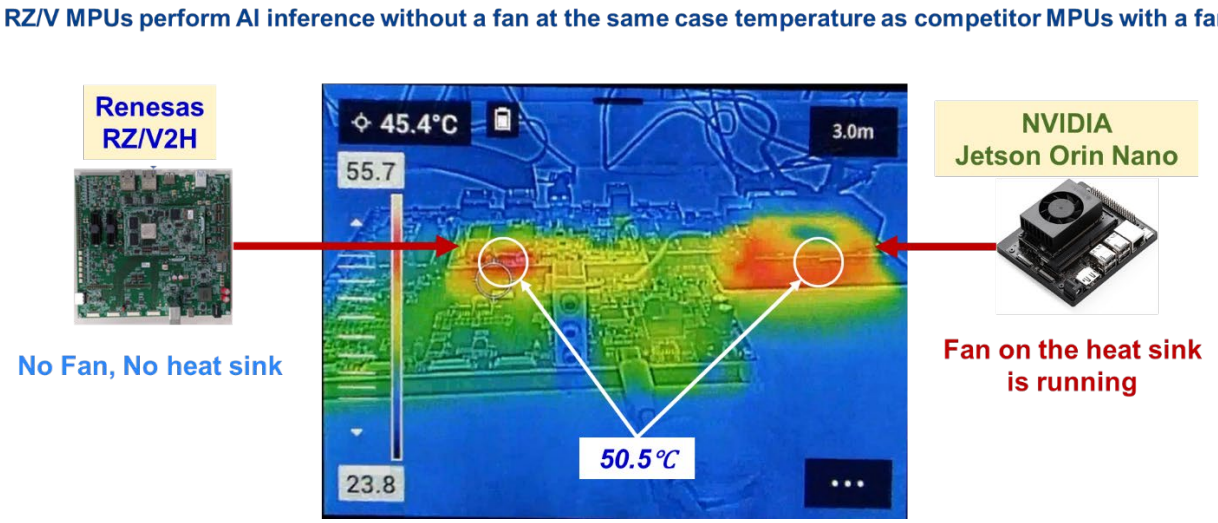

Figure 6: Comparison of RZ/V2H running the same AI inference without a fan as other MPU with a fan

RA8D1 MCUs are part of the high-performance RA8 MCU family, the first processors on the market to be equipped with Arm's CM85 core and Helium technology. These devices provide performance that far exceeds previous microcontrollers and enables cost-efficient Tiny Vision AI implementation. Figure 7 shows example results from the MLPerf Inference Tiny Benchmark Suite, which provides standard comparison.

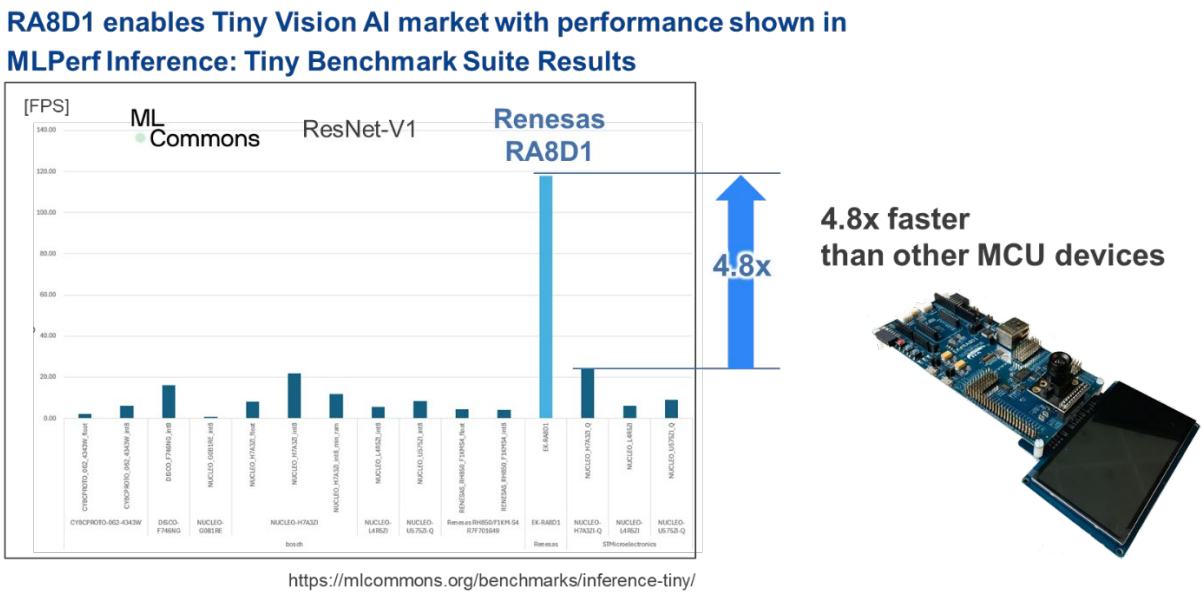

Figure 7: RA8D1 based on the Arm® Cortex®-M85 core accelerate neural network processing and outperform other MCUs

In terms of the AI development environment, Renesas provides AI tools that cover all customer journeys, from AI beginners to experts, addressing AI modeling to deployment on devices (Figure 8).

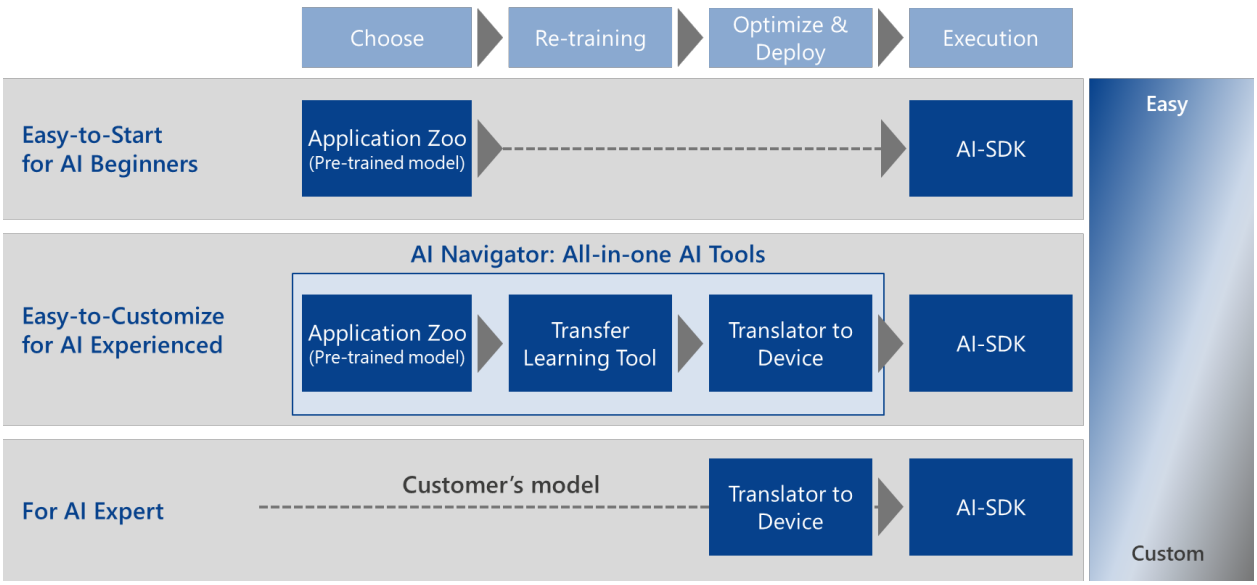

Figure 8: Renesas AI tools address all phases of the development environment for a range of AI backgrounds and experience

Renesas will continue contributing to the growth of the embedded AI system market with scalable AI solutions centered on hardware platforms, including differentiated new products (Figure 9).

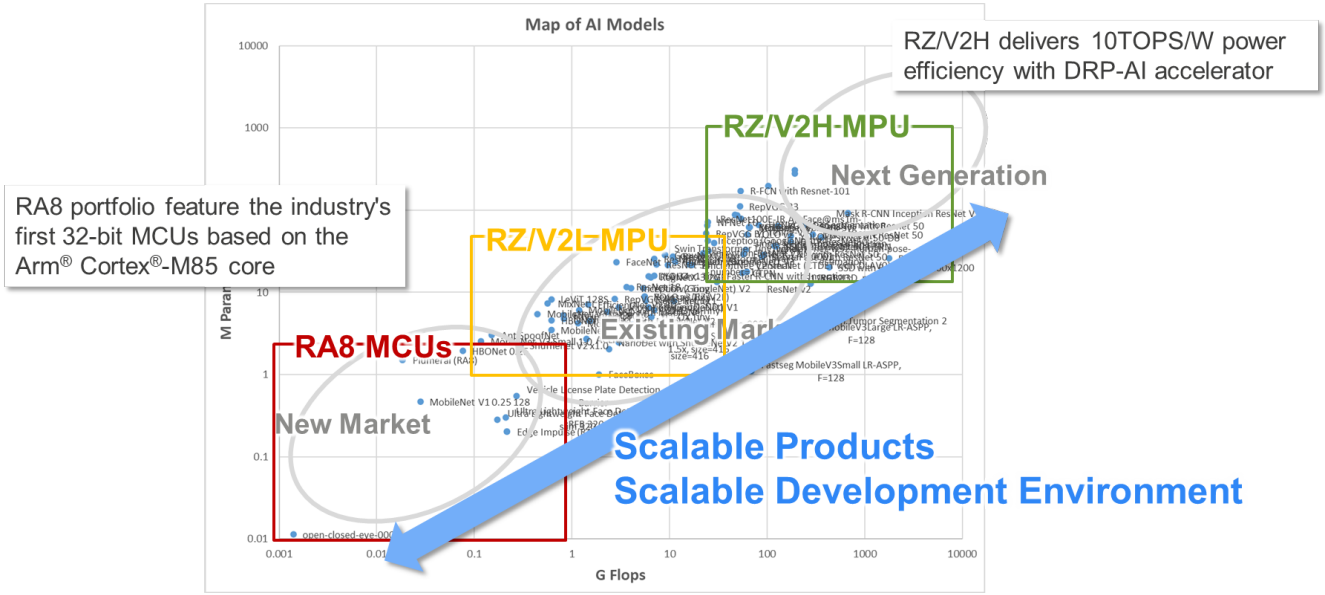

Figure 9: Renesas provides scalable products and development environments for embedded AI systems

Conclusion

The continued evolution of embedded AI requires both scalable hardware devices and software development environments for market growth. Because the development costs of embedded processor software are already higher, AI application software is the largest component, including DevOps. AI solutions from Renesas will continue to evolve and provide a software designer-oriented development environment that cover all MCUs and MPUs to simplify the process. To learn more about AI technology, as well as voice, vision and real-time analytics solutions from Renesas, visit renesas.com/AI.

Corporate Headquarters
TOYOSU FORESIA, 3-2-24 Toyosu, Koto-ku, Tokyo 135-0061, Japan
https://www.renesas.com

Contact Information
For further information on a product, technology, the most up-to-date version of a document, or your nearest sales office, please visit:
https://www.renesas.com/contact-us

Doc Number: R01WP0028EU0100