Next Generation Highly Power-Efficient AI Accelerator (DRP-AI3):
10x Faster Embedded Processing in Advanced AI for Autonomous Systems

Takao Toi, Senior Manager, Product Management, Embedded Processing Product Group, Embedded Processing 1st Business Div, Embedded Processor Product Management Dept Section 2, Renesas Electronics Corporation

Masayuki Shimobeppu, Senior Manager, Product Management, Embedded Processing Product Group, Embedded Processing 1st Business Div, Embedded Processor Product Management Dept Section 3, Renesas Electronics Corporation

Kentaro Mikami, Principal Software Engineer, Software& Digitalization Group, Software Development Division, System Solution Department 1, AI Solution Section, Renesas Electronics Corporation

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

Overview

As the working population decreases due to falling birthrates and a growing proportion of the population being elderly, advanced artificial intelligence (AI) processing, such as recognition of the surrounding environment, decision of actions, and motion control, will be required in various aspects of society, including factories, logistics, medical care, service robots operating in the city, and security cameras. Systems will need to handle advanced artificial intelligence (AI) processing in real time in various types of programs. In particular, 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 accelerator processor technology, cultivated over many years, is embedded in the RZ/V series of MPUs targeted at AI applications. The DRP-AI3 is the next-generation of the DRP-AI, achieving power efficiency approximately 10 times higher than that of the previous generation. The DRP-AI3 is able to respond to the further evolution of AI and the sophisticated requirements of applications such as robots. This white paper introduces the key technologies developed for DRP-AI3 and demonstrates how the DRP-AI3 solves heat generation challenges, enables high real-time processing speed, and realizes higher performance and lower power consumption for AI-equipped products.

DRP-AI3 Accelerator Features

1. Hardware(H/W)-Software(S/W) coordination for AI model lightening (pruning) achieves about 10 times the power efficiency compared to conventional models.
   - AI accelerator (DRP-AI3 hardware) introducing high-speed and low-power technology for pruning models
   - Software to easily generate pruning models suitable for DRP-AI3 and optimally implement them in H/W
2. DRP-AI, DRP, and the CPU work together in a heterogeneous architecture to accelerate a variety of algorithms in addition to AI.

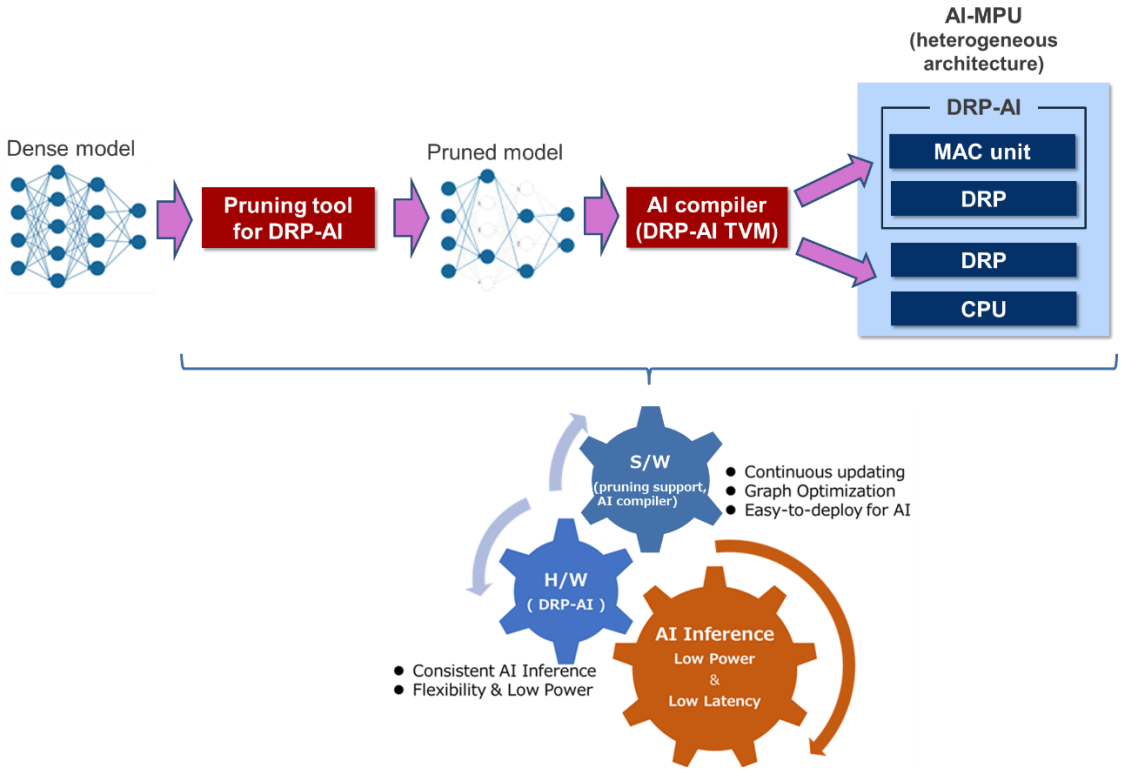

Figure 1. Cooperative Design of the Hardware and Software for DRP-AI3

High-Speed, Low-Power Hardware Features of the Pruning AI Model

- A hardware architecture that supports the main lightweight technology of bit count reduction (INT8) as well as pruning technology
- The flexibility of DRP-AI3 allows for 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.

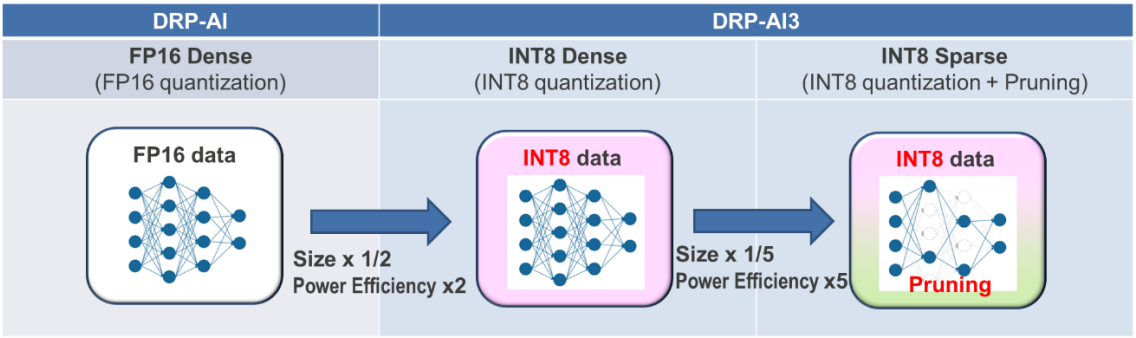

Figure 2: Lightweight Technology Applied to DRP-AI3

DRP-AI3 introduces high-speed and low-power methods that support major AI model lightening methods. Specifically, the following lightweighting methods are supported:

1) 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).
2) Pruning: A technique to skip calculations by setting weight information (branches) that do not affect recognition accuracy to zero.

(1) 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 a lighter weight in relation to the number of bits. (2) 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 a large number of simultaneous sum-of-products operations to efficiently process large sum-of-products matrix operations of neural networks. Since the locations of weights that do not affect recognition accuracy are randomly located in the matrix, even if some of the weights become zero inside the parallel sum-of-integration operation, the parallel computation is still performed for these, together with non-zero weights. Therefore, not reducing the number of computations by pruning branches (Figure 3).

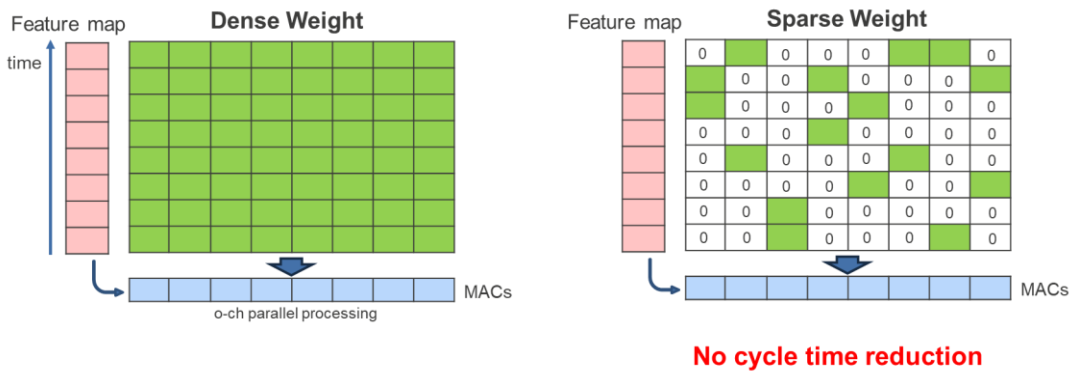

Figure 3: Lightweight (Pruned) Model Processing with General Parallel Architecture

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. Another method is to select and calculate one of two adjacent weights in the weight matrix(1). This method also reduces the amount of information in the weights by at most 1/2, so its effect on increased speed and lower power consumption is limited.

Therefore, Renesas developed the flexible N:M pruning method as a method that has the flexibility to skip operations even in more random pruning.

As shown in Figure 4, the basic concept of 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 that allows the number of operation cycles to be adjusted by 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.

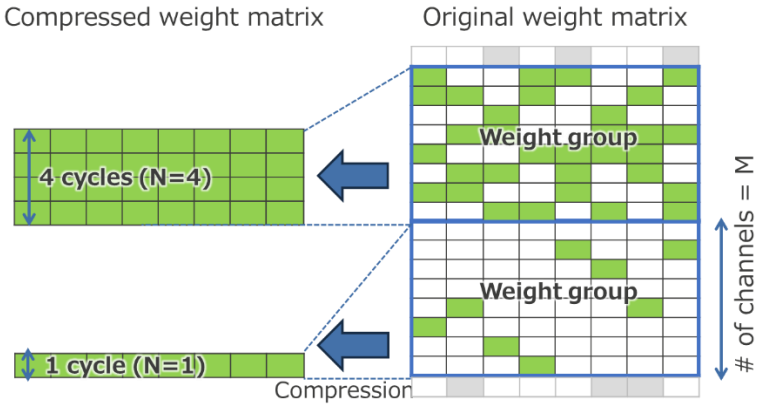

Figure 4: Compression of a Lightweight (Pruned) Model Using DRP-AI3

By analyzing the relationship between the pruning method and the recognition accuracy for actual AI models, we identified the optimal parameters that take into account the trade-off between accuracy, area and power increase, and reflected them in the hardware architecture.

This technology reduces the number of processing cycles for AI models by at least 1/16 while also reducing power consumption by at least 1/8, resulting in a significant improvement in processing efficiency compared to conventional AI accelerator configurations (Figure 5). This solves the problem of heat generation with conventional AI processors and enables implementation inside robots and small AI devices without requiring a cooling mechanism.

	General AI accelerators		50% pruning arch. (ref [1])	This work (DRP-AI)
	zero weight / non-zero weight		Skip	Skip
	Cycle		Cycle / Cycle	Cycle / Cycle
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

- Hardware/software co-design based on DRP-AI3 architecture features, considering where to prune to improve performance efficiently.
- Development of the DRP-AI Extension Pack (pruning tool)
- Development of the DRP-AI TVM (INT8 quantization tool and compiler)

As mentioned above, pruning is a method of skipping calculations by setting the weight information (branches) which do not affect recognition accuracy to zero. The more branches that are pruned in an AI model, the faster the performance and the lower the power consumption is to be expected, but the recognition accuracy tends to decrease.

A pruning flow is generally applied, as shown in Figure 6, in order to improve the pruning rate while suppressing the degradation of recognition accuracy. After the initial training, pruning is performed by selecting the pruning points, and retraining is performed using the weight information after pruning. Retraining is expected to have the effect of suppressing the degradation of recognition accuracy caused by pruning.

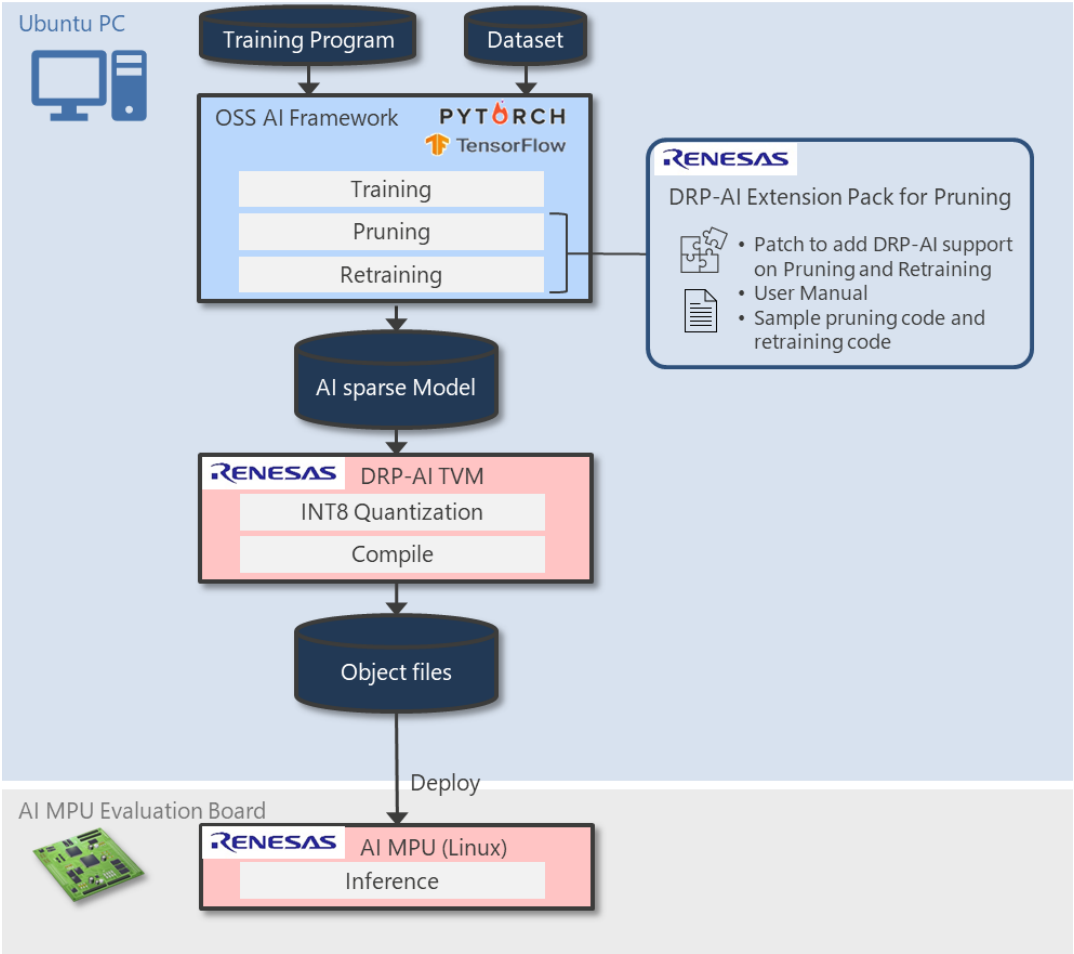

Figure 6: DRP-AI3 AI Model Lightening and Implementation Flow

Generally, after the initial training, the pruning points are selected to have the least impact on recognition accuracy, i.e., the points with small absolute values of weights or with small gradients during training. In addition to these factors, Renesas has 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 in the form of libraries in OSS AI frameworks (Pytorch, Tensorflow), which enables pruning and retraining (Pruning and Retraining in Figure 6) by simply adding a few lines to the user's existing training scripts.

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

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. DRP-AI is a powerful AI accelerator, but the number of operations it can perform is limited. By introducing a heterogeneous computing mechanism, the types of AI models that can be supported (AI model coverage) can be greatly expanded.

In this way, Renesas provides software environments such as the DRP-AI Extension Pack, which minimizes the time and effort required for users to introduce pruning while maximizing the DRP-AI hardware architecture through hardware-software co-design to improve pruning efficiency, as well as the DRP-AI TVM, which enables significant expansion of AI model coverage through heterogeneous computing to promote UX improvement.

Heterogeneous Architecture Features in which DRP-AI, DRP, and CPU Operate Cooperatively

- Multi-threaded and pipelined processing with AI accelerators, DRPs, and CPUs
- Low jitter and high-speed robot applications with DRP (dynamically reconfigurable wired logic hardware)

Service robots, for example, require advanced AI processing to recognize the surrounding environment. On the other hand, algorithm-based processing that does not use AI is also required for deciding and controlling the robot's behavior. However, current embedded processors (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, the dynamically reconfigurable processor (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 performance degradation and 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, programs such as robot behavior decision and control require processing while changing conditions and processing details in response to changes in the surrounding environment. CPU software processing may be more suitable for this than hardware processing such as in the DRP. It is important to distribute processing to the right places and to operate in a coordinated manner. Renesas' heterogeneous architecture technology allows the DRP and CPU to work together.

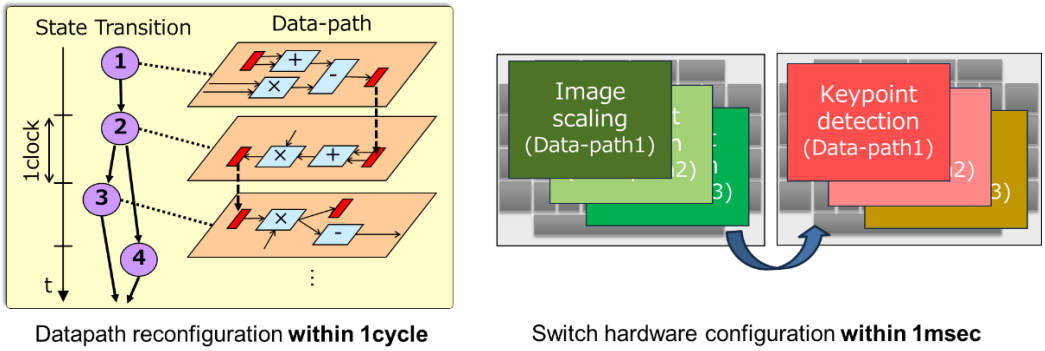

Figure 7: Flexible Dynamically Reconfigurable Processor (DRP) Features

An overview of the MPU and AI accelerator (DRP-AI) architecture is shown in Figure 8. Robotic applications use a sophisticated combination of AI-based image recognition and non-AI decision and control algorithms. Therefore, a configuration with a DRP for AI processing (DRP-AI) and a DRP for non-AI algorithms will significantly increase the throughput of the robotic application.

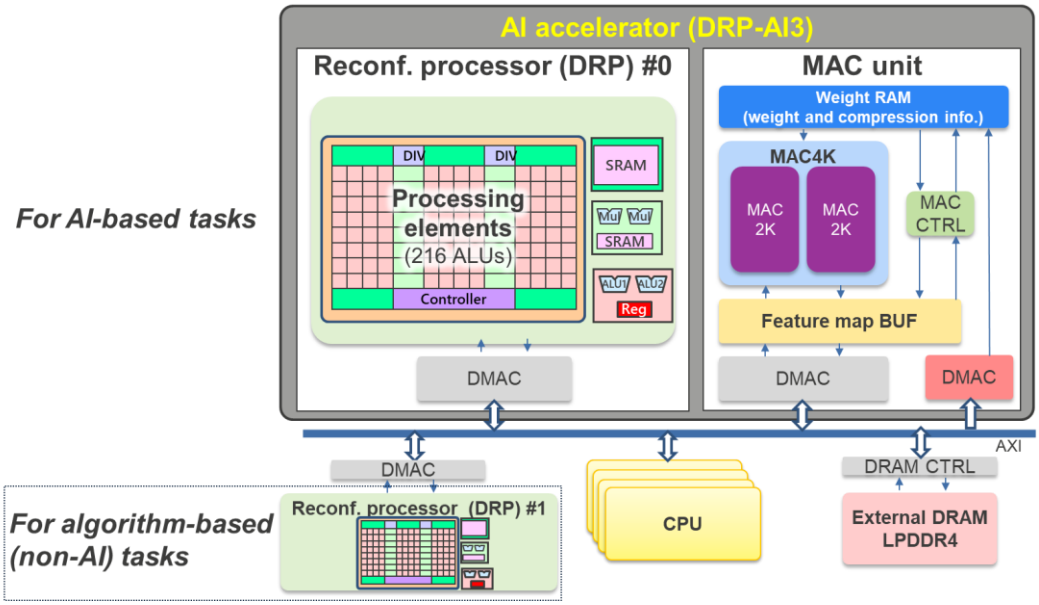

Figure 8: DRP-AI3-based Heterogeneous Architecture Configuration

Evaluation Results

(1) Evaluation of AI model processing performance

A prototype test chip equipped with this technology has achieved a maximum of 8 TOPS (8 trillion sum-of-products operations per second) for the processing performance of the AI accelerator. Furthermore, for AI models that have been pruned, the number of operation cycles can be reduced in proportion to the amount of pruning, thus achieving 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 that can sufficiently keep pace with the rapid evolution of AI (Figure 9).

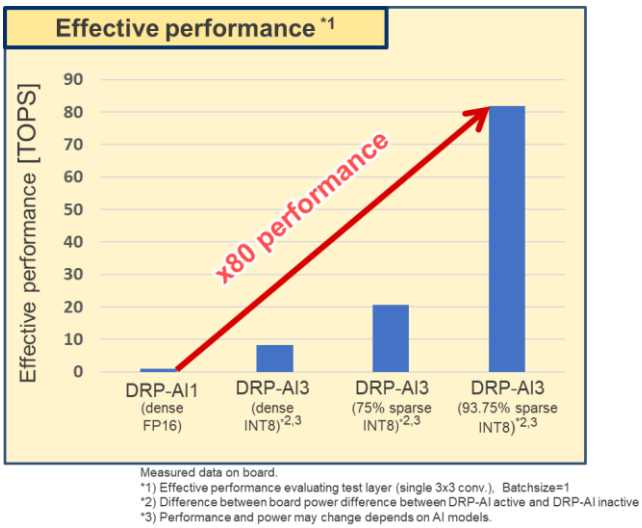

Figure 9: Comparison of Measured Peak Performance of DRP-AI

On the one hand, as AI processing speeds up, the processing time for algorithm-based image processing without AI, such as pre- and post-AI 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).

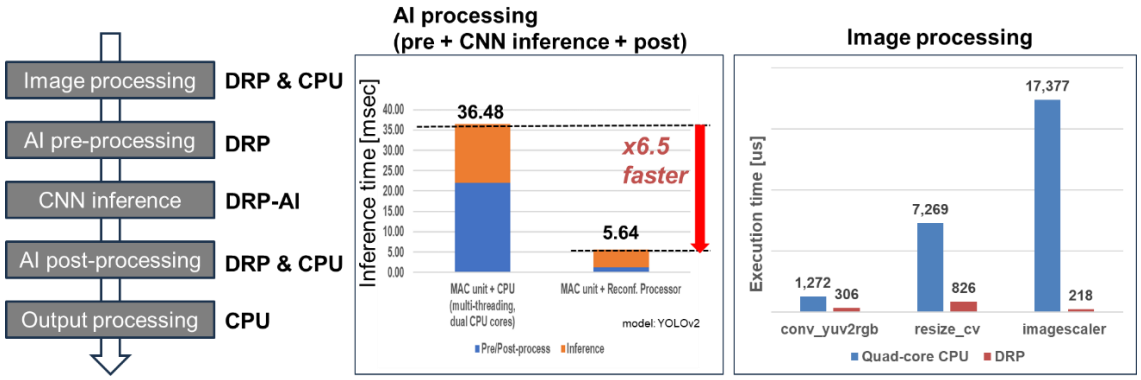

Figure 10: Heterogeneous Architecture Speeds Up Image Recognition Processing

In terms of power efficiency, the performance evaluation of the AI accelerator alone demonstrated a theoretical maximum performance of approximately 23 TOPS/W, and the world's top level power efficiency (approximately 10 TOPS per watt) when running major AI models (Figure 11).

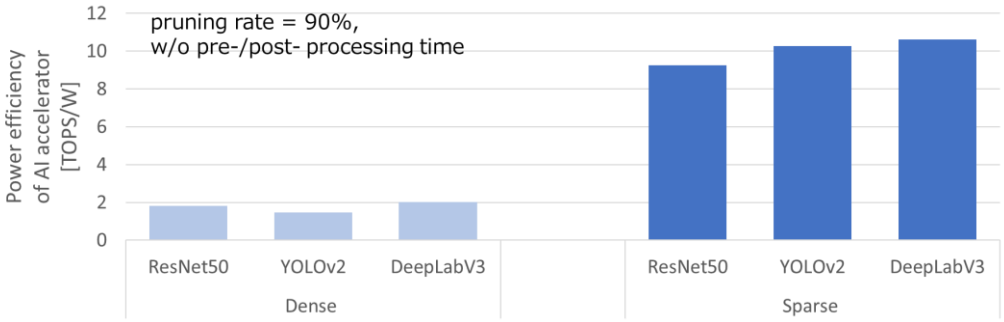

Figure 11: Power Efficiency of Real AI Models

We also showed that the same AI real-time processing could be performed on an evaluation board equipped with the prototype chip, without a fan at temperatures comparable to competitor products equipped with fans (Figure 12).

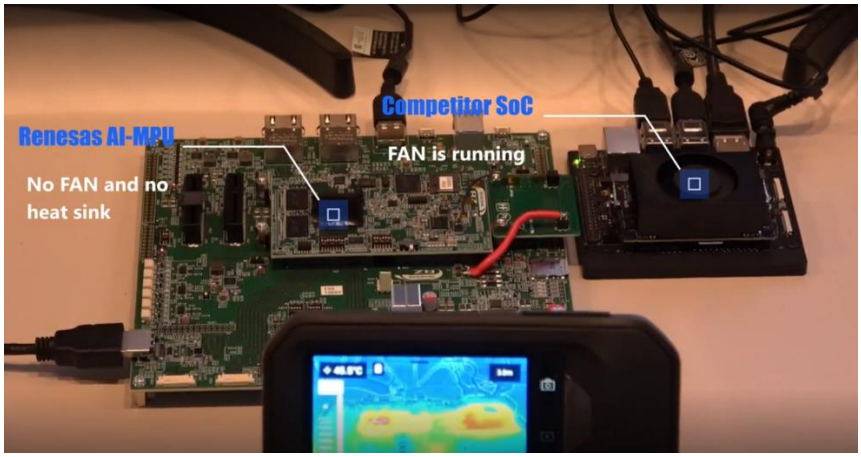

Figure 12: Comparison of Heat Generation between a Fanless DRP-AI Test Board and a GPU with Fan

(2) Examples of applications with robot applications

For example, SLAM (Simultaneously Localization And Mapping), one of the typical robot applications, has a complex configuration that requires multiple program processes for robot position recognition in parallel with environment recognition by AI processing. The Renesas DRP enables the robot to switch programs instantaneously, and parallel operation with an AI accelerator and CPU has proven to be about 17 times faster than CPU operation alone, and to reduce power consumption to 1/12 the level of CPU operation alone.

Conclusion

Renesas developed DRP-AI3, an advanced version of DRP-AI (Dynamically Reconfigurable Processor for AI), a unique AI accelerator that combines 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. Renesas will continue to expand MPU line ups to provide scalable MPU products (RZ/V series) equipped with this superior AI accelerator.

Renesas will release products in a timely manner responding to the AI evolution, which is expected to become increasingly sophisticated, and will contribute to deploy systems that respond to end-point products in a smart and real-time manner.

More detailed technical content was presented at ISSCC 2024 (International Solid-State Circuits Conference 2024), the top international conference on semiconductor circuits, held from February 18, 2024.

Reference

[1] “NVIDIA Jetson AGX Orin Series tech. brief”, 2022

https://www.nvidia.com/content/dam/en-zz/Solutions/gtcf21/jetson-orin/nvidia-jetson-agx-orin-technical-brief.pdf

Related Information

- RZ/V2H: Quad-core Vision AI MPU Delivering High-Efficiency AI Inference and High-Responsive Real-time Control
- DRP-AI: Renesas' proprietary AI accelerator that combines high AI inference performance with low power consumption

(Rev.1.0 Mar 2020)

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: R01WP0022EU0100