White Paper

Embedded AI-Accelerator DRP-AI

Hideaki Abe, Cognitive Product Department, IoT and Infrastructure Business Unit, Renesas Electronics Corporation

Koichi Nose, Cognitive Product Department, IoT and Infrastructure Business Unit, Renesas Electronics Corporation

Kazutaka Kikuchi, Cognitive Product Department, IoT and Infrastructure Business Unit, Renesas Electronics Corporation

June 2021

Abstract

With the remarkable advances in computing power in recent years, AI (Artificial Intelligence) is penetrating our lives, starting from cloud services. Reflecting this, according to Gertner's report[1], AI semiconductor market size is expected to grow from $12 billion in 2019 to $43 billion in 2024.

A new trend is the implementation of AI into endpoints from cloud. This is because endpoints, such as IoT devices and robots, are required to be smarter and react in real time. The AI required for endpoints is inference processing based on deep learning that replaces human perception such as vision and hearing.

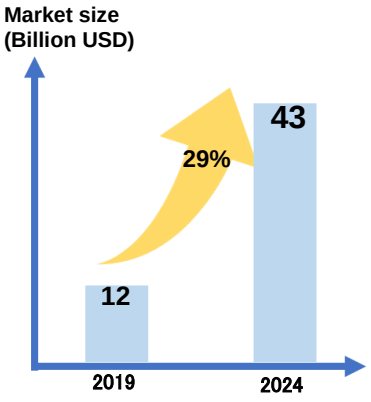

Figure 1: AI Semiconductors Market Dynamics

To implement AI in endpoints, two major challenges need to be overcome: First, power consumption limitations, and second, flexibility. While the cloud can be equipped sufficient power and cooling, endpoints are strictly required to limit power consumption which can cause shorter runtimes, generate heat, or increase costs. The idea of power consumption saving is to utilize dedicated hardware that is specialized for specific AI processing; however, the hardware will soon become obsolete since AI models are evolving day by day. Therefore, AI acceleration in endpoints is required to provide the flexibility to support newly developed AI models.

Renesas has developed the DRP-AI (Dynamically Reconfigurable Processor for AI) as an AI accelerator with high-speed AI inference processing that achieves the low power and flexibility required by endpoints based on the reconfigurable processor technology it has cultivated over many years.

[1] Graph created by Renesas based on Gartner Research, Source: Forecast Analysis: AI Neural Network Processing Semiconductor Revenue, worldwide, Alan Priestley, 20 Apr 2020, Revenue Basis

DRP-AI Features

- AI accelerator dedicated AI inference
- High power efficiency by the collaboration HW (DRP-AI) and SW (DRP translator)
- Supported AI model extension by continuous update of DRP-AI translator

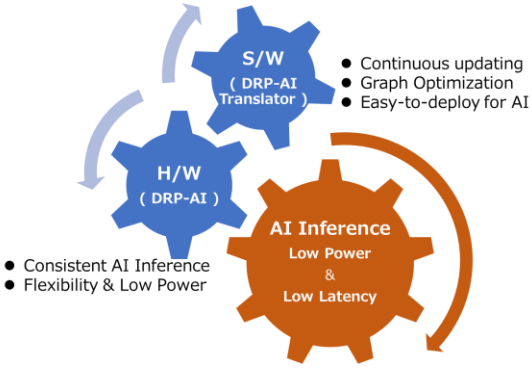

Figure 2: DRP-AI Features

The DRP-AI translator is a hardware dedicated to AI inference, but it achieves flexibility, high-speed processing, and power efficiency by utilizing Renesas' unique dynamic reconfigurable technology. The DRP-AI translator is provided to enable users the ability to easily implement AI models optimized to maximize the performance of DRP-AI on this flexible hardware. Multiple executables output by the DRP-AI translator can be placed in external memory. This makes it possible to dynamically switch between multiple AI models as a system. In addition, the DRP-AI translator can be continuously updated to support newly developed AI models without hardware changes.

DRP-AI Hardware Architecture

- DRP (Dynamically Reconfigurable Processor): Programable hardware
- AI-MAC (multiply-and-accumulate): Hardware dedicated for MAC computing
- DMAC (Direct Memory Access Controller)

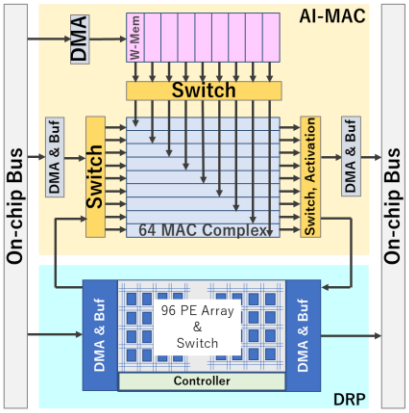

Figure 3: DRP-AI H/W Architecture

DRP-AI is composed of AI-MAC and DRP (Dynamically Reconfigurable Processor), which can efficiently process operations in convolutional and all-combining layers by optimizing data flow with internal switches. The DRP can process complex processing such as image preprocessing and AI model pooling layers flexibly and quickly by dynamically changing the hardware configuration. The DRP-AI translator automatically allocates each process of the AI model to the AI-MAC and DRP, thus allowing the user to easily use DRP-AI without being aware of the hardware.

DRP-AI Translator

- Tool to generate DRP-AI optimized executables from trained ONNX model
- Optimizing the graph structure of AI models to minimize memory access and improve computing efficiency
- Extension of supported AI models by continuous updates

The DRP-AI translator is a tool that generates DRP-AI optimized executables from trained models based on the ONNX format, which is independent of various AI frameworks. The tool's internal tasks are:

1. Scheduling of each operation to process the AI model
2. Hiding the overhead such as memory access time that occurs during the transition of each operation in the schedule defined in 1.
3. Optimization of graph structure in the network (Layer fusion, DRP & AI-MAC processing allocation)

Using the DRP-AI translator, users can implement automatically optimized AI models from ONNX AI models into DRP-AI without knowledge of the DRP-AI hardware configuration. The user can then simply make calls through the supplied driver to run the high-performance AI model.

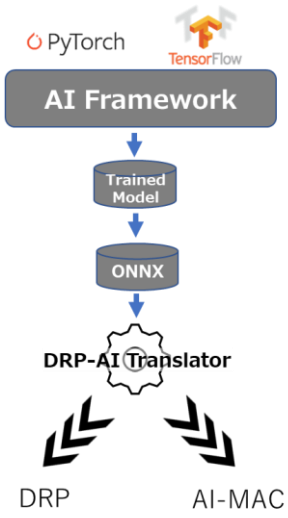

Figure 4: AI model implementation flow by DRP-AI

Architecture for High Power Efficiency

- Reduction of external memory communication volume by data reuse technique
- Low power control using inputted zero data
- Scheduling of operation flow

Reduction of external memory communication volume by data reuse technique

The power consumption of AI accelerators is not only due to the enormous matrix operations, but also the power components due to data transactions between the accelerator's internal components or external memory becoming large. In addition, as shown in **Figure 5**, the ratio of the amount of data related to weight/input/output are different depending on the image size or types of models etc., and the power bottleneck factors are diverse. Therefore, for comprehensively reducing the power consumption of AI models execution, it is necessary to reduce the amount of memory access for all data types.

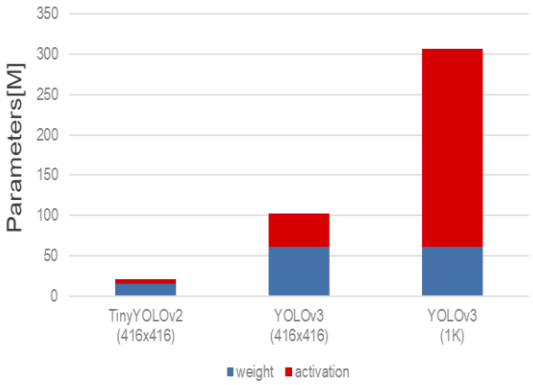

Figure 5: Data Structure of AI Model

As an effective way to reduce the amount of external memory access, DRP-AI employs a technology that efficiently reuses data which input to AI-MAC one time.

For example, in a convolutional operation using a 3x3 filter, one pixel of data is used for nine filter operations. im2col, which is widely used as a highly parallel operation method in GPUs, expands all the image data in the order of matrix operations as a pre-processing step for input to the GPU. At this time, the data information of one pixel appears nine times, so the number of data increases by a factor of nine. This causes an increase in power consumption and communication bandwidth. On the other hand, AI-MAC can reuse the data by shifting the data taken into the register corresponding to the MAC arithmetic unit to the adjacent register. The specific flow is explained using **Figure 6**.

1. Data stored in the external memory (blue) is loaded into the AI-MAC buffer (Figure 6-left)
2. Transfer data (blue) from buffer to register (Figure 6-center)
3. The data in the register is used to perform operations in the corresponding MAC arithmetic unit (Figure 6-right)
4. Shift data (blue) to the next register down (data reuse)

By adopting this configuration, the number of data loads from external memory and internal buffer to the AI-MAC can be reduced by up to a factor of nine compared to the GPU. As a result, the power and communication bandwidth required for data movement is significantly reduced. In addition, AI-MAC can reuse data not only for input data, but also for output and weight information, reducing access to external memory by more than one order of magnitude.

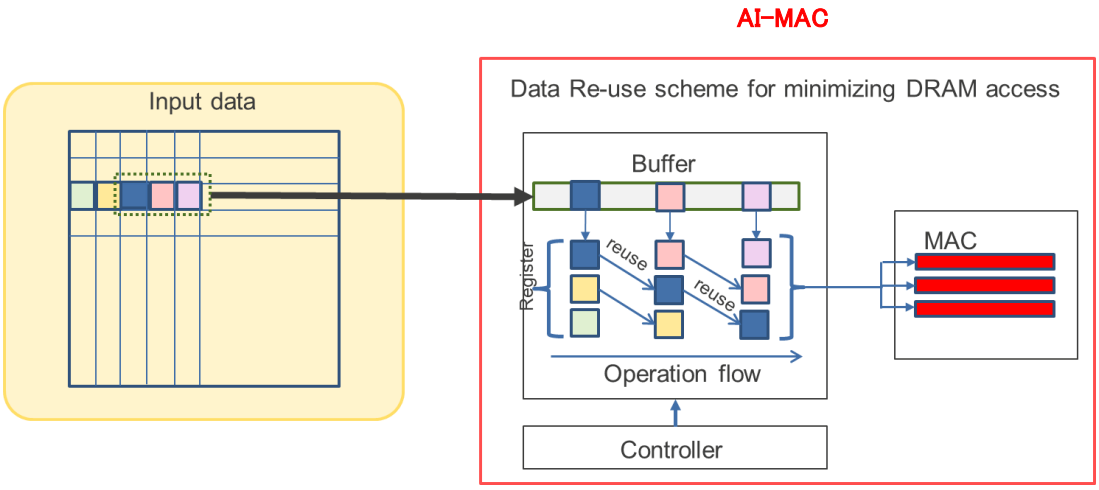

Figure 6: Data Reuse Diagram

Low power control using inputted zero data

One of the characteristics in AI model computation is the high ratio of "zero" values in the weight data and input/output data of each layer (it is called sparsification). For example, as shown in **Figure 7**, in the image recognition model, more than 50% of the input and output data of all layers are zero values on average. This is because many AI models use an activation function (ReLU) that replaces all negative results of the sum-of-products operation with zero. In DRP-AI, unnecessary computational power is reduced by introducing a switching technology that detects in advance when zero is entered in the input for each element of the operation and prevents unnecessary operations.

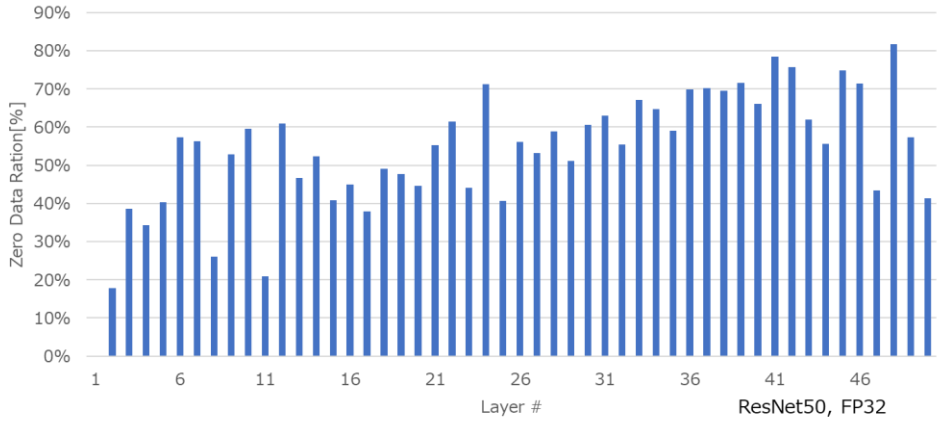

Figure 7: Zero Data Ratio of AI Models

Scheduling of operation flow

In addition to the data reuse techniques mentioned above, optimization of the order and timing of operations such as external data access or MAC operation processing, etc. is essential for efficient AI execution. In other words, scheduling the operation flow can maximize the performance of DRP-AI. One example is described below. By scheduling the external memory access timing so that the weight information for the next operation is read ahead and stored in the buffer during AI-MAC operation, the external memory access latency can be hidden. Such cases also occur in the timing of internal memory access and any internal arithmetic processing, and scheduling can avoid unnecessary waiting time and power generation between each process. Since the DRP-AI translator automatically generates this optimized scheduling, the user can easily handle the DRP-AI.

Evaluation

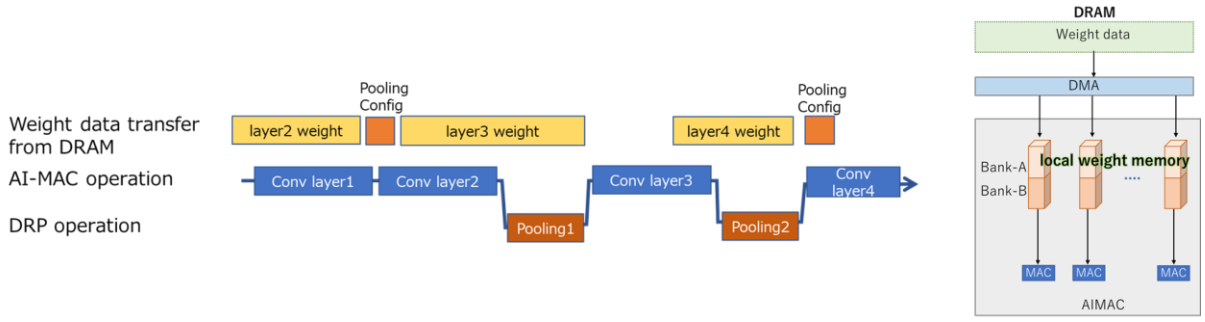

Figure 8: Example of Operation Flow Scheduling

We implemented TinyYolov2[2] in a DRP-AI test vehicle, the high power-efficient architecture which has been previously described, and measured the surface temperature of the device using thermography. In the experiment, we ran the AI under the same conditions using a commercially available GPU as a comparison to express the performance of DRP-AI in a digestible way. The results are shown in **Figure 9**. You can see that the surface temperature of our DRP-AI test vehicle is clearly lower than that of the commercial GPU.

The surface temperature of the DRP-AI test vehicle was 40.9° without heat sink when running TinyYolov2 at 42 fps[3][4]. On the other hand, that of the commercial GPU was 79.0° despite having a heat sink when running AI at a lower framerate than DRP-AI.

In this experiment, similar to a real use case, we hope that you have understood that the product with DRP-AI is a product that can withstand mass production of endpoint products under severe temperature constraints even when running practical level AI.

[2] Tiny Yolov2: https://pjreddie.com/darknet/yolov2/

[3] The value is AI inference only

[4] The performance varies from product to product due to the difference in device configuration

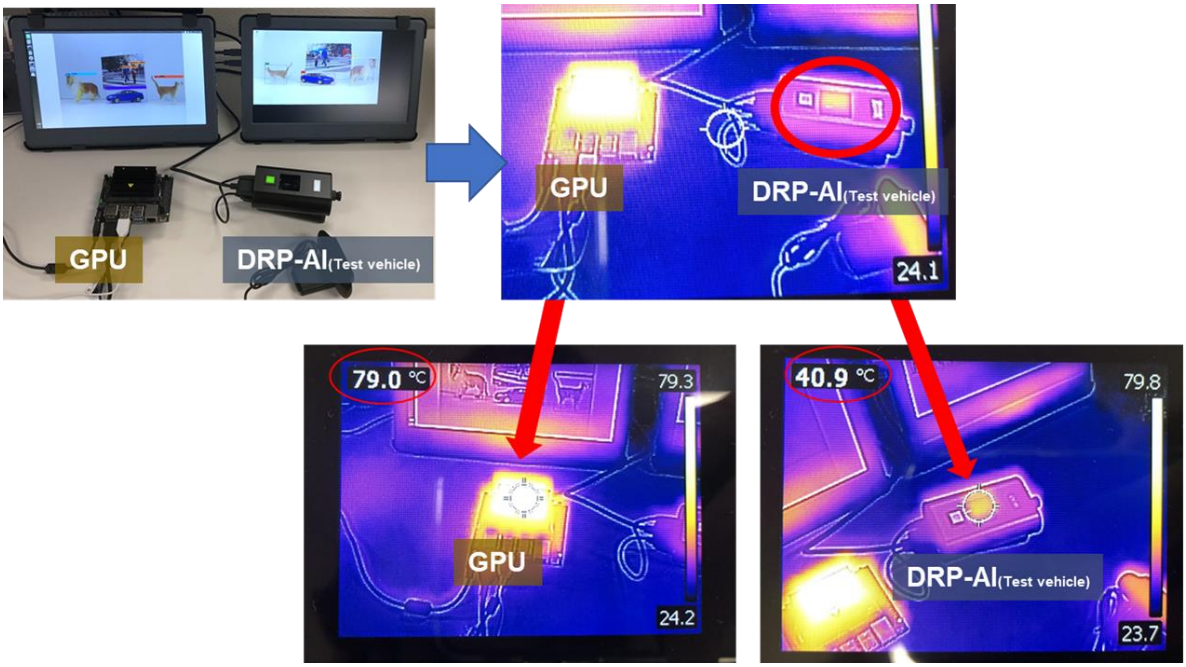

Figure 9: Surface Temperature of the Device During AI Execution

Conclusion

Renesas has developed the DRP-AI (Dynamically Reconfigurable Processor for AI) as an AI accelerator with high-speed AI inference processing that achieves the low power and flexibility required by endpoints. We will deploy MPU products equipped with this superior AI accelerator in a scalable manner. This will help endpoint products to be equipped with AI that reacts intelligently and in real time.

Learn More

RZ/V2M Renesas' original AI-dedicated Accelerator (DRP-AI), 4K-compatible Image Signal Processor (ISP), Vision-AI ASSP for real-time human and object recognition

RZ/V2L General-purpose microprocessor with Renesas' original AI-dedicated Accelerator "DRP-AI", 1.2GHz Dual-Core Arm® Cortex®-A55 CPU, 3D Graphics, and Video Codec Engine