<News Release>

New Processing-In-Memory Technology for Next-Generation AI Chips

e-AI business operation department
Industrial Solution Planning & Strategy division, IBU
Renesas Electronics

June 13, 2019

2BP-C-19-0070

2019 Symposia on VLSI Technology and Circuits.

Oral presentation

Date:	at 9:20am on 13th June (JST)
Session:	Technology and System for AI
Title :	“A Ternary Based Bit Scalable, 8.80 TOPS/W CNN Accelerator with Multi-Core Processing-in-Memory Architecture with 896K Synapses/mm^2”

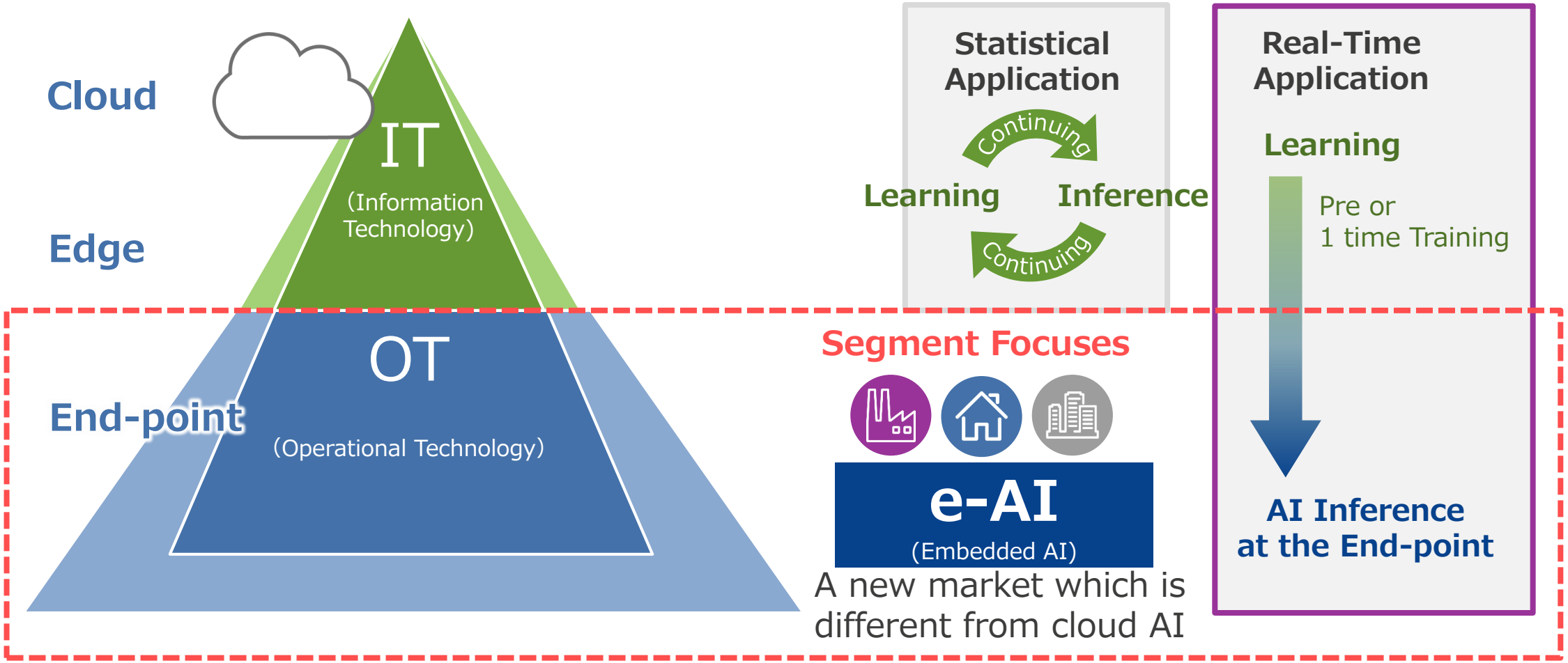
Demonstration

Date:	5:30pm on 10th June (JST)
Session:	Demo Session & Reception

RENESAS e-AI (embedded-AI)

OT Applications are expanding with e-AI

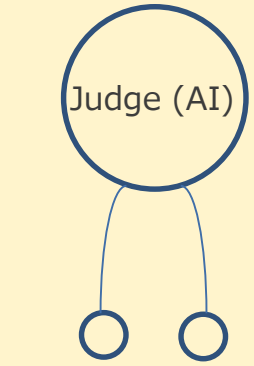
Difference between cloud AI and e-AI on OT



e-AI: Specific to Inference for End-point

Concept release in 2015; e-AI tools introduced in July 2017

- Learning mainly
- Big data
- High-speed processing
- AI designated



Statistical Big Data Use of AI

Maximize

Cloud AI

Optimize

Edge AI

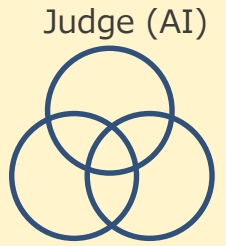
Continuous Learning

Pre- or One-Time Learning, Inference

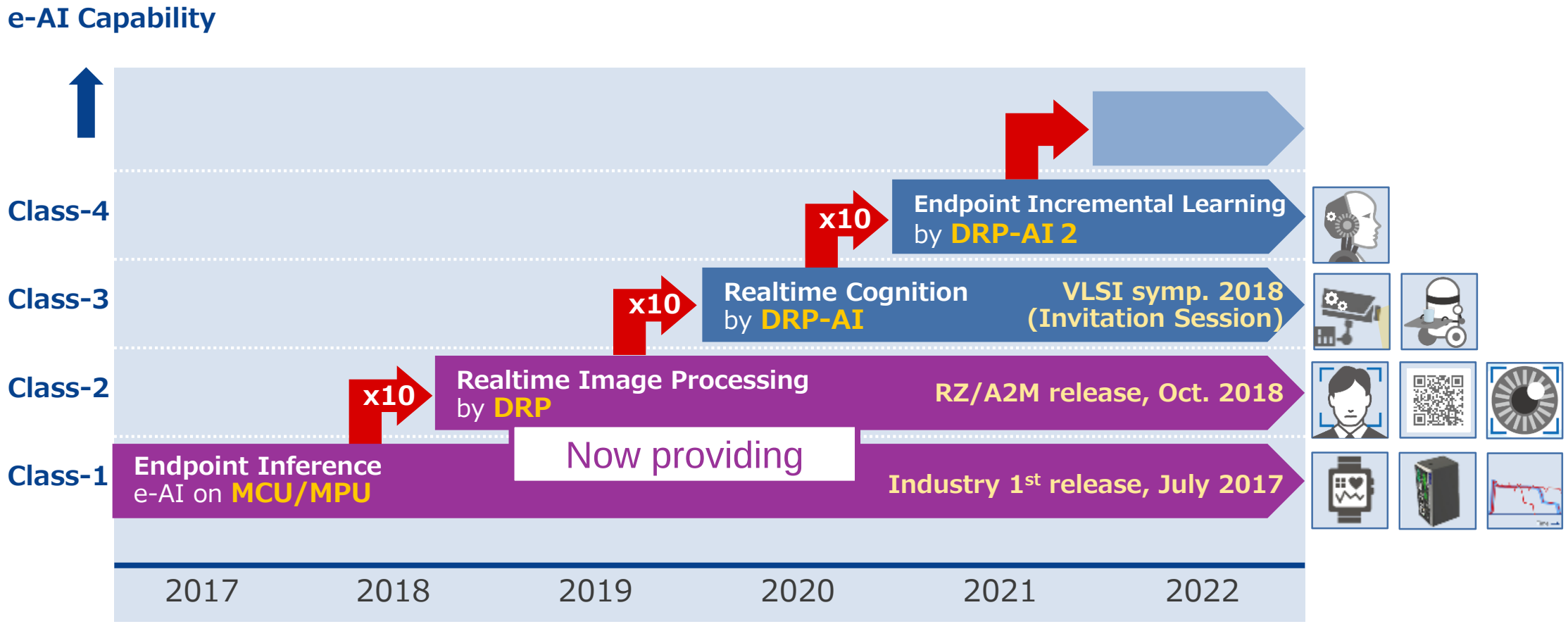
End-point e-AI

Realtime data use of AI

- AI specific to inference
- Fusion with embedded technology
- Application AI



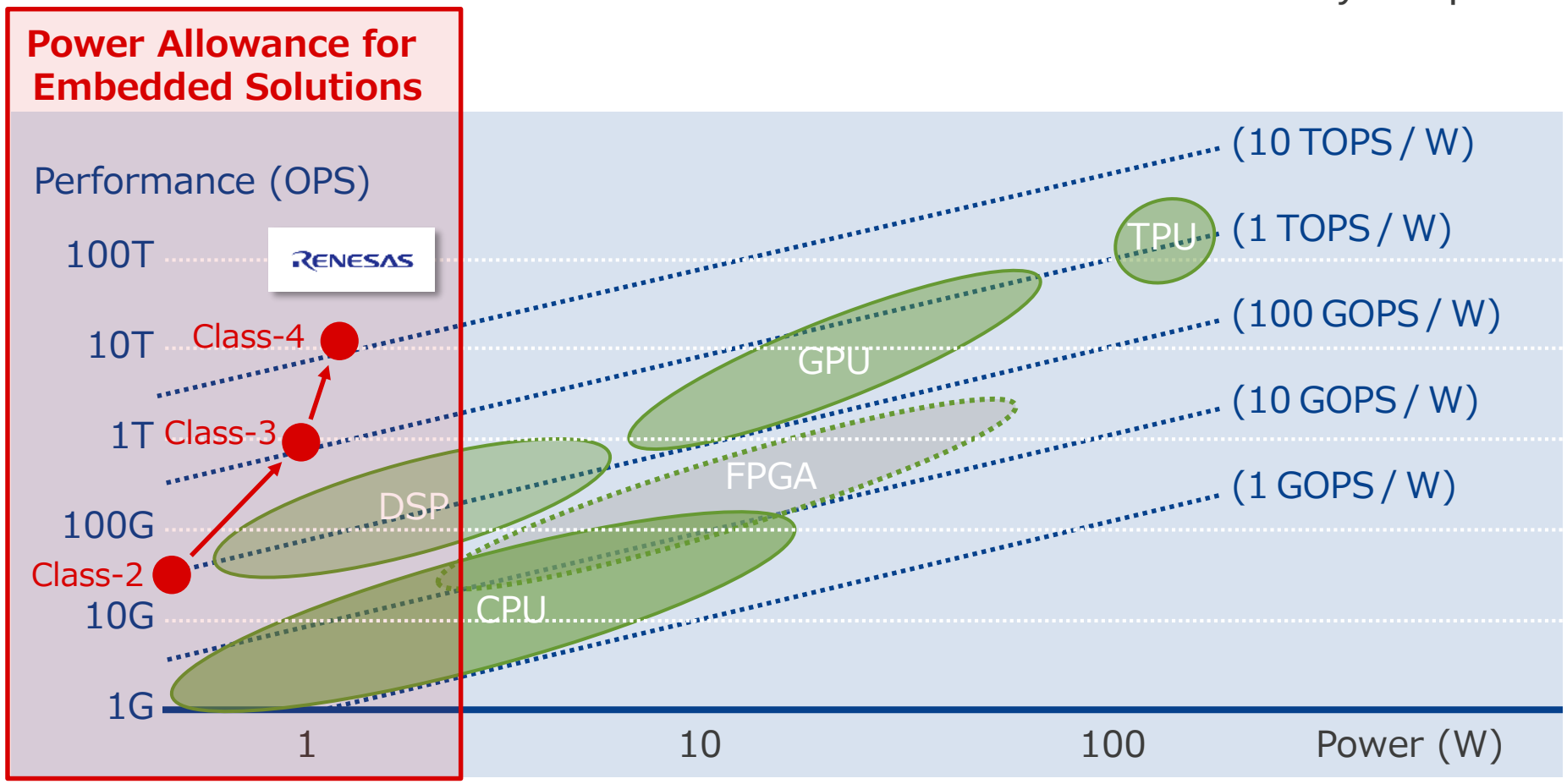
DRP Accelerates e-AI Roadmap
10 Times the Performance Per 1.5 Years



DRP: Dynamically Reconfigurable Processor

Power Efficiency of e-AI

DRP realizes both flexibility and power efficiency



Technical Details

Overview of the News on 13th June

Renesas Electronics Develops New Processing-In-Memory Technology for Next-Generation AI Chips and Demonstrates AI Processing Performance of 8.8 TOPS/W

AI Accelerator Achieves Both CNN Processing Speeds and Reduced Power Consumption

Renesas Electronics has developed an AI accelerator that performs CNN (convolutional neural network) processing at high speeds and low power to move towards the next generation of Renesas embedded AI (e-AI). A test chip featuring this accelerator has achieved the power efficiency of 8.8 TOPS/W – the industry's highest class of power efficiency.

The accelerator is based on the processing-in-memory (PIM) architecture and features three new developments:

1. Ternary-valued (-1, 0, 1) SRAM structure PIM technology that can perform large-scale CNN computations.
2. SRAM circuit to be applied with comparators that can read out memory data at low power.
3. Technology that prevents calculation errors due to process variations in the manufacturing.

These technologies achieve both a reduction in the memory access time in deep learning processing and a reduction in the power required for the multiply-and-accumulate operations, while maintaining an accuracy ratio of over 99 percent when evaluated in a handwritten character recognition test (MNIST).

Opportunities for Class-4 e-AI

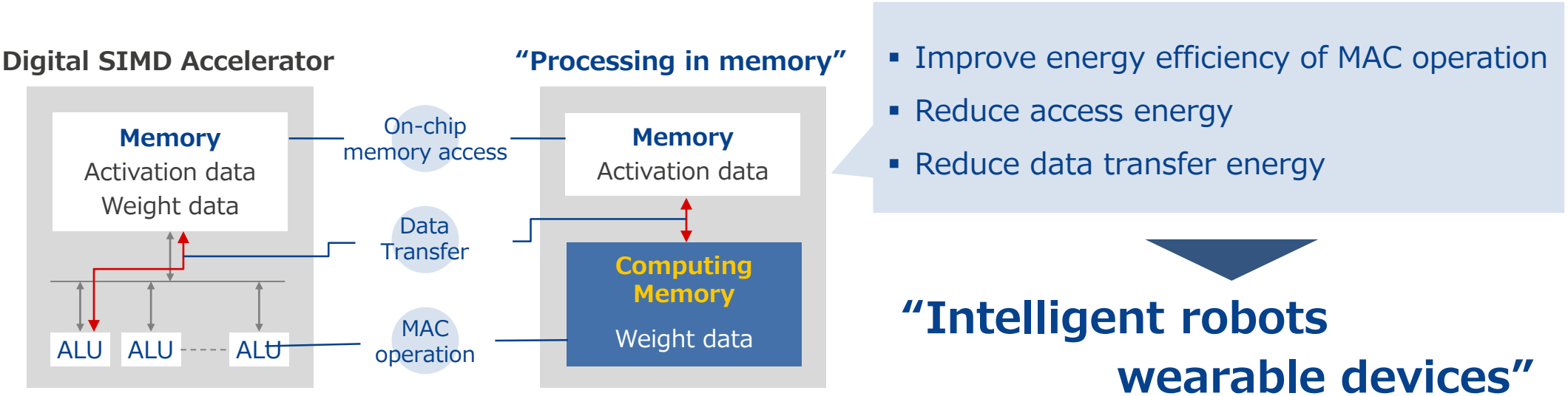
- The new technology could be **one of the key technologies** to realize **e-AI Class-4 (10 TOPS/W class)** performance

- To realize e-AI Class-4 (10 TOPS/W class) performance, two issues need to be achieved: 1) **A reduction in the memory access time** and 2) **a reduction in the power for the multiply-and-accumulate operations** in deep learning processing.

Technology Architecture

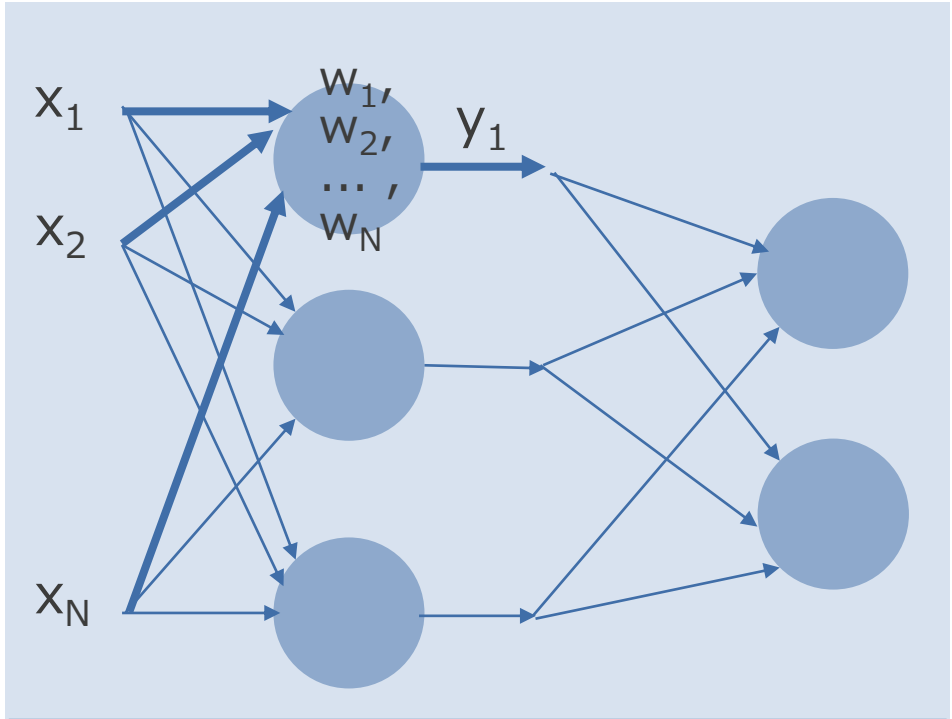
- Advanced SRAM structure PIM technology is applied.

- PIM architecture enables “multiply-and-accumulate operations in the memory circuit as data is read out from that memory.”

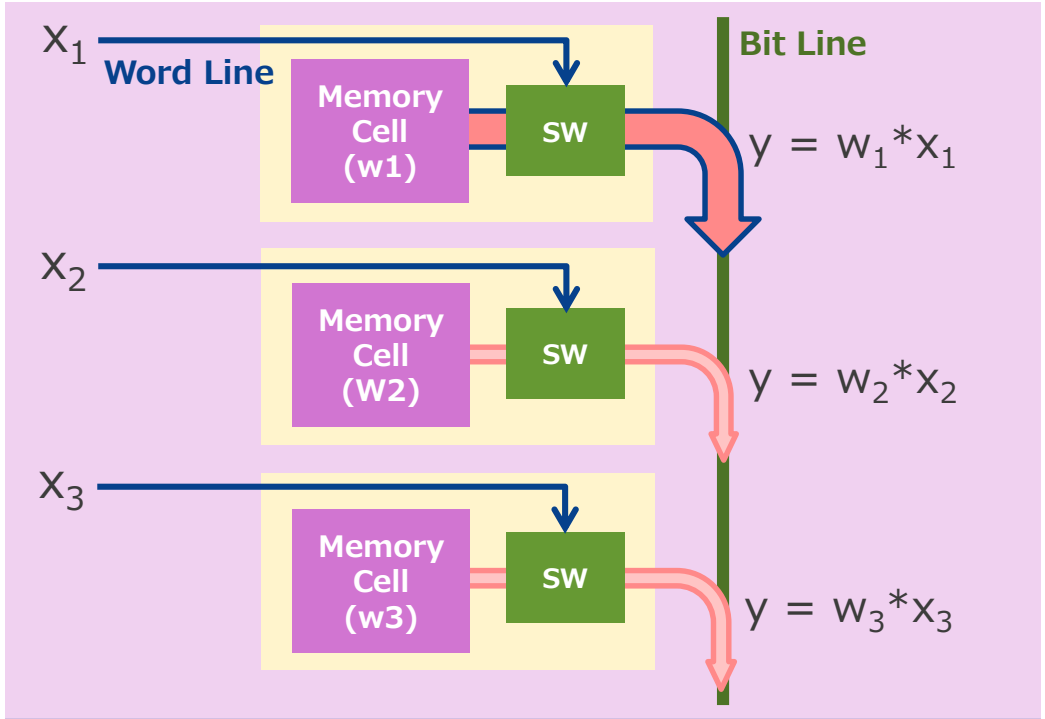


Processing-in-Memory (Circuit Principle)

MAC (Multiply-and-Accumulate) operations in the memory circuit as data is read out from that memory, which enables a reduction of both the memory access time and power consumption.



=



MAC result at 1 neuron (node)

$y = w_1 * x_1 + w_2 * x_2 + ... + w_N * x_N$

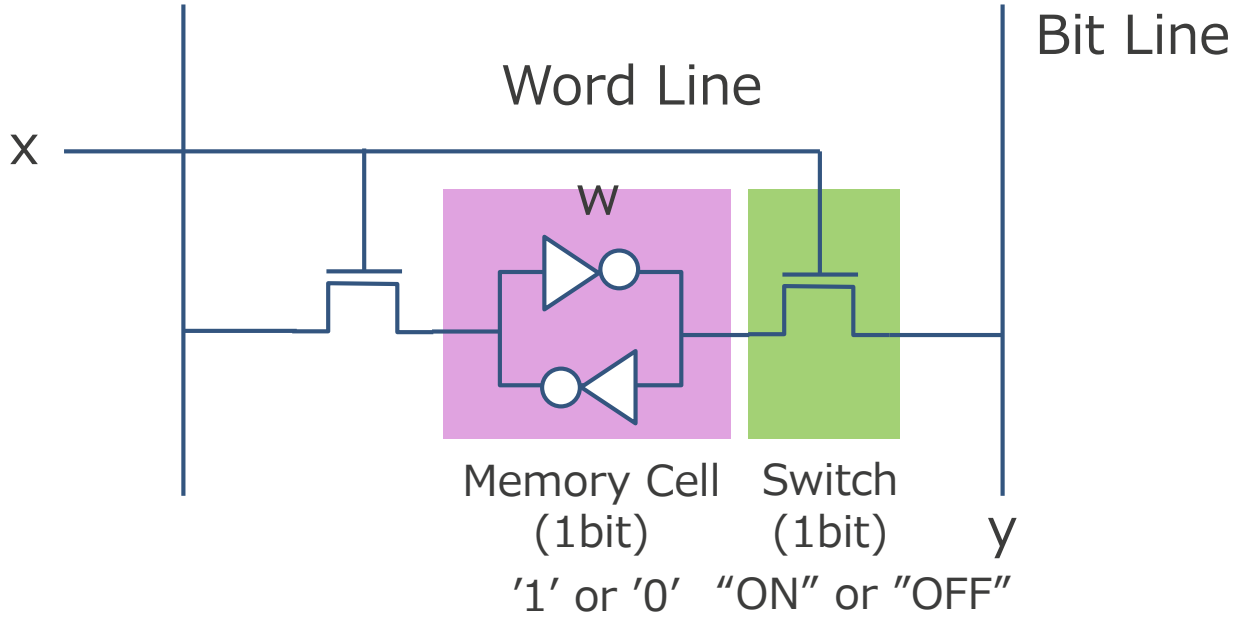
Whole amount of current

$y = w_1 * x_1 + w_2 * x_2 + ... + w_N * x_N$

PIM Challenges

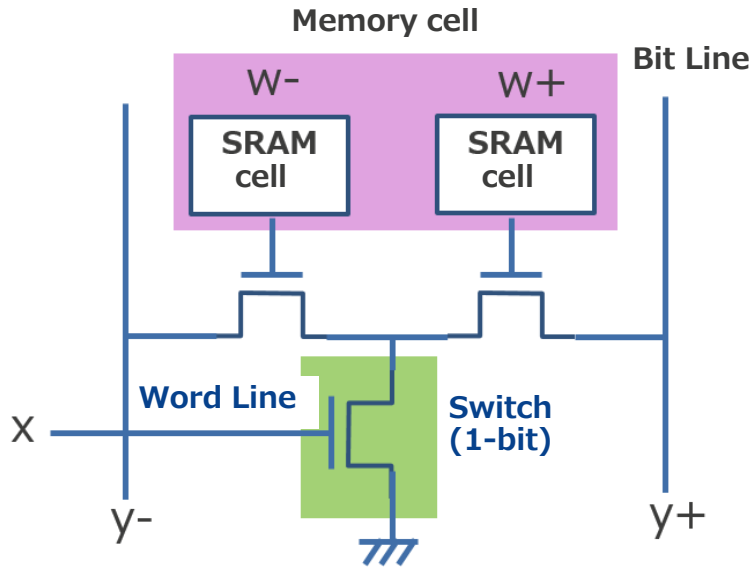
Traditional SRAM structure PIM technology:

- 1 bit (binary (0, 1)) calculation is basis.
- Calculation bit number is up to bit number of cell.
- Process variation occurs the errors of MAC results.



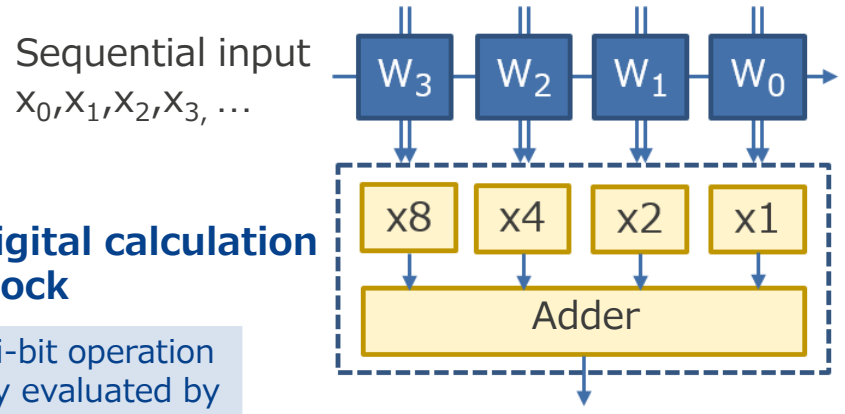
Ternary-Valued SRAM Structure PIM Technology That Can Perform Large-Scale CNN Computations

- Ternary SRAM-based (-1, 0, 1) PIM architecture allows switching the number of bits between, for example, 1.5-bit (ternary) and 4-bit calculations, according to the required accuracy.

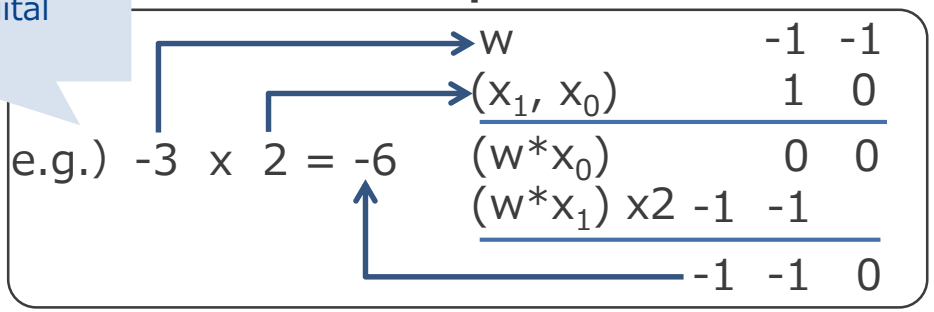


w	w+	w-
0	“0”	“0”
-1	“0”	“1”
+1	“1”	“0”

Increase the number of states from binary to ternary values (“-1”, “0”, “+1”)



Signed multi-bit operation can be easily evaluated by using ternary memory cells and simple digital calculation blocks

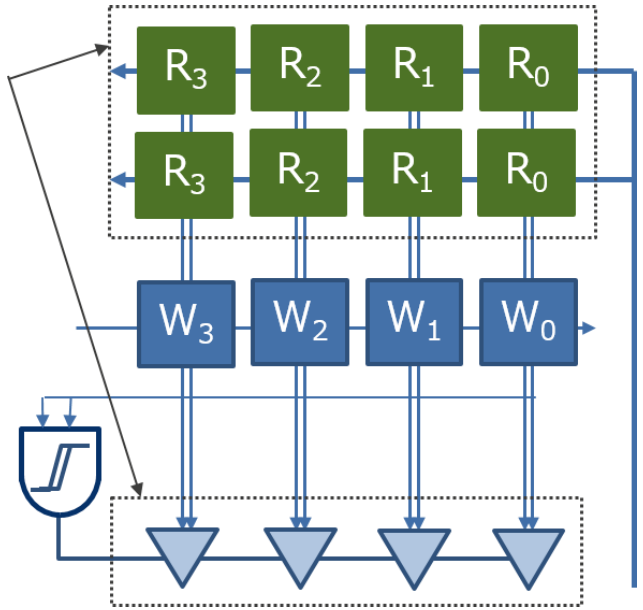


Low-Power Memory Data Readout Circuit That Combines Comparators and Replica Cells

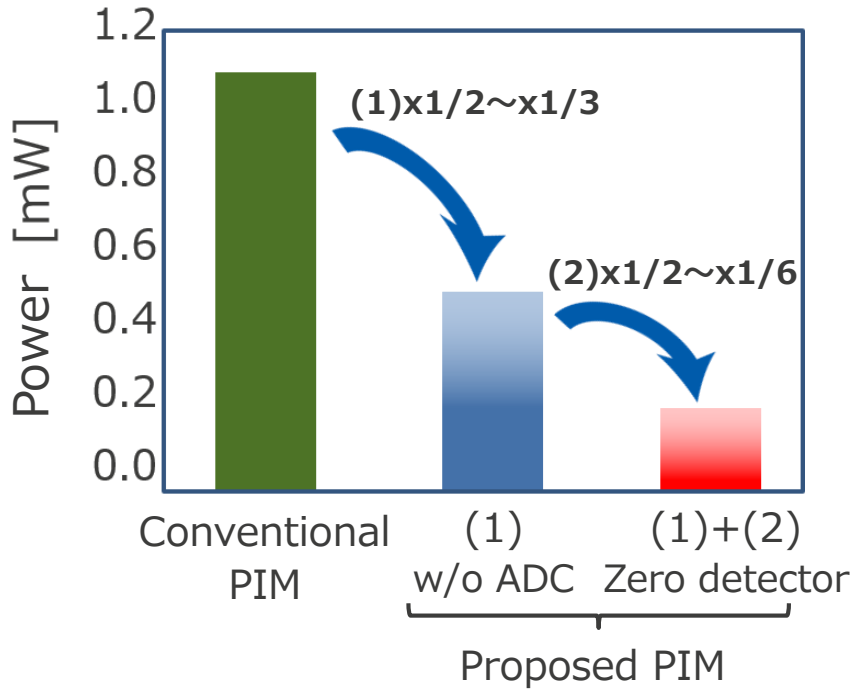
- High-accuracy memory readout circuits without using area- and power-consuming A/D converters.
- The number of activated nodes (neurons) in neural network is very small; lower power operation is achieved by stopping operation of the readout circuits for inactive nodes.

(1)
Multi-level current detection with comparators and replica cells

(2)
"Zero-detector" detects the state that MAC result is equal to zero, and stops operation of comparators

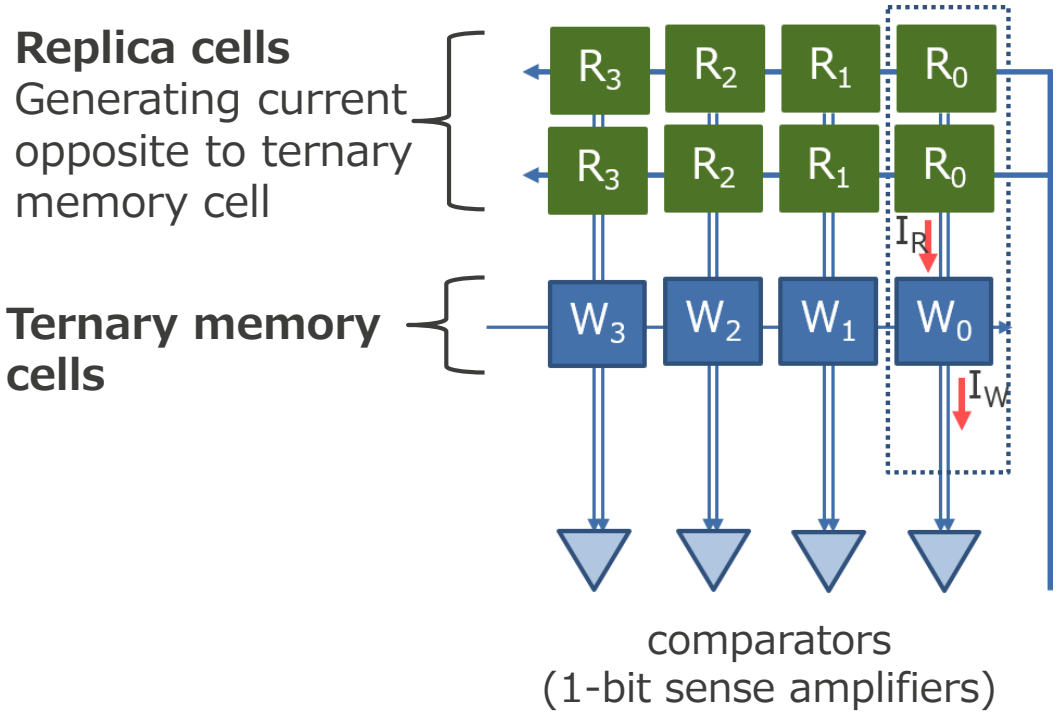


comparators
(1-bit sense amplifiers)

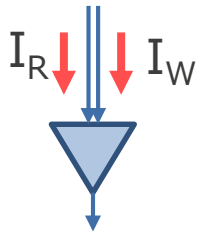


Low-Power Memory Data Readout Circuit Cont.

In order to measure ternary memory cell current with comparators, current comparison between ternary memory cells and replica cells, in which the current can be controlled flexibly, is iterated by bit serial.

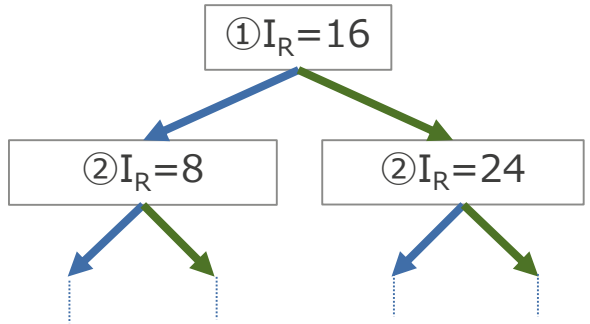


Role of comparator



Comparison between replica cell current (I_R) and ternary cell current (I_W).

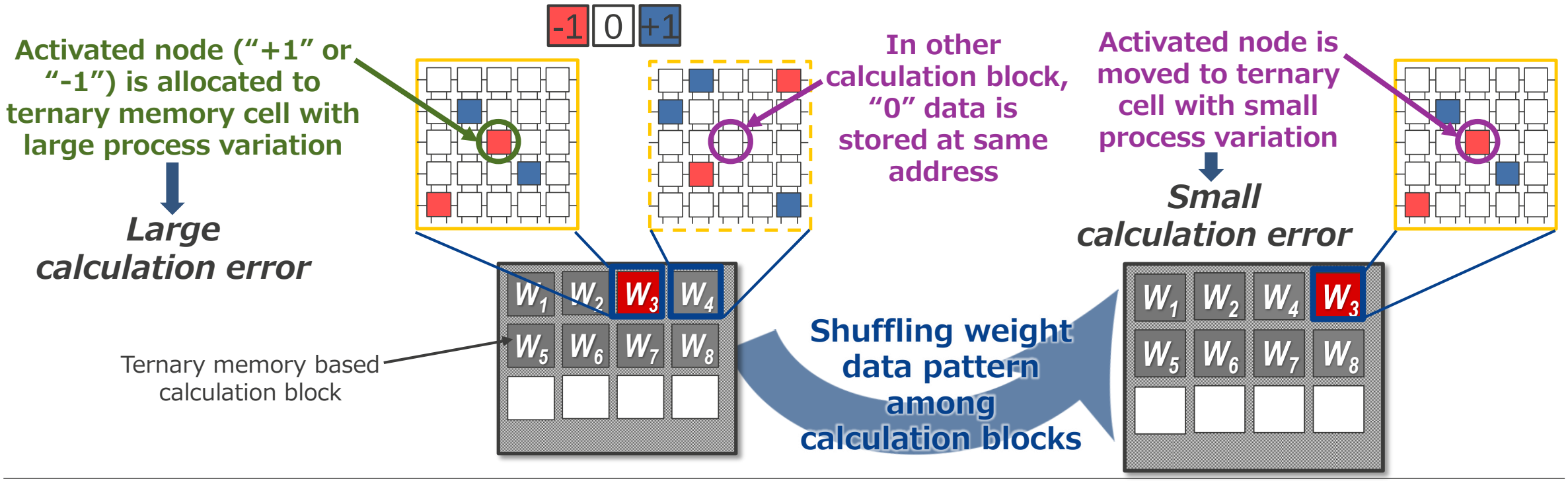
Operation flow



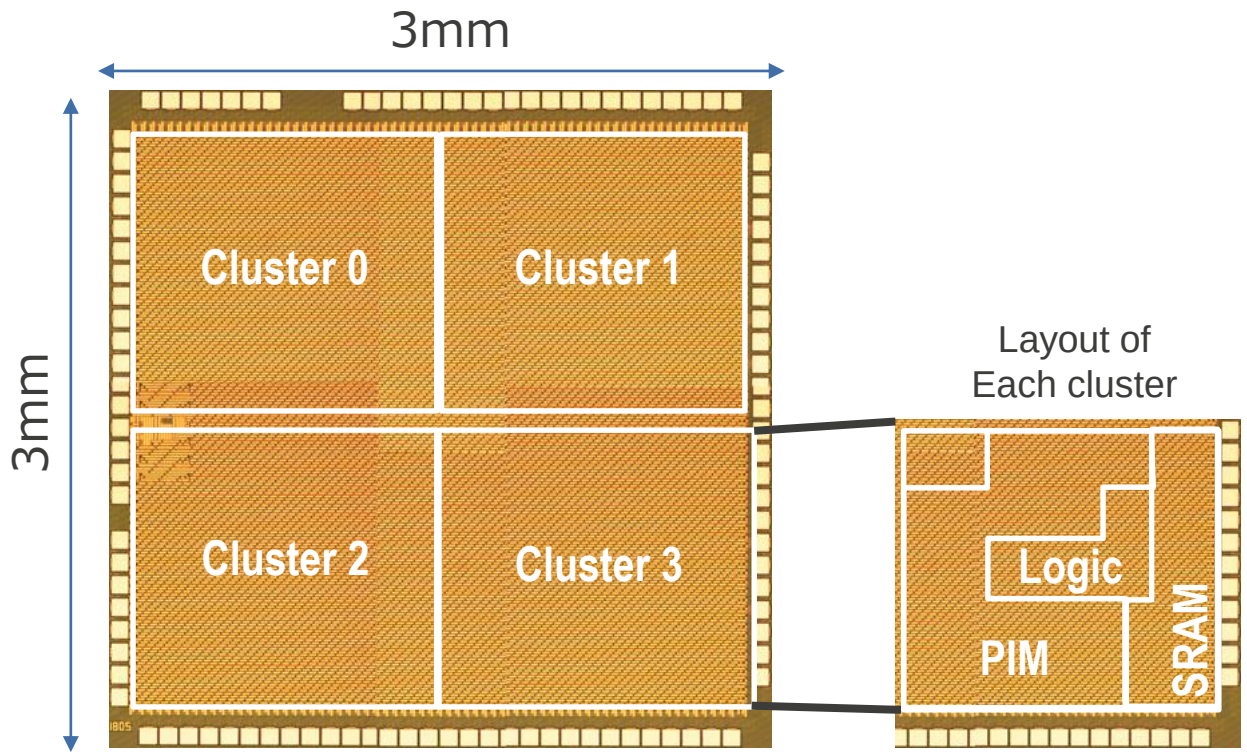
Bit-serial current measurement to search I_W

Technology That Prevents Calculation Errors Due to Process Variations in the Manufacturing

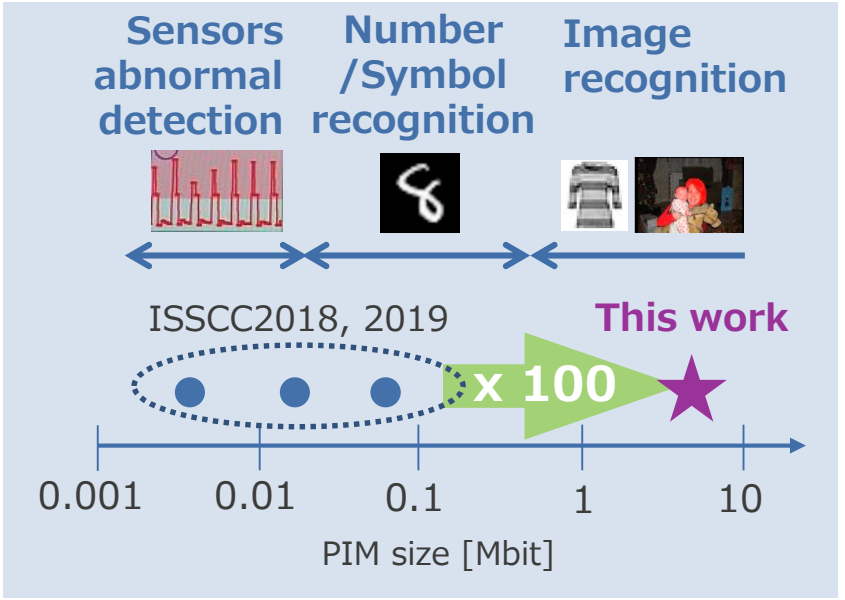
- Activated nodes are allocated selectively to SRAM calculation blocks that have minimal manufacturing process variations.
- Accurate calculation can be evaluated without additional circuits in SRAM macros, such as redundant cell arrays.



Test Chip Overview



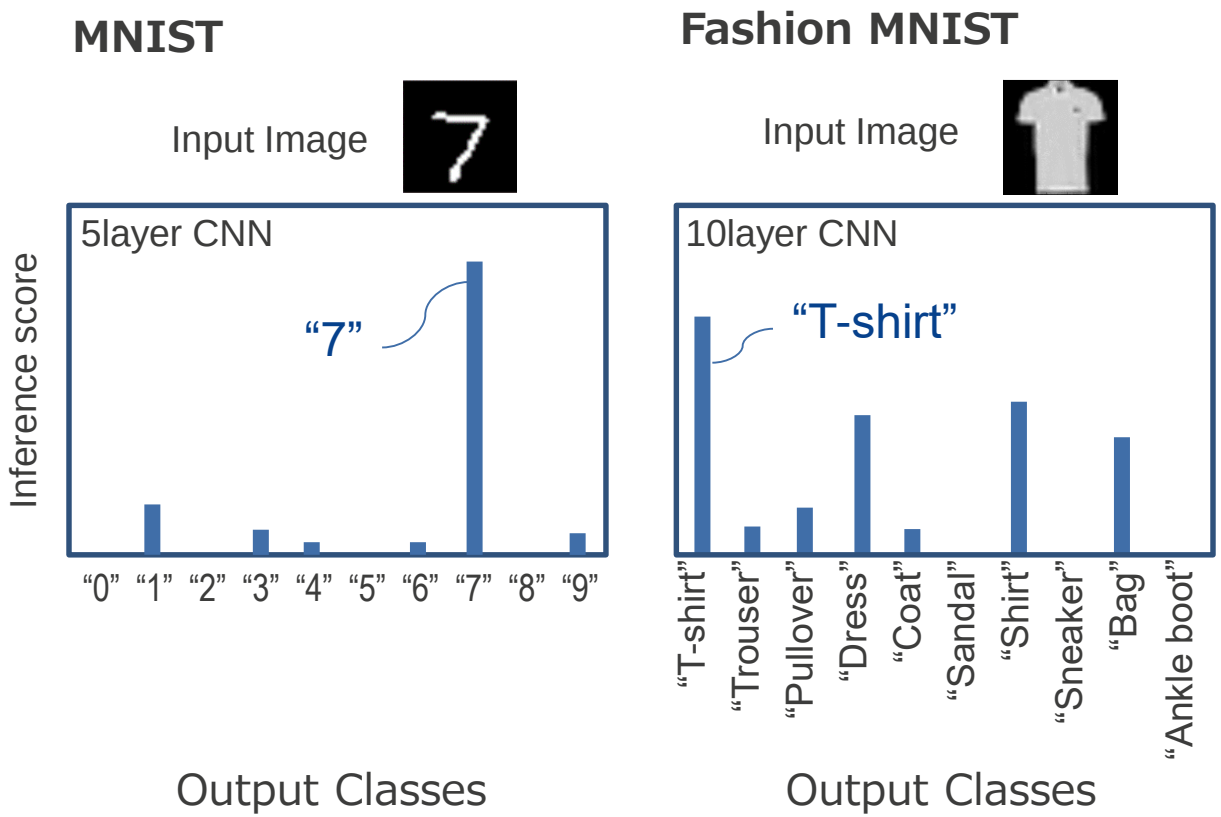
4 clusters incorporated, which is equivalent to CPU core
->4 kinds of AI processing can be executed simultaneously



Process [nm]	12
Area [mm²]	9.0
# CNN Layer	1-128
PIM [Mb]	4.74
SRAM [Mb]	12.58

Test Chip Evaluation Results

- An accuracy ratio of more than 99 percent when evaluated in a handwritten character recognition test (MNIST).
- Calculation errors is reduced to a level where they can be essentially ignored



MNIST Acc.: accuracy

	Acc. (SW)	Acc. (Chip)	Acc.(SW) – Acc.(Chip)
VLSI '16[6]	---	90	---
ISSCC '18[3]	99*2	96*2	3*2
VLSI '18[4]	98.75	98.58	0.17
This work	**98.91 (99.0*3)**	**99.0*3**	**0*3**

*2 100 test image *3 300 test image

Performance Comparison

In this work, both bit scalability and power efficiency are highly established compared with prior works

★★Extremely superior ★Superior △Same level ×Inferior

	Digital Architecture				Processing-in-Memory					
	ISCA '16		ISSCC '18		ISSCC '18		VLSI '18		**Our chip**	
Process node（nm）		28 (Sim.)		40		65		65		12
Bit number (Scalability)	★	4bit	★★	1-4bit	△	1bit	△	1bit	★★	1.5-4bit
On-chip memory available for calculation	★★	41MB	★	7.6MB	×	10Kb PIM (not adequate)	×	2Mb PIM 8KB SRAM (not adequate)	★	4Mb PIM 1.5MB SRAM
Power efficiency @NAM macro		No data		No data	★	28.1 TOPS/W	★★	658 TOPS/W	★	79.3 TOPS/W
Power efficiency @system	★	4.1 TOPS/W (Sim.)	△	2.2 TOPS/W (1bit)	×	not evaluated	×	not evaluated	★★	8.8 TOPS/W (Ternary)

APPENDIX

Demonstration Overview

Renesas is demonstrating real-time image recognition using a prototype AI module in which the test chip, powered by small battery, are connected with a microcontroller, a camera and other peripheral devices and combined with development tools.