SMART FACTORY WITH ARTIFICIAL INTELLIGENCE AT ENDPOINT

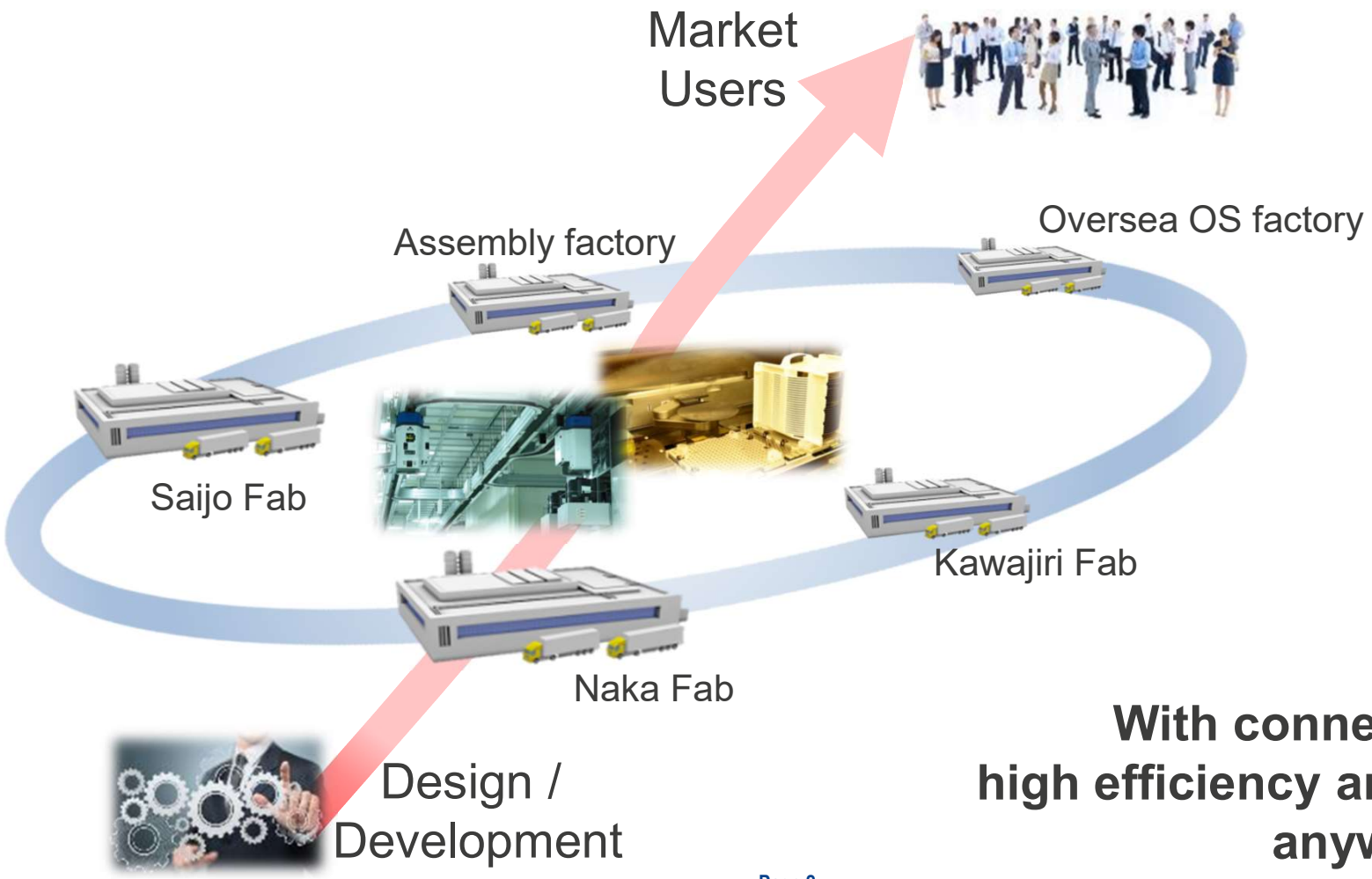
ATSUSHI HASEGAWA

IOT AND INFRASTRUCTURE BUSINESS UNIT
RENESAS ELECTRONICS CORPORATION

MPSOC'19
July 8th, 2019

BIG IDEAS FOR EVERY SPACE RENESAS

SMART FACTORY CHALLENGE

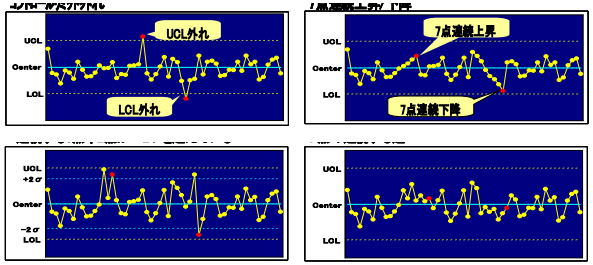


**With connected factories,
high efficiency and high quality
anywhere, anytime**

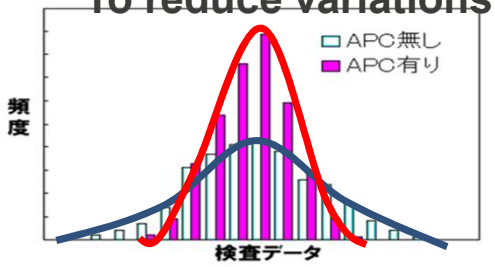
FURTHER HIGH-QUALITY: FDC CHALLENGE

FDC： Fault Detection and Classification

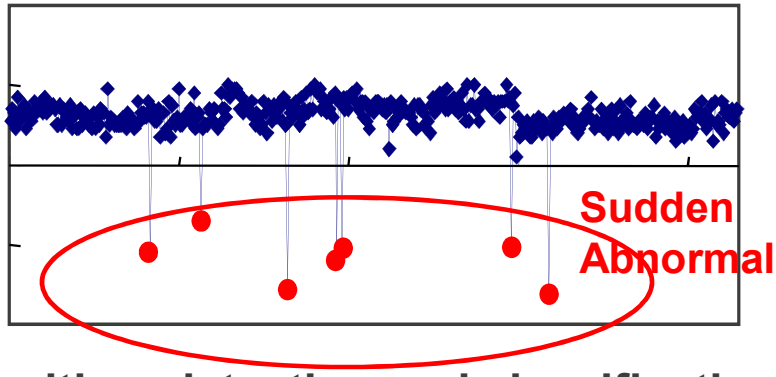
SPC（Statistical Process Control）
24/7 Auto Trend Monitoring



APC
Feedback/Feed-forward
To reduce variations



SPC not handle Sudden Abnormal well

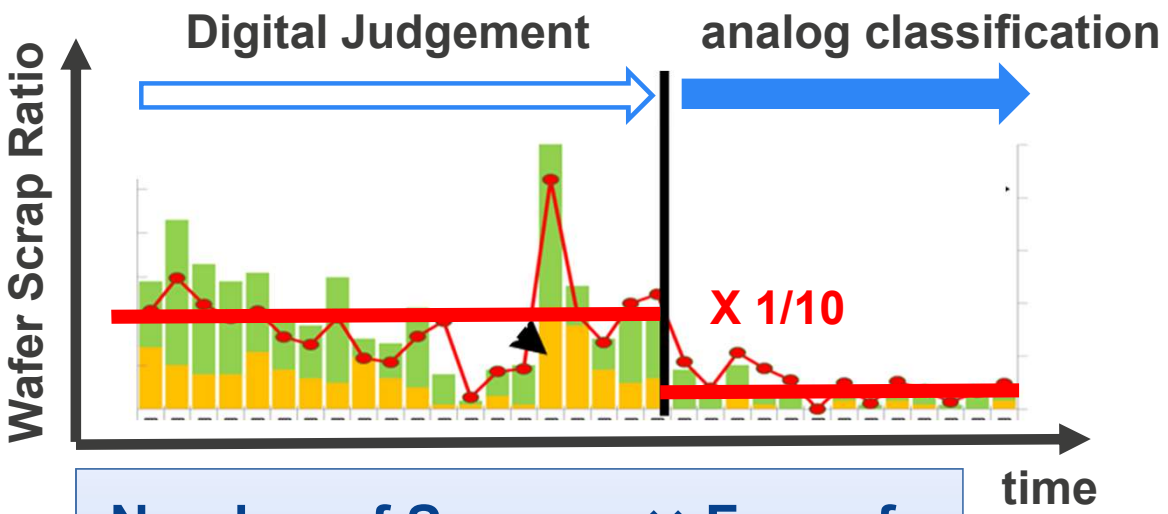


Realtime detection and classification of sudden abnormal to stop movement to next process

Automated FDC system

Predictive Manufacturing

EFFECT OF AUTOMATED FDC SYSTEM



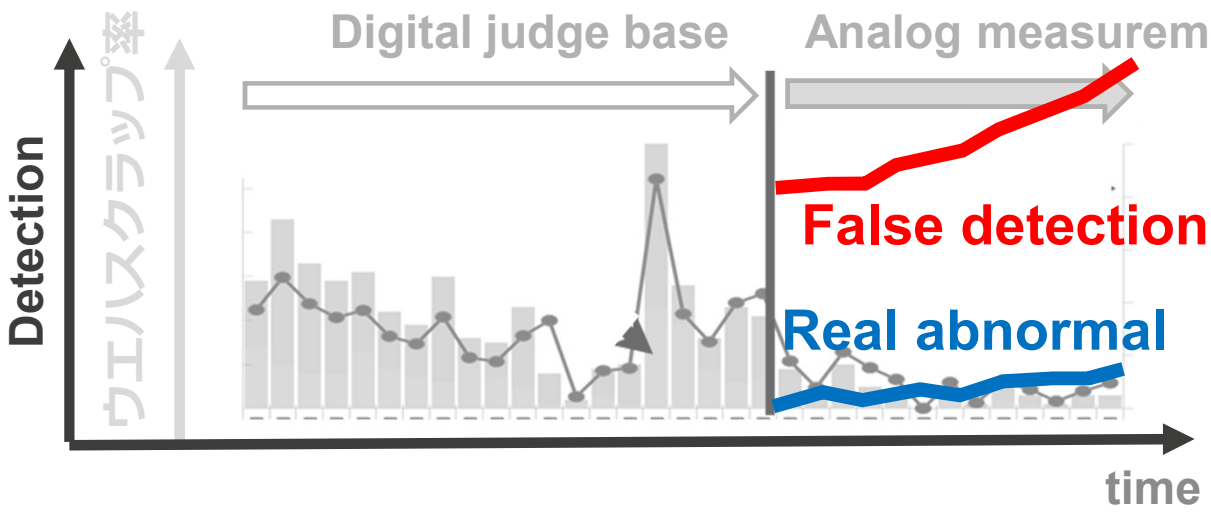
Number of Sensors × Freq of measurement

Digital 20M/day

Analog 10M/day

1/10 reduction of scrap

ISSUE OF FDC AUTOMATED DETECTION



Issues

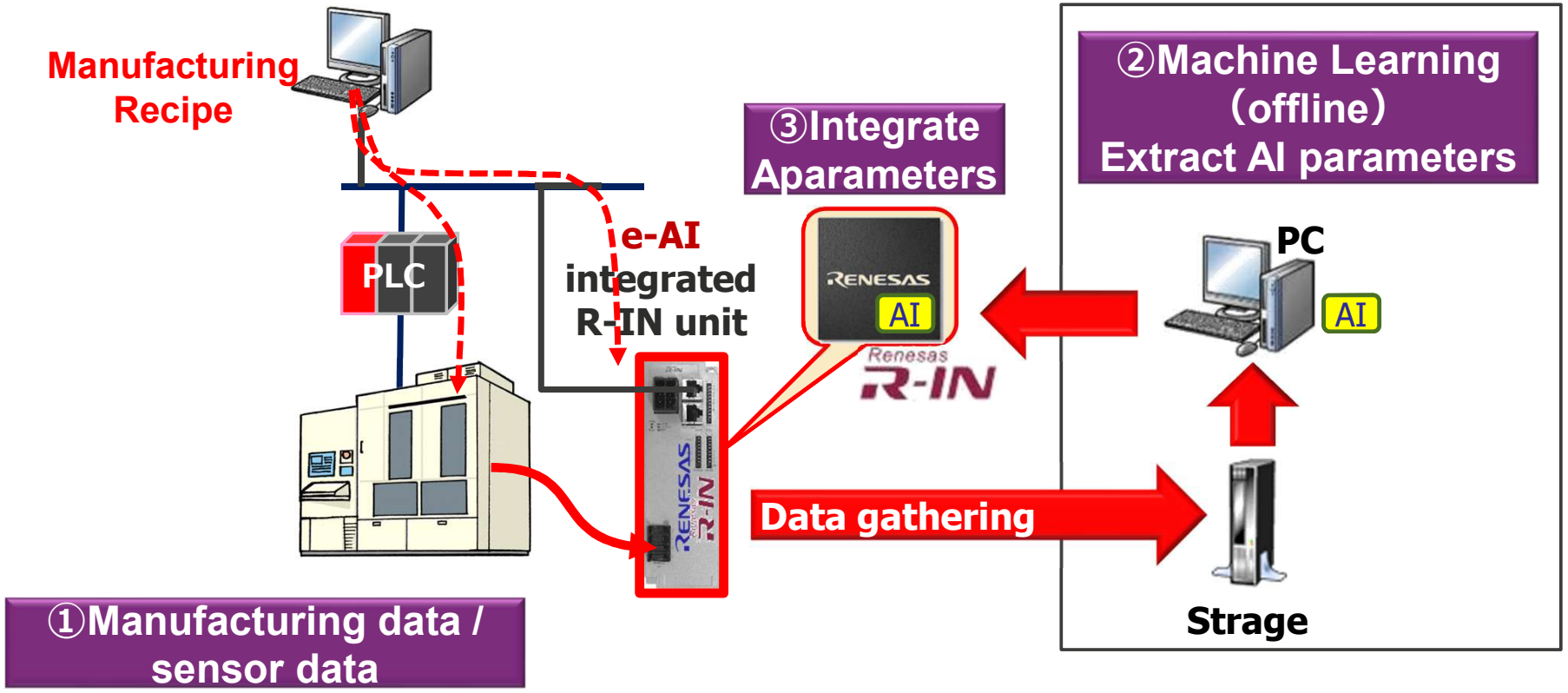
- Staff engineers spend time to derive threshold value
- Staff engineers spend time to distinguish false and real
- Improve detection algorithm to avoid error

Overlooking by improper threshold

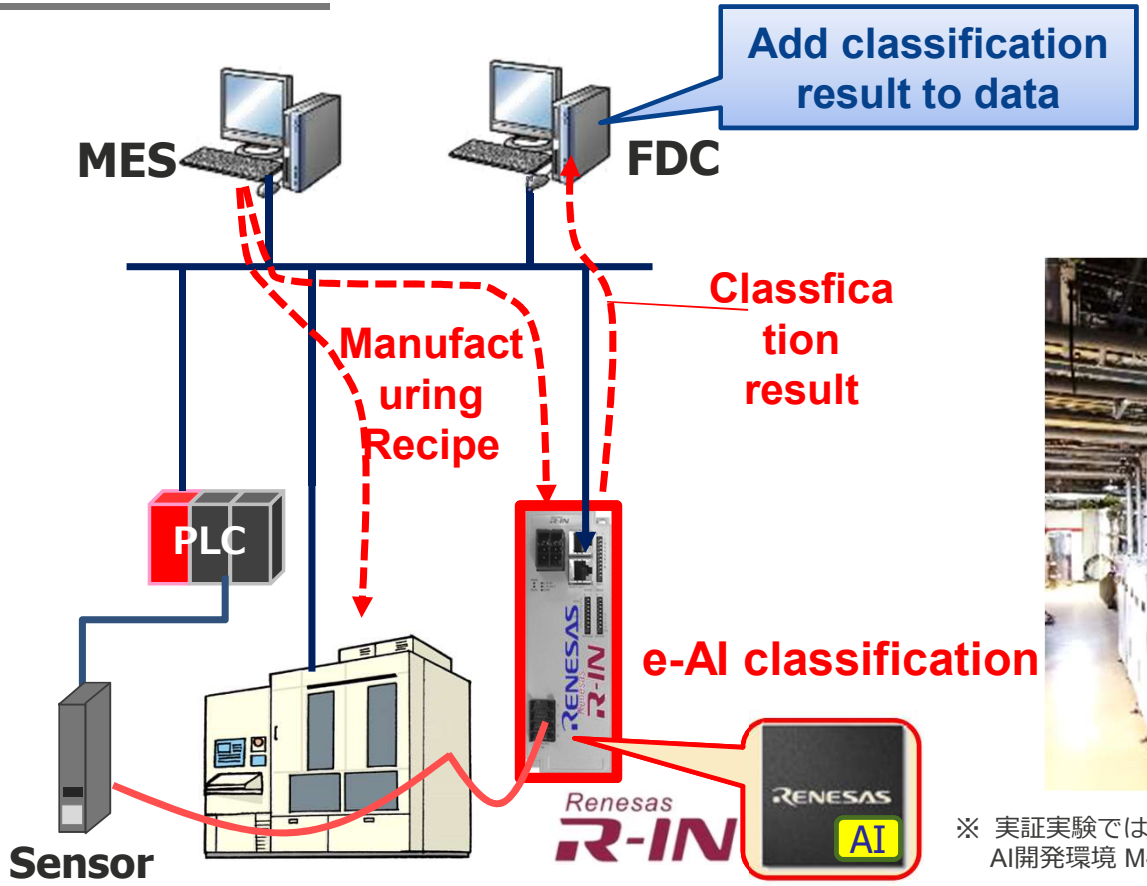
Trade off between error detection and false detection

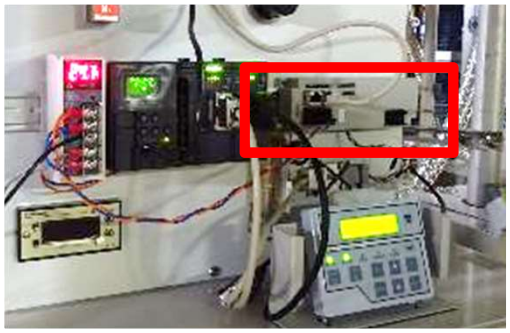
New approach

「E-AI」 LEARNING ①gathering ②Machine Learning ③Integrate parameters



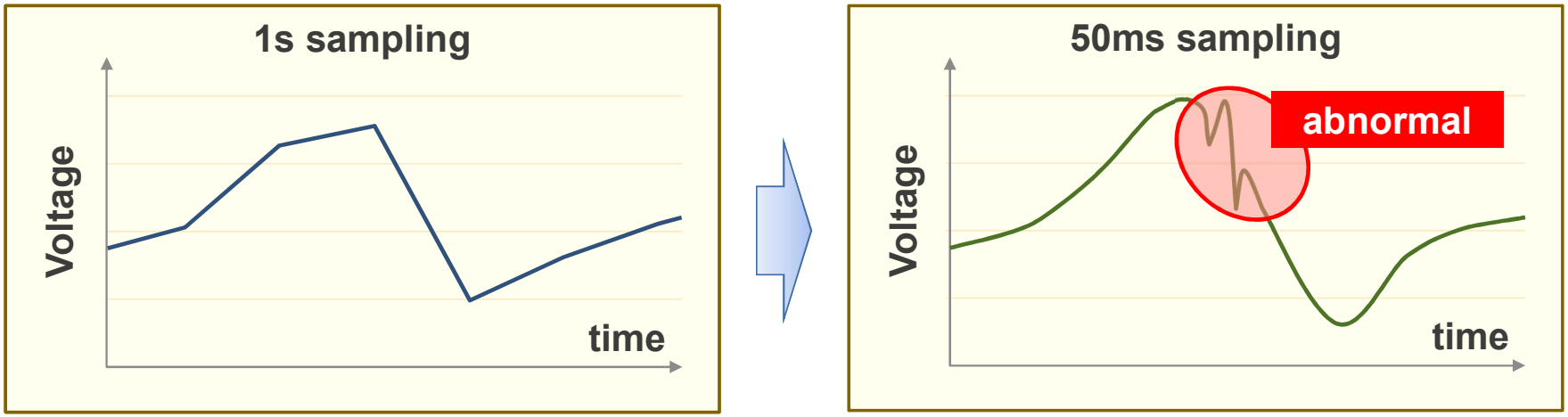
STRUCTURE OF NEW FDC





※ 実証実験では当社パートナー（クロスコンパス・インテリジェンス社）の AI開発環境 M-IX で開発した人工知能を搭載

ADDITIONAL RESULT : HIGHER SAMPLING RATE



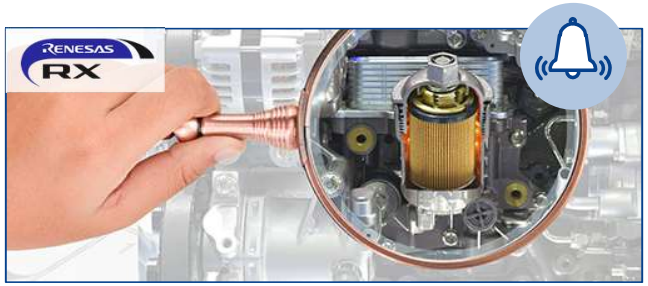
Analog value and higher sampling rate makes Invisible to visible

Can we find more valuable fact by higher sampling rate as 1ms, sub msec ??

USE CASE CLASS-1: e-AI ANOMALY DETECTION
FOR HUNDREDS MILLION MOTORS

Benefits:

- Improve service quality
- Avoid downtimes
- Reduce maintenance cost



e-AI detects and pre-warns anomalies everywhere every time

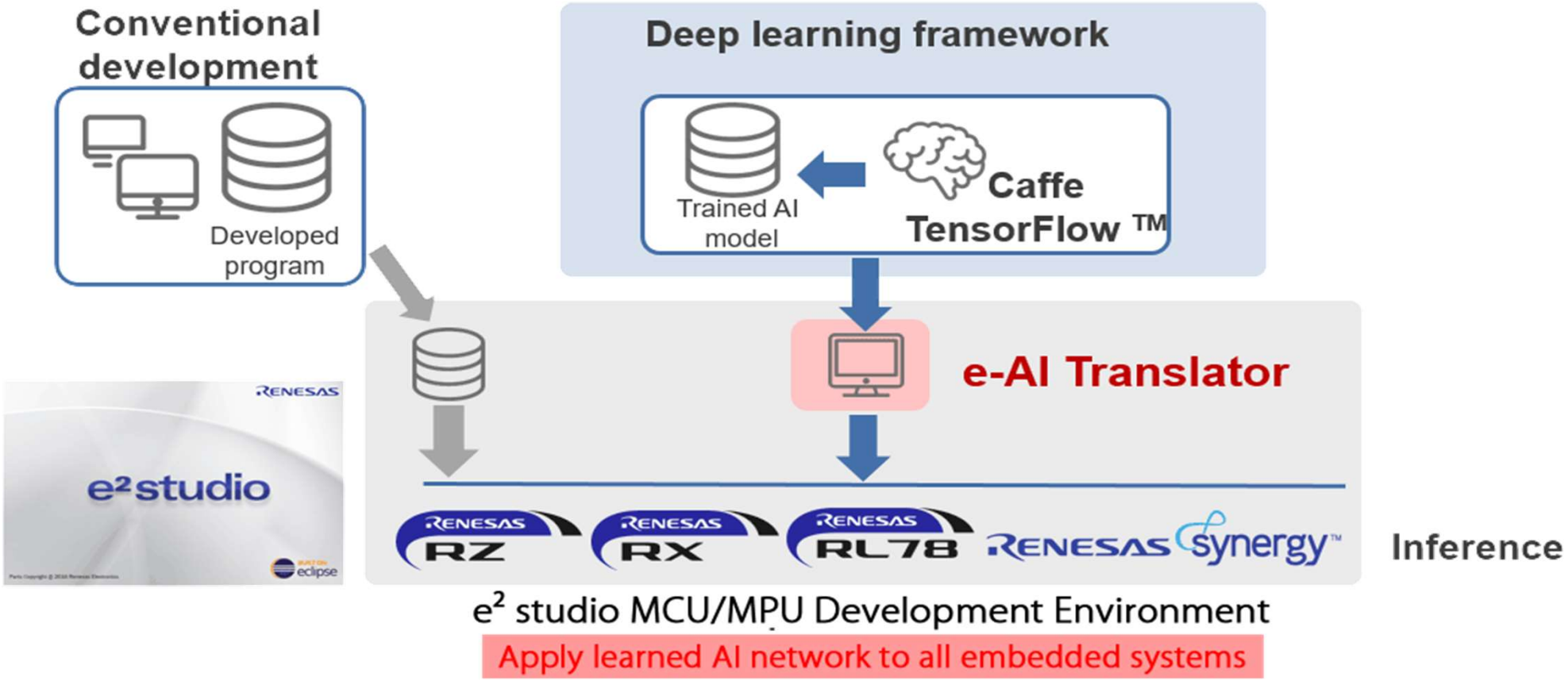
Smart Factory

Smart Living

Smart Infrastructure

Renesas is shipping 200M+ motor control MCU per year.
New product series “RX66T”(Nov.) will enable e-AI Anomaly Detection.

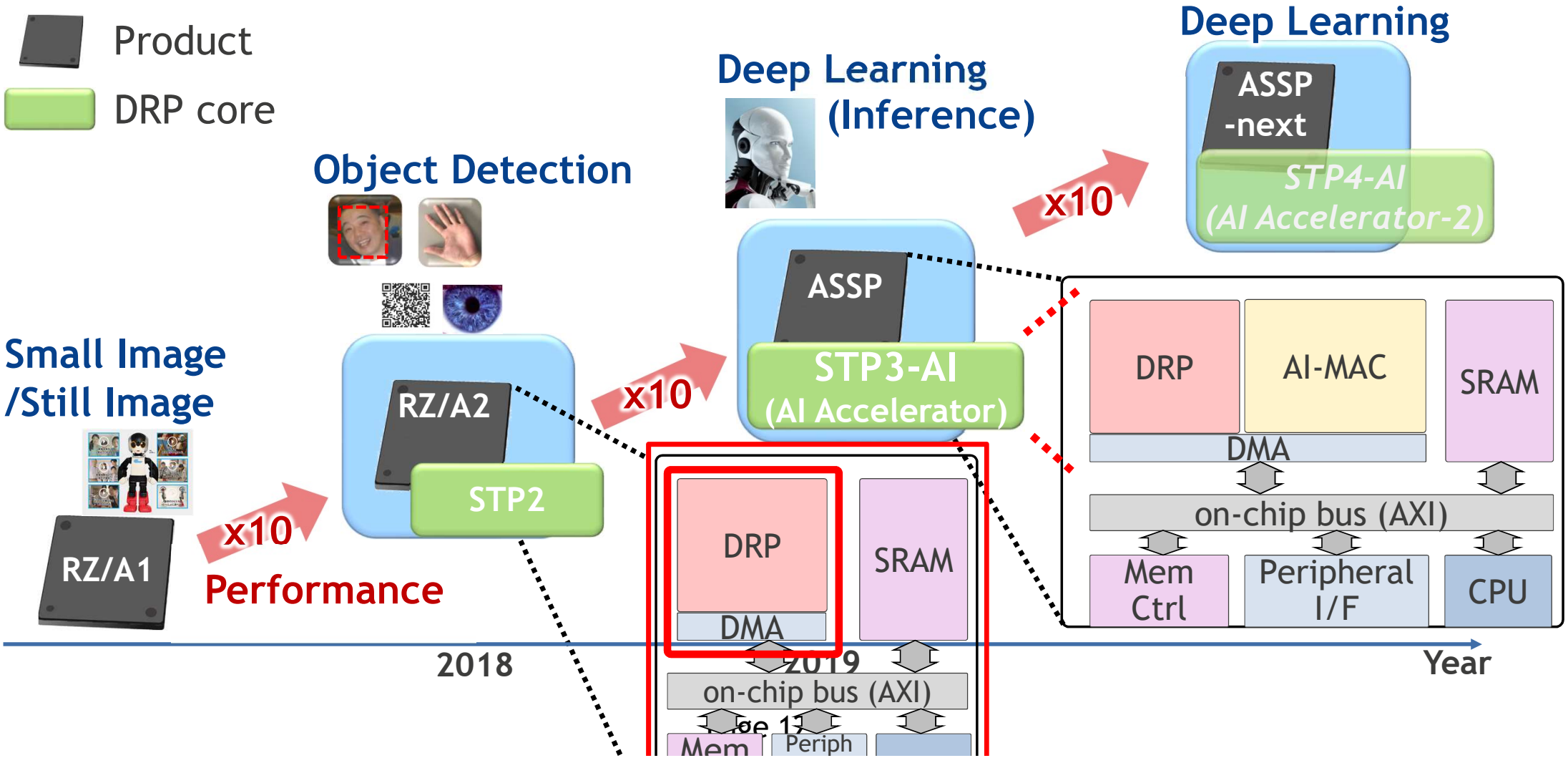
Renesas e-AI Solution



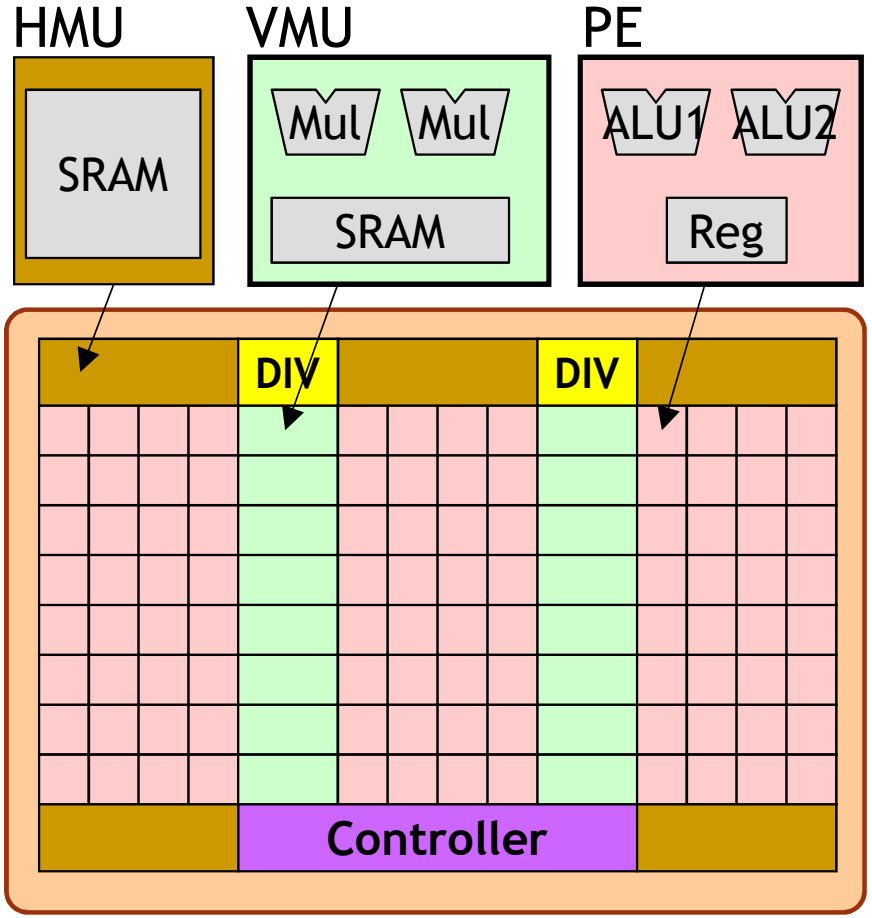
Significant performance improvement under low power and reduction of memory usage are required for the factory.

TARGET

`10x Performance' for End-point Intelligence



DRP: Dynamically Reconfigurable Processor



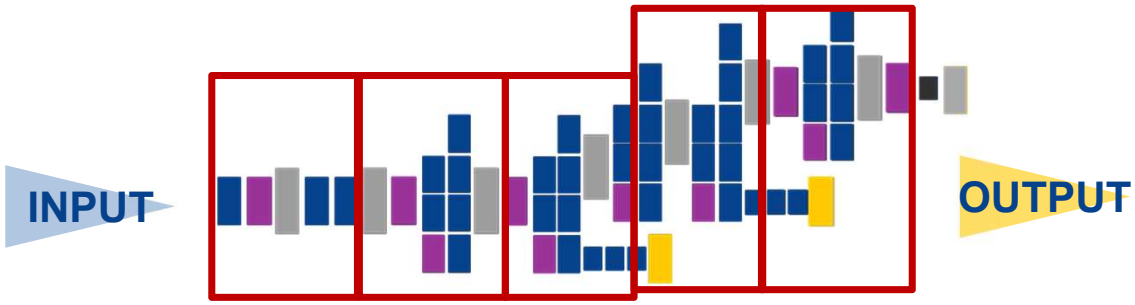
DRP (3rd gen. 96PE)

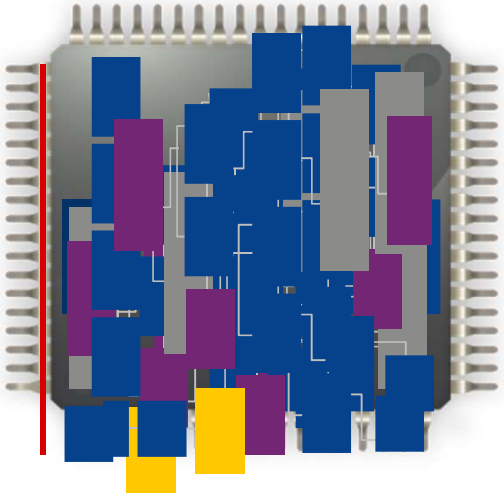
DRP: State Transition Controller (STC)
+ Coarse grained reconfigurable data-path

Data-path Resources (16bit width)

- Array of ALU:
  - 96x PE
- Linear memory and multiplier units:
  - 16x VMU (512word 2port mem, 2x {MUL16, FP16 ALU})
  - 10x HMU (8Kword 2port mem)
  - 2x {DIV32, FP32 ALU}
- Programmable Wires (not shown)

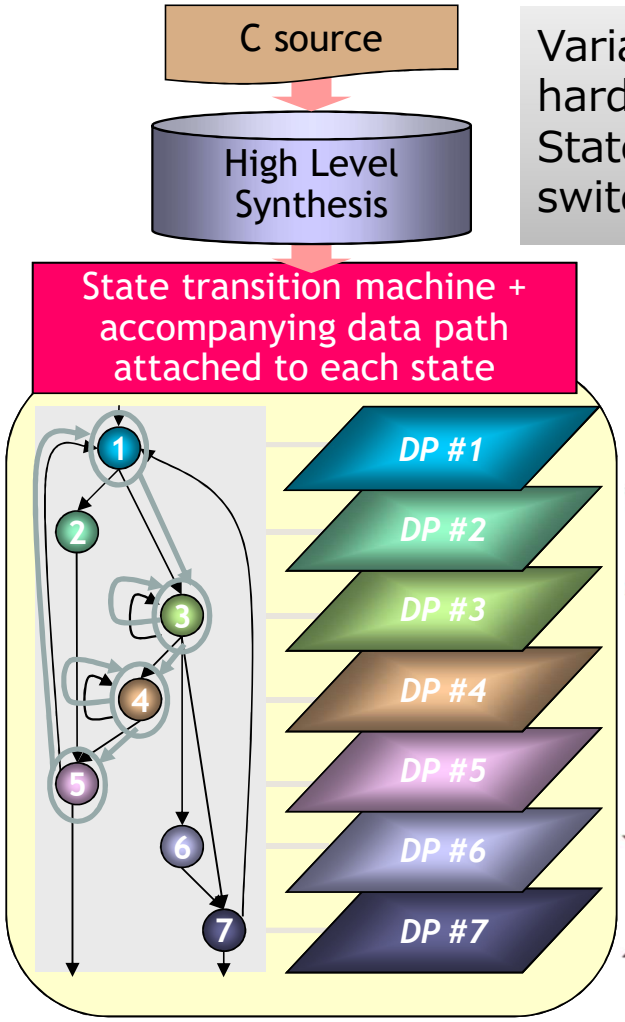
e-AI Accelerator
DRP: Dynamically Reconfigurable Processor





*DRP; Dynamically Reconfigurable Processor

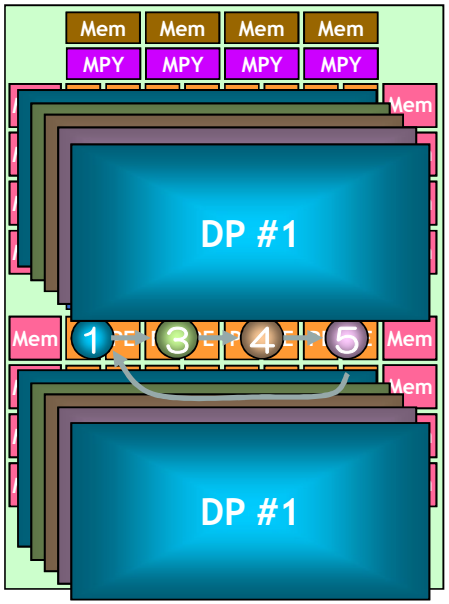
DRP Mechanism



Variables and operators in C language dedicated compiler maps to hardware resources.
State Transition Controller (STC) realizes an arbitrary function by switching the connection therebetween every one clock.

Realized in architecture

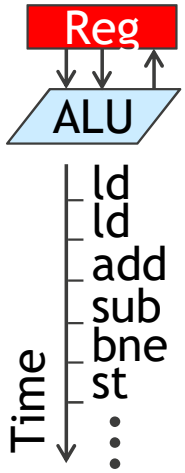
Actual operation is :



STC that controls state transition
+
Arithmetic unit array having configuration information of plural planes switchable in 'ns'

Spatial + Time-multiplexed computing

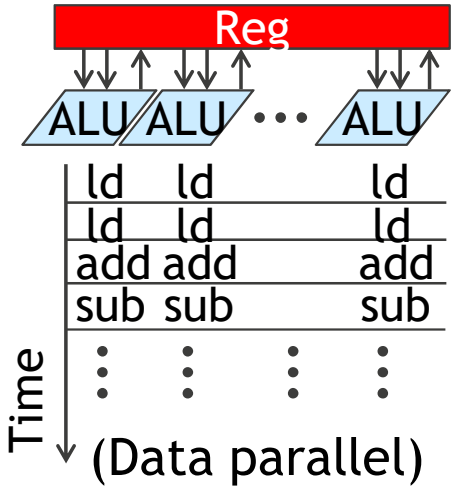
CPU



(Sequential)

- High flexibility
- Limited Performance/core
- High frequency

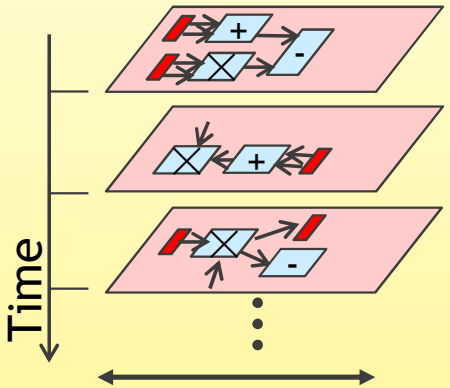
GPU



(Data parallel)

- High performance
- Vector data processing

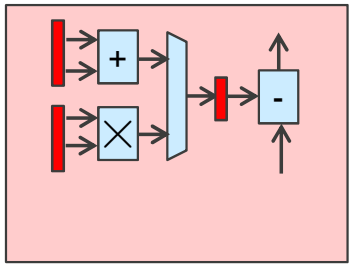
DRP



Spatially pipelined
[ALU array]

- High performance
- Area efficiency

FPGA



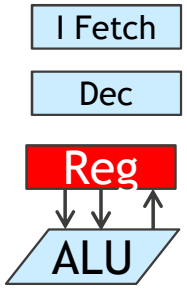
Spatially Pipelined
[LUT array]

- High performance
- Large area
- Partly time multiplex

Time-multiplexed ⟷ Spatial

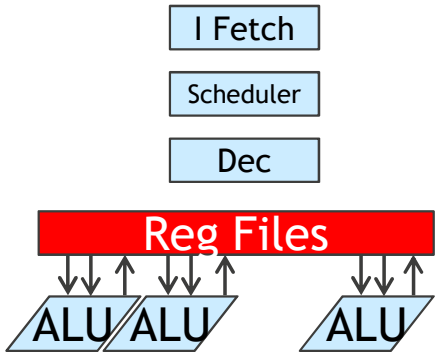
Control vs Data Processing

CPU



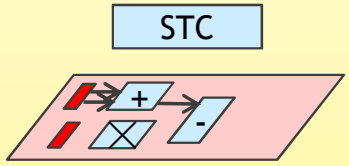
Complex instruction processing
- I cache
- Branch predict
- Pipelining

GPU



- SIMD(SIMT)
- Sophisuticated scheduler & dispatcher

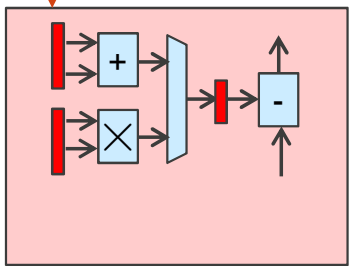
DRP



- STC stores multiple configuration and sequence information
- STC control data can load on interval

FPGA

Load bit stream on boot



Spatially Pipelined [LUT array]

- Configuration information is stored in array

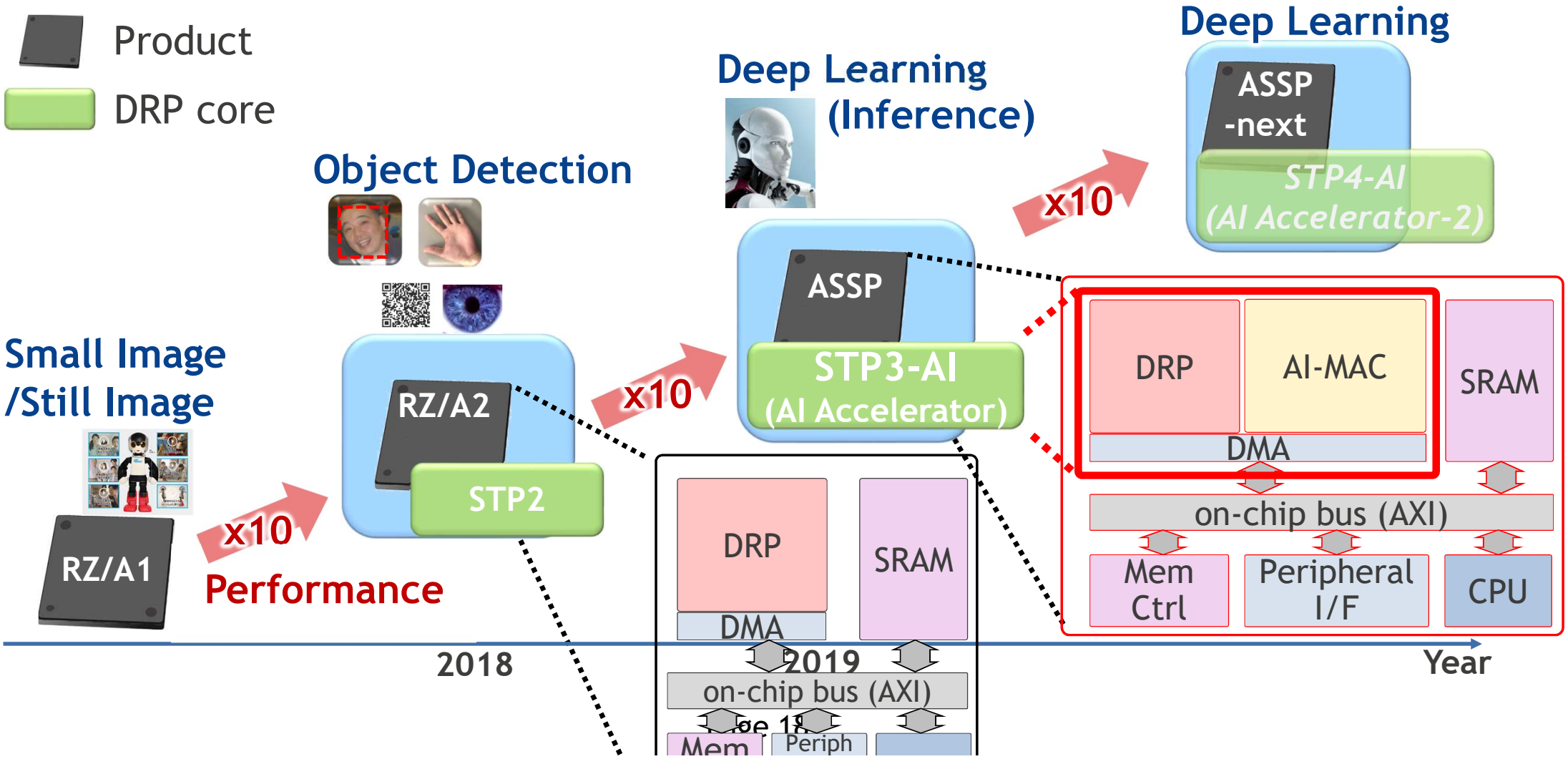
Flexible
Coarse

Flexibility of configuration
Granularity of configuration

Fixed
Fine

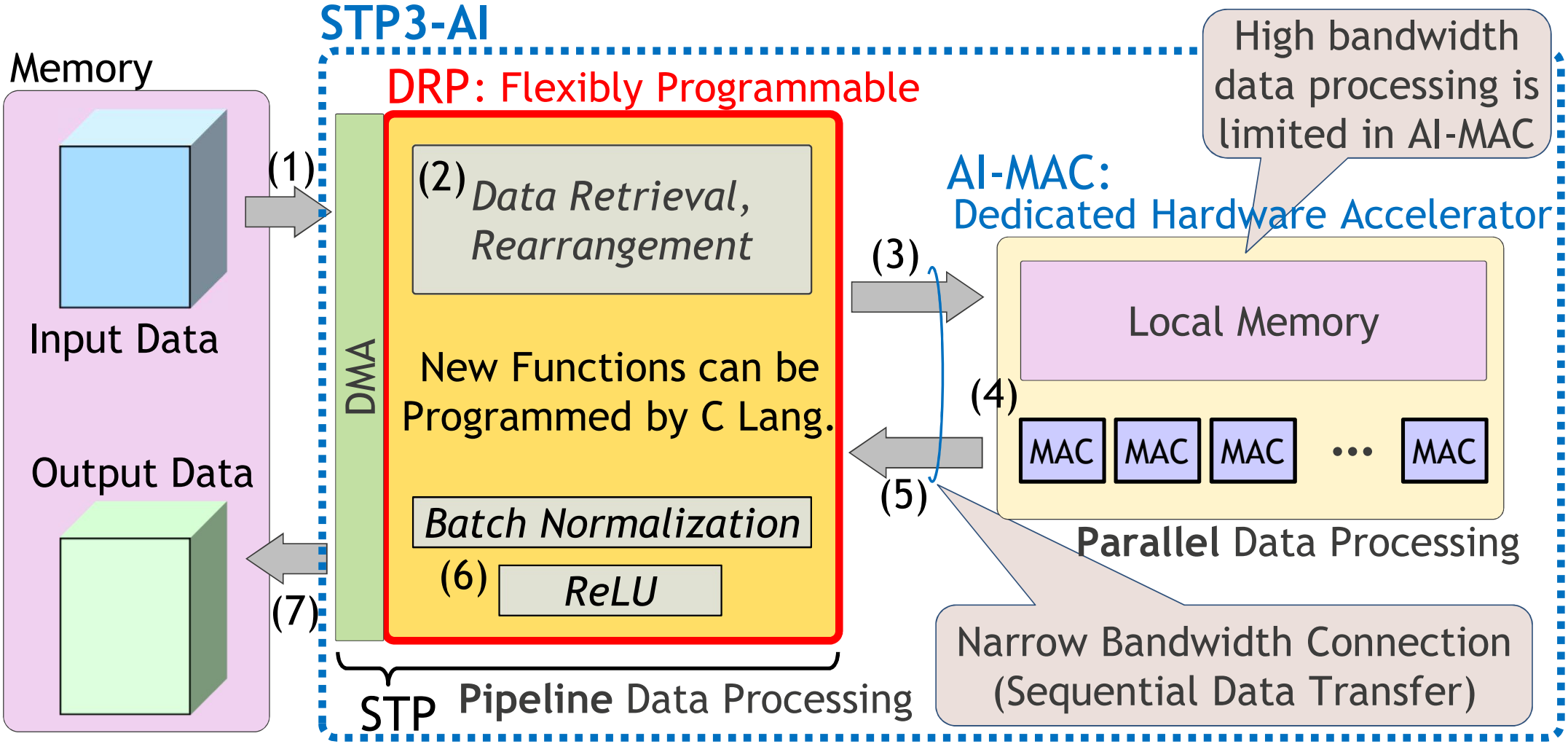
TARGET

`100x Performance' for End-point Intelligence

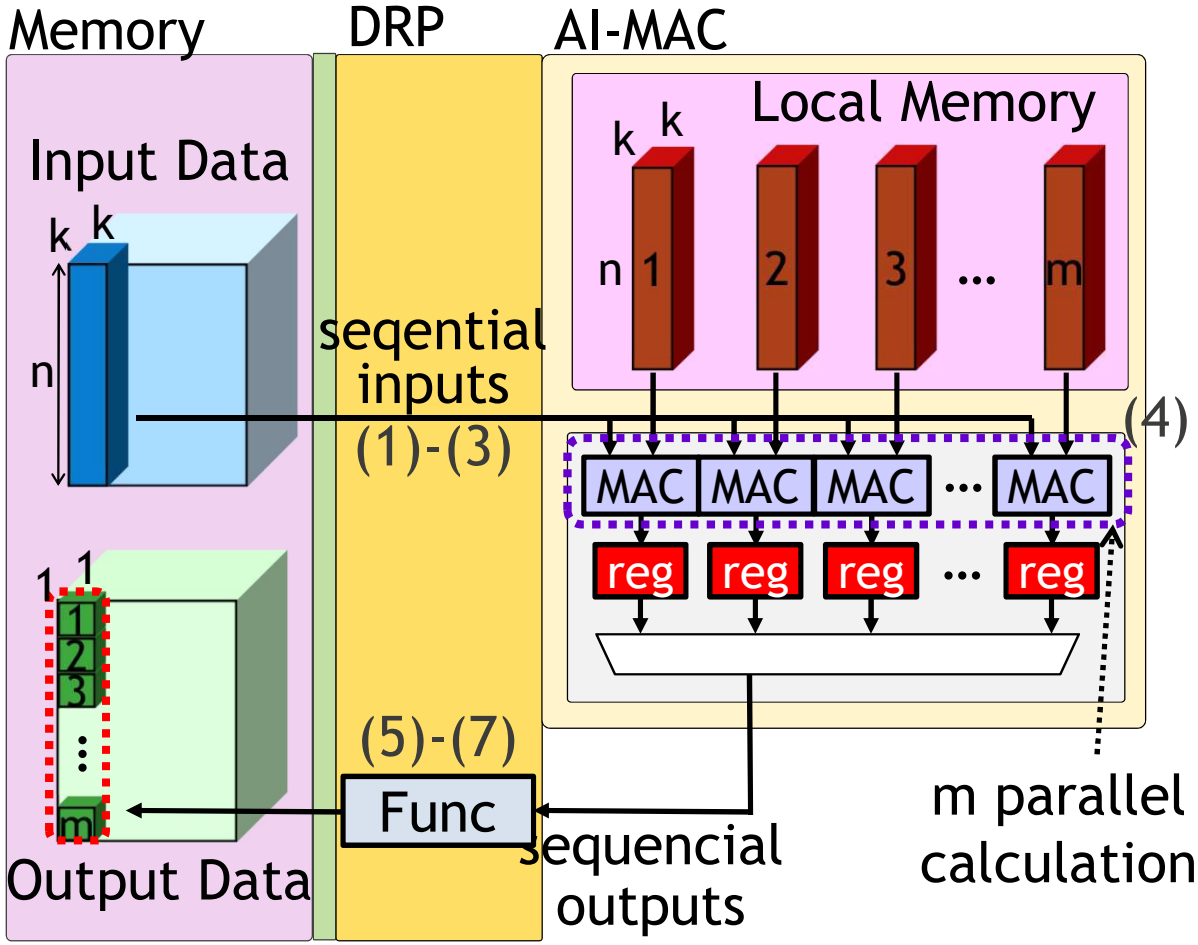


STP3-AI Architecture

( ): Data Flow Order

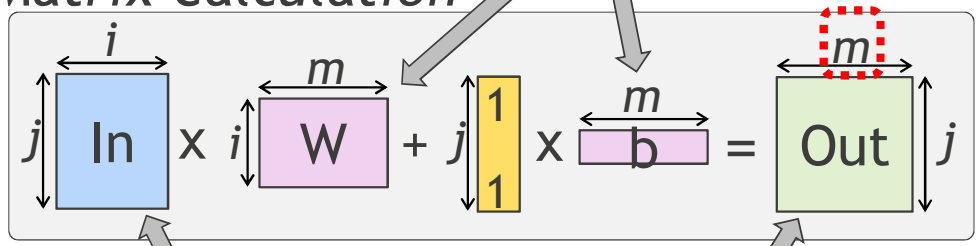


AI-MAC Data Processing



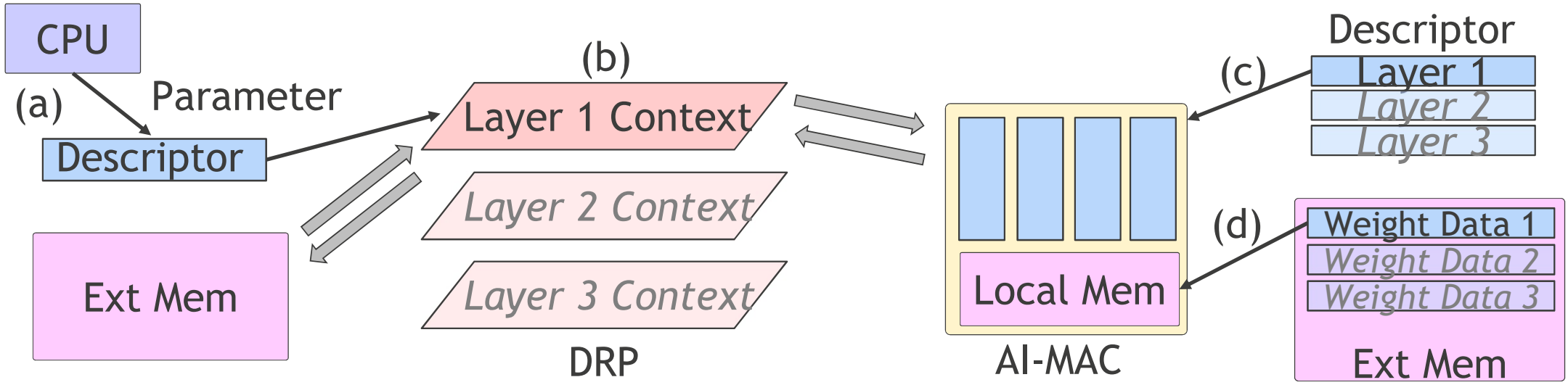
Fixed Data
(Stored in Local Memory)
・Trained weight parameter

Matrix Calculation



Stream Data
(DRP send/receive sequentially)
・Layer Input/Output data

Multi Layer Processing: Layer 1



(a) CPU prepares a DRP descriptor

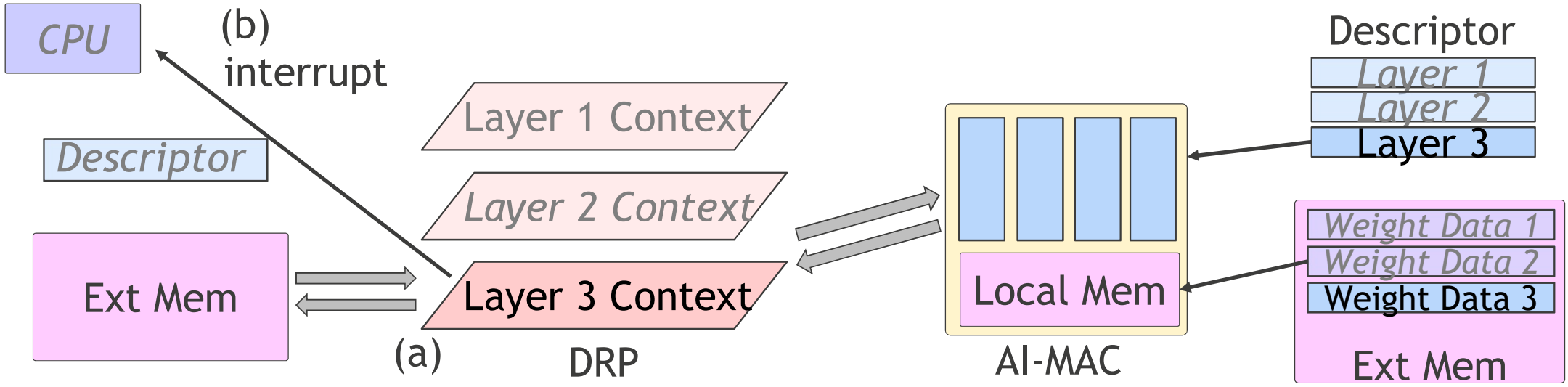
(b) DRP starts processing the Layer 1

– By using optimized Layer 1 context

(c) AI-MAC is configured by AI-Mac Layer 1 descriptor

(d) Weight data 1 is transferred to prepare the calculation

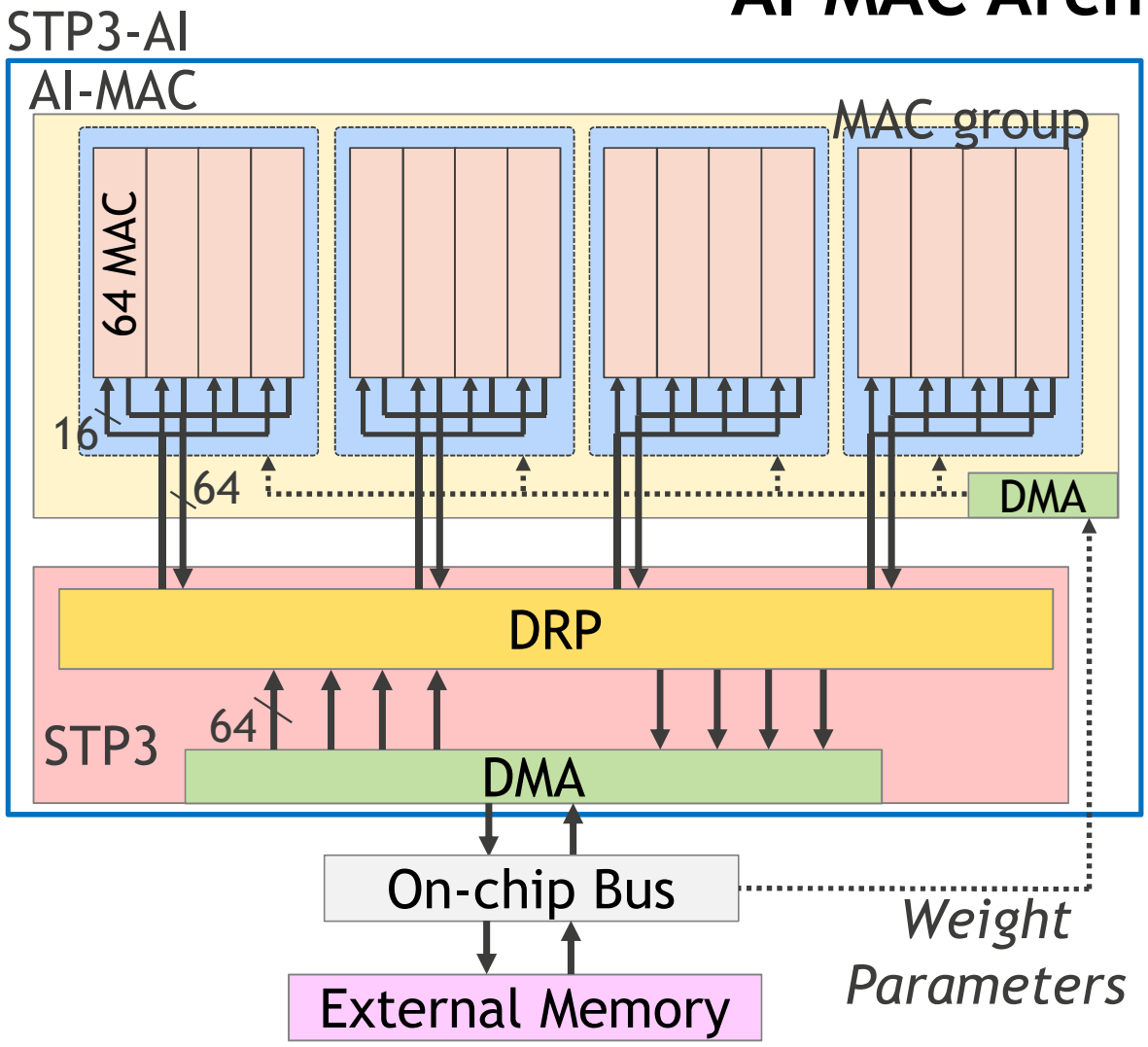
Multi Layer Processing: Layer 3



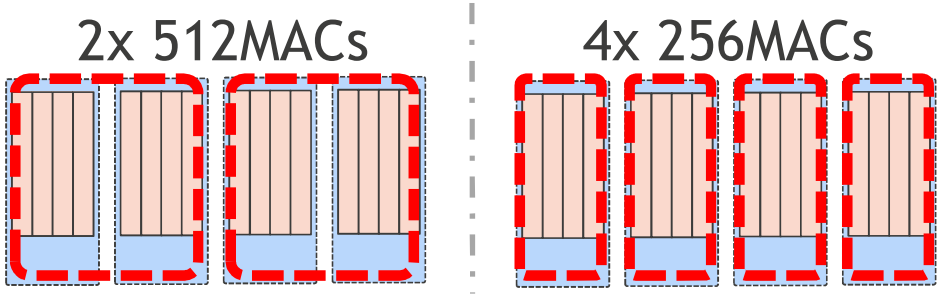
(a) DRP repeats dynamic reconfiguration and layer processing.

(b) Finally, DRP generate an interrupt signal to CPU to notice the end of all layer processing.

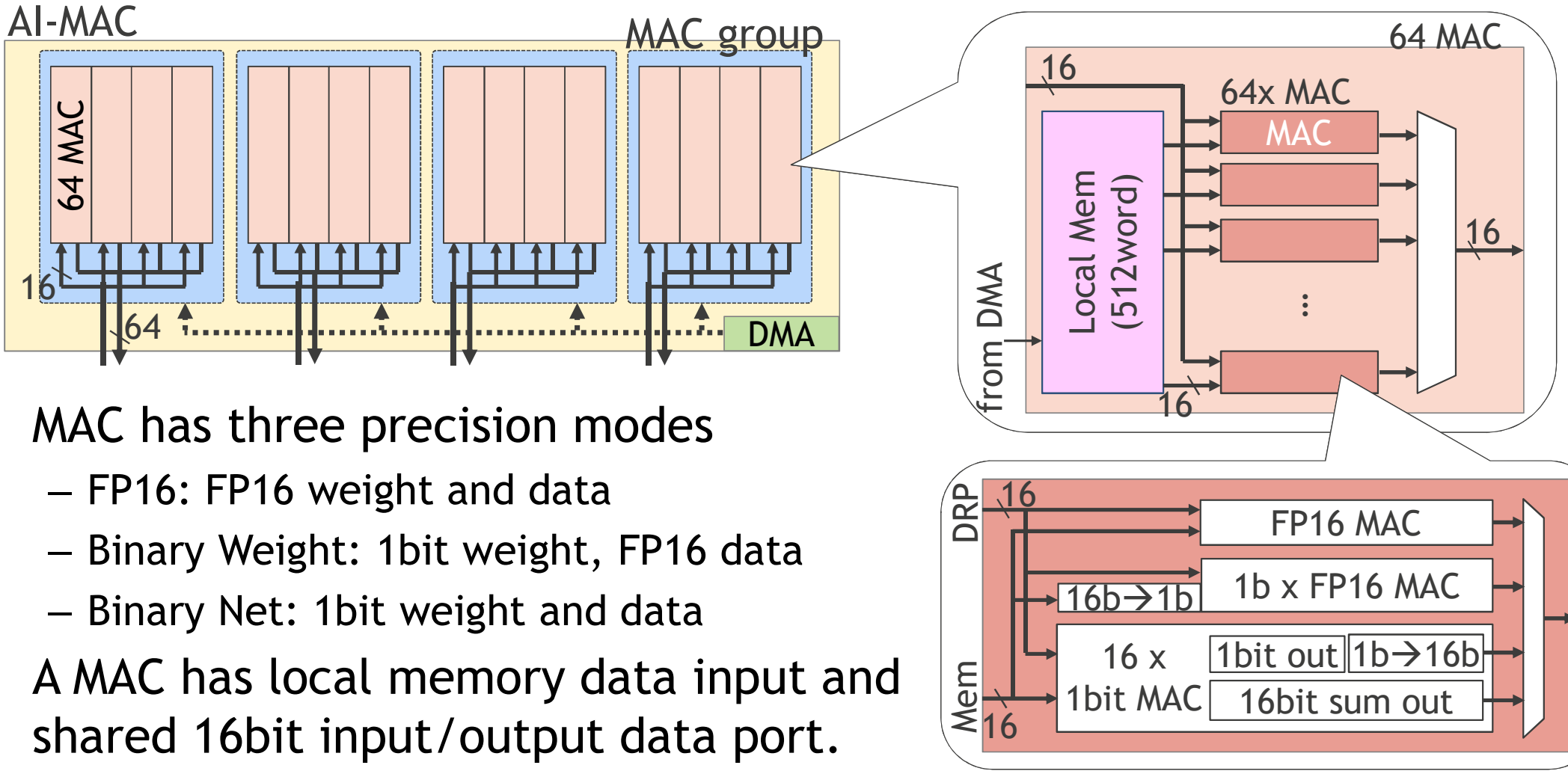
AI-MAC Architecture



- AI-MAC has 1024 MACs
  - MACs are Divided into four MAC groups.
  - A MAC group has four of 64-MAC.
- MAC group is connected with DRP by 64bit in/out data port.
- AI-MAC can handle up to four independent matrix.

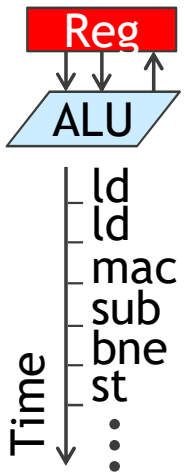


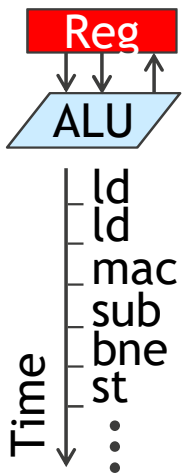
AI-MAC Architecture



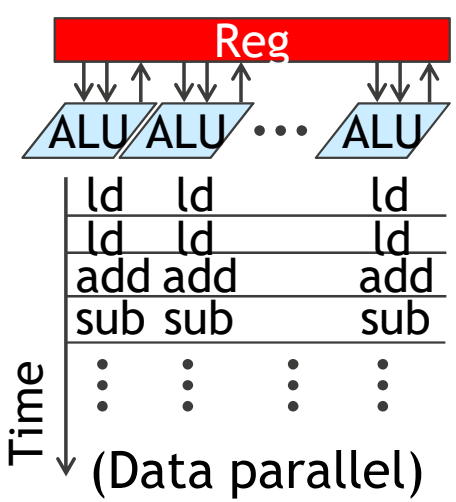
- MAC has three precision modes
  - FP16: FP16 weight and data
  - Binary Weight: 1bit weight, FP16 data
  - Binary Net: 1bit weight and data
- A MAC has local memory data input and shared 16bit input/output data port.

Machine Learning, Deep Learning adaption

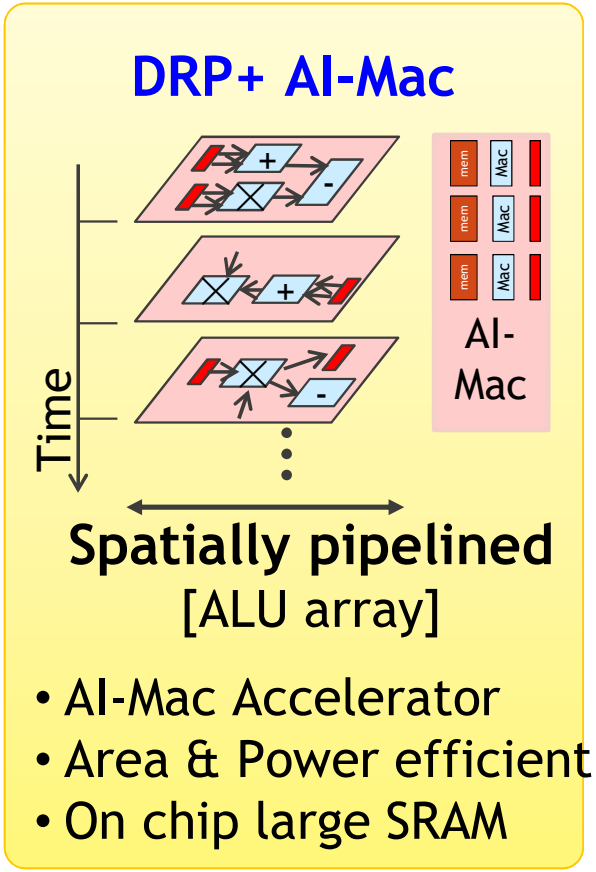


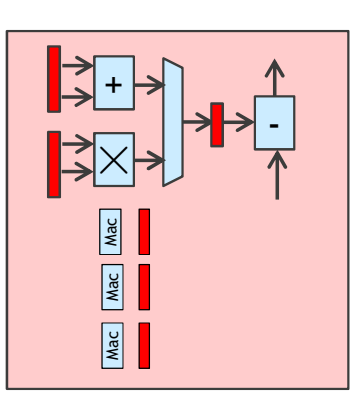


Many cores
NN ISA extension
NN accelerator



More cores
Tensor Core

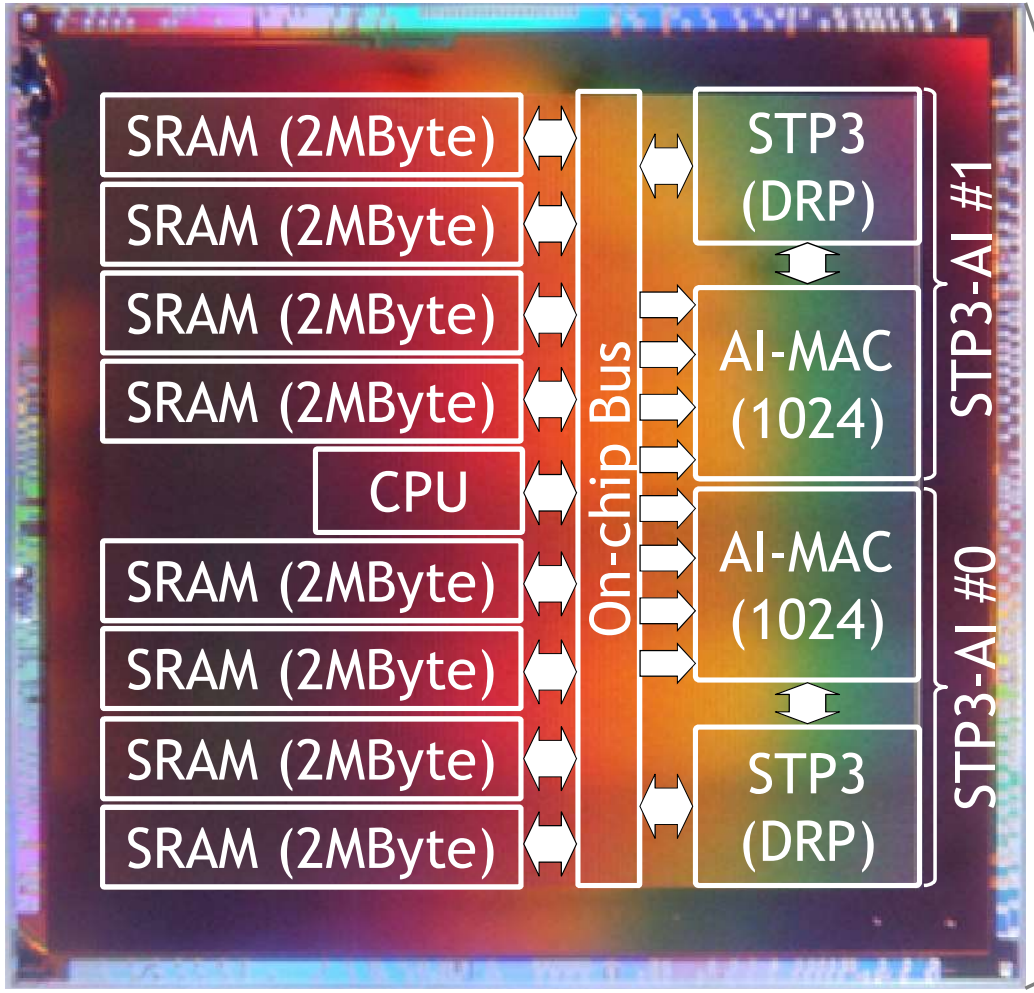




- Sophisticated DSP array
- AI Mac systric array

AI-Mac

Prototype Chip



Chip Specification

Technology: UMC 28HPC, 10-Metal

Package: 324pin BGA

Core: Dual STP3-AI, ARM CPU, etc.

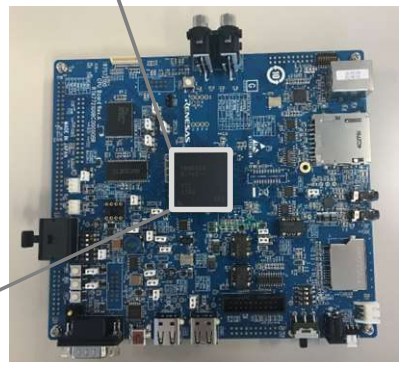
On-Chip Memory: 128Mbit SRAM

AI-MAC: max 500MHz

DRP: max 333MHz

One STP3-AI core peak performance (designed): 1,000 GOPS

 # Built upon micro-computer chip



Measured 960 GOPS

Evaluation Board

Evaluation (Some VGG Layers)

Layer	Data Size (in, out)	Mode	Batch	[GOPS]	Bottleneck
Convolution, 3x3	56 x 56 x 256ch	FP16	1	524.5	MAC Usage Rate (DRP Program)
Convolution, 3x3	28 x 28 x 512ch	FP16	1	754.5	
Convolution, 3x3	28 x 28 x 512ch	Binary Weight	1	754.5	
Full Connection	4096	FP16	1	19.3	Weight Data Transfer Time
Full Connection	4096	FP16	16	268.9	
Full Connection	4096	FP16	64	687.1	
Full Connection	4096	Binary Weight	1	461.5	

*One STP3-AI core used, Exclude initial mem transfer time , **Ideal: 1,000GOPS***

- Convolution throughput depends on the utilization rate of 1024 MACs.
  - Data processing design and DRP program determine.
- Full Connection throughput depends on weight memory transfer time from ext. mem.
  - Binary Weight mode (1/16 weight data size) or batch data processing are effective.

Conclusion

- There’s a certain application to match various range of ML performance. In smart factory and other embedded segment have good opportunity to utilize relatively lower performance Machine Learning solution.
- Dynamic reconfigurable Processor architecture can provide 10x performance acceleration compare to standard MCUs for AI inference.
- Dedicated AI-Mac accelerator and DRP architecture can improve another 10x boost of AI inference performance.
- Part of these performance improvement came from limited flexibility of programing, less generality compare to generic CPU or GPU. But we think such restriction will not harm the adaption for future Machine Learning application.

Thank you for your attention

Atsushi Hasegawa
Atsushi.hasegawa.gx@Renesas.com
IoT and Infrastructure Business Unit
Renesas Electronics Corporation