An Energy-efficient Floating-Point DNN Processor using Heterogeneous Computing Architecture with Exponent-Computing-in-Memory

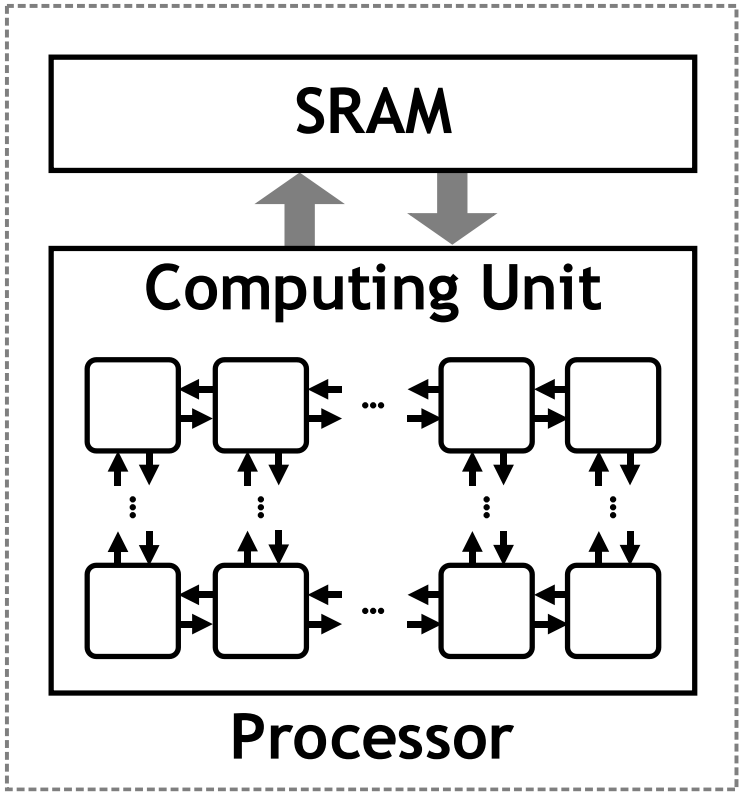
Juhyoung Lee, Juhyoung Lee, Jihoon Kim, Wooyoung Jo,
Sangyeob Kim, Sangjin Kim, Donghyeon Han, Jinsu Lee and Hoi-Jun Yoo

Semiconductor System Lab.
School of EE, KAIST

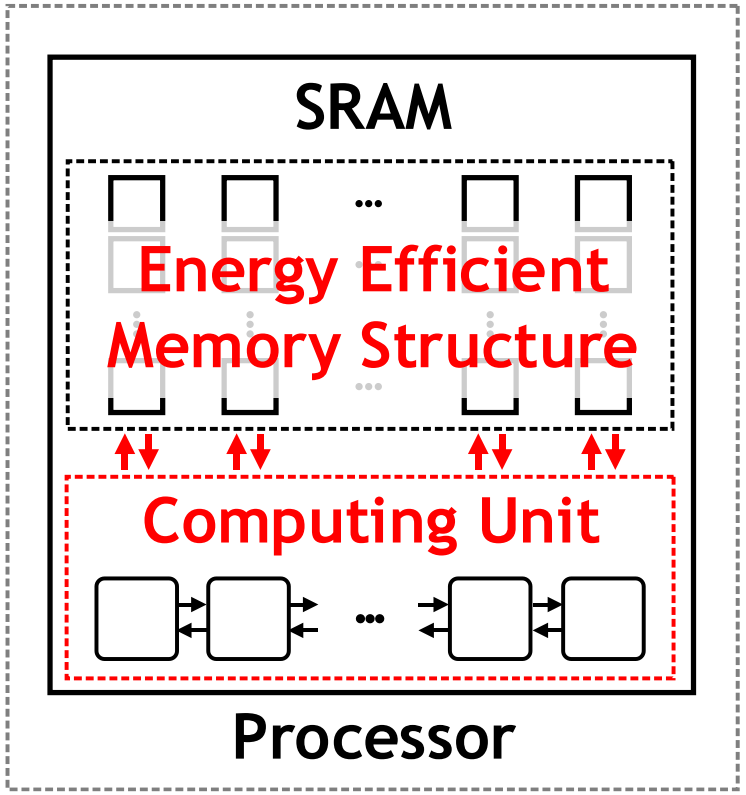
Computing-in-Memory (CIM) Design

- **Moving Computation into Memory**

<Von-Neumann Arch.>

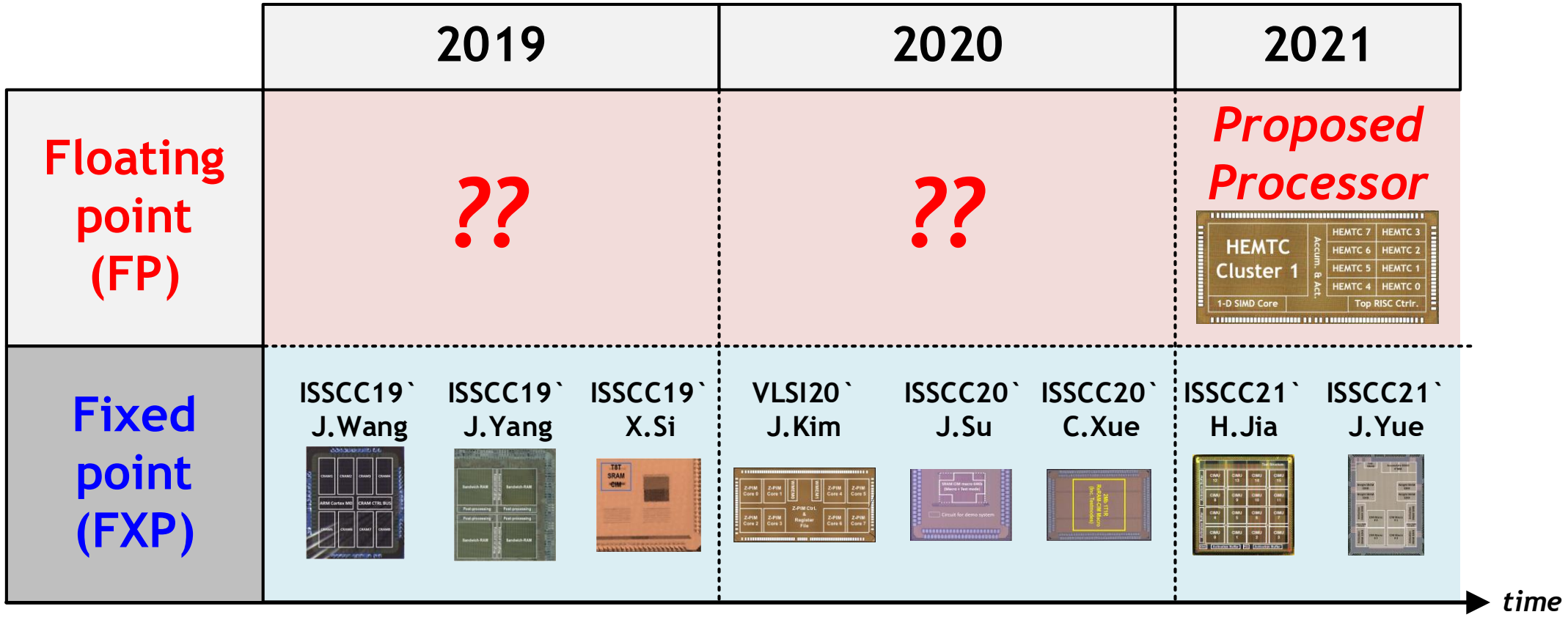


<Computing-in-Memory Arch.>



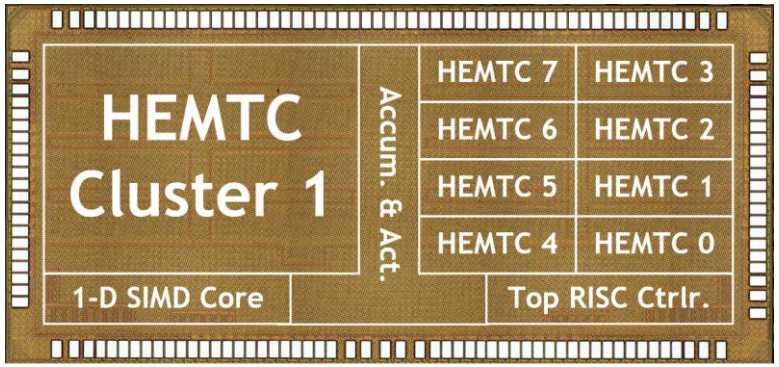
Limitation of Previous CIM Hardwares

- **Cannot Support Floating-point (FP) Data Representation**
  - Achieved SOTA energy efficiency, but no fixed-point (FXP) precision



Abstract of Proposed FP CIM Processor

<Chip Photo & Spec>



Technology	28nm Logic CMOS
Die Area	1.62 mm × 3.6 mm
Precision	BFloat16
Energy Efficiency [TFLOPS/W]	1.43* – 13.7** @ (40MHz, 0.76V)

* = Activation Sparisty 0%, ** = Activation Sparisty 90%
(1 MAC = 2 OP)

<Key Feature List>

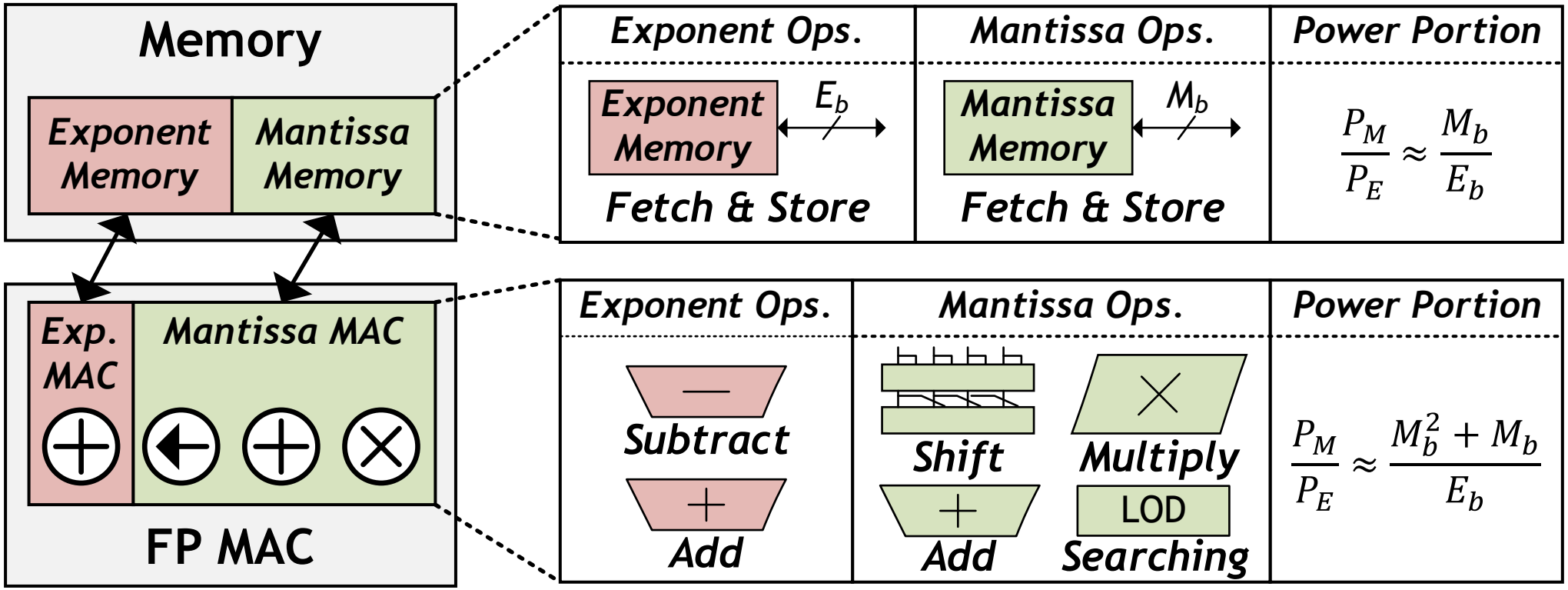
(1) Heterogeneous FP Computing Arch.
: Separate optimization of FP computing
: Realize 2 cycles FP MAC w/ CIM ☺

(2) Exponent Computing-in-Memory
: In-memory AND/NOR + BL charge reusing
: Total memory power 46.4 % ↓

(2) Mantissa Free Exponent Calculation
: Removing redundant normalization
: Total MAC power 14.4 % ↓

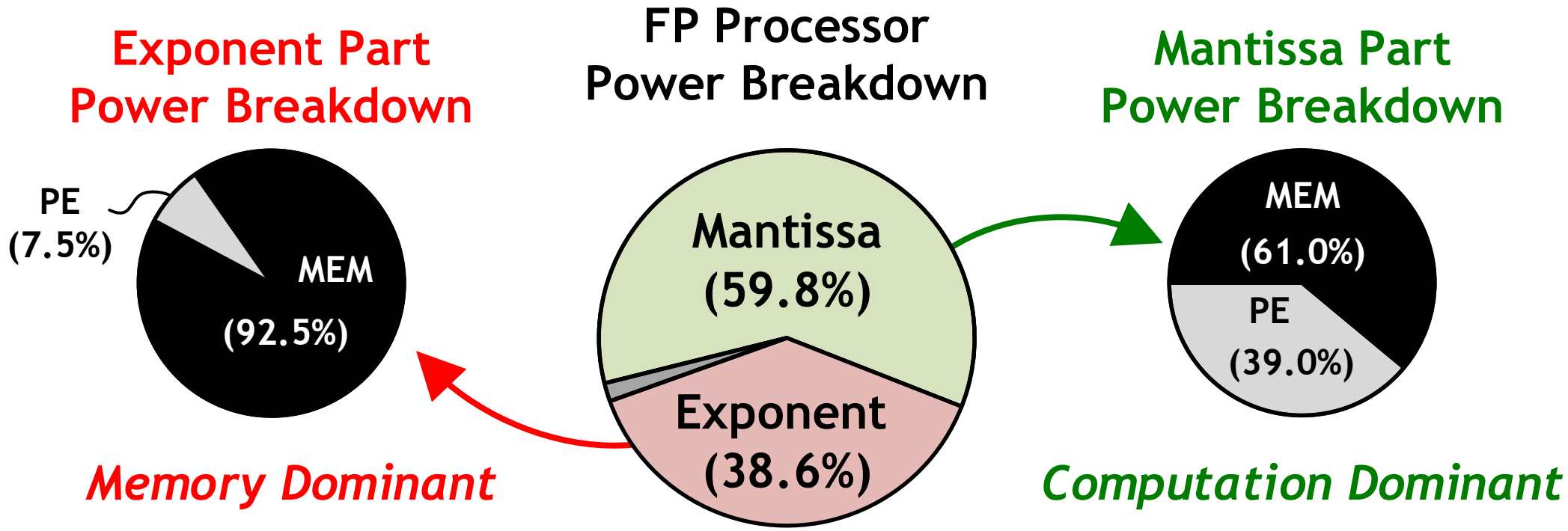
Floating-point (FP) Computing System

- **Memory Part : Same Op. for Mantissa and Exponent**
- **MAC Part : Complex Mantissa Op. ⇔ Simple Exponent Op.**



Heterogeneous Characteristics!

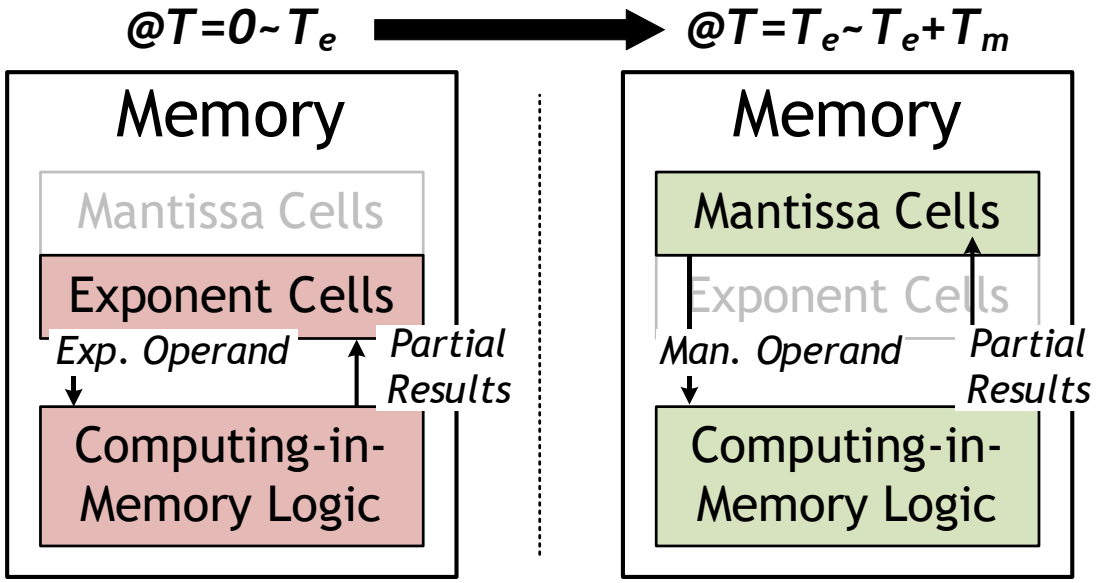
- **Exp. : memory Intensive ⇔ Man. : computation Intensive**
  - Hard to optimize both mantissa and exponent w/ CIM at once! ☹



Previous FP CIM Architecture*

- **Unified FP Computing-in-memory Architecture**
  - Complex mantissa op. ➔ Repetitive simple op. ➔ **Throughput ↓ ☹**

<Prev. FP Computing Arch.>



<Operation Cycles Analysis>

1) BFP16 Multiplication

: Mantissa Mult. ➔ Repetitive Addition

> 100 cycles required

2) FP32 Accumulation

: Mantissa Shift, Mantissa Norm. ➔ Repetitive Addition

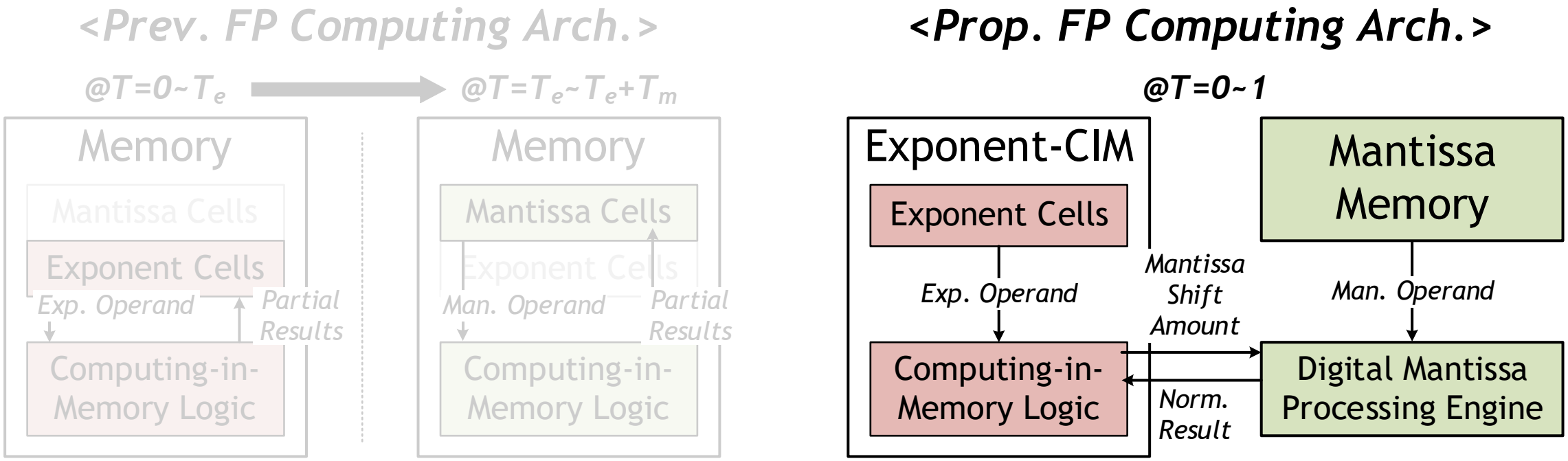
> 4900 cycles required

Total > 5000 cycles required ☹

** J. Wang et al., "A Compute SRAM with Bit-Serial Integer/Floating-Point Operations for Programmable In-Memory Vector Acceleration," ISSCC 2019.*

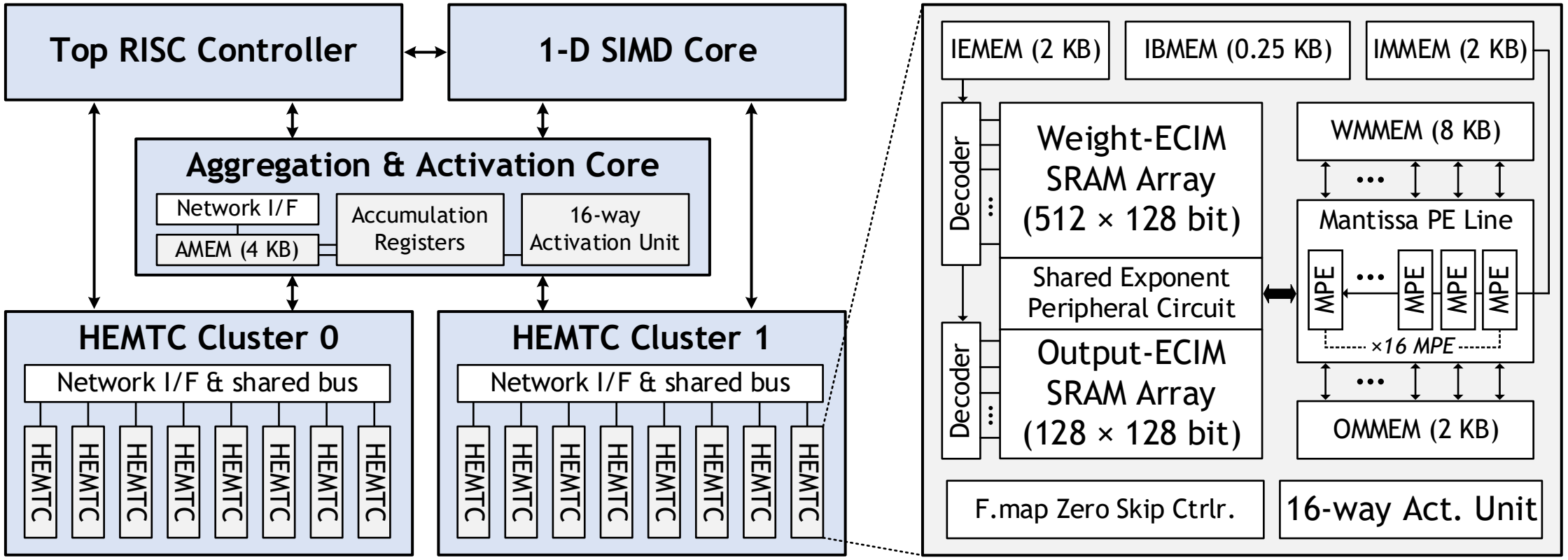
Proposed Heterogeneous FP CIM Arch.

- **Only exponent operations optimized w/ CIM**
  - Cycles for FP MAC: >5000 cycles ➔ **2 cycles ☺**



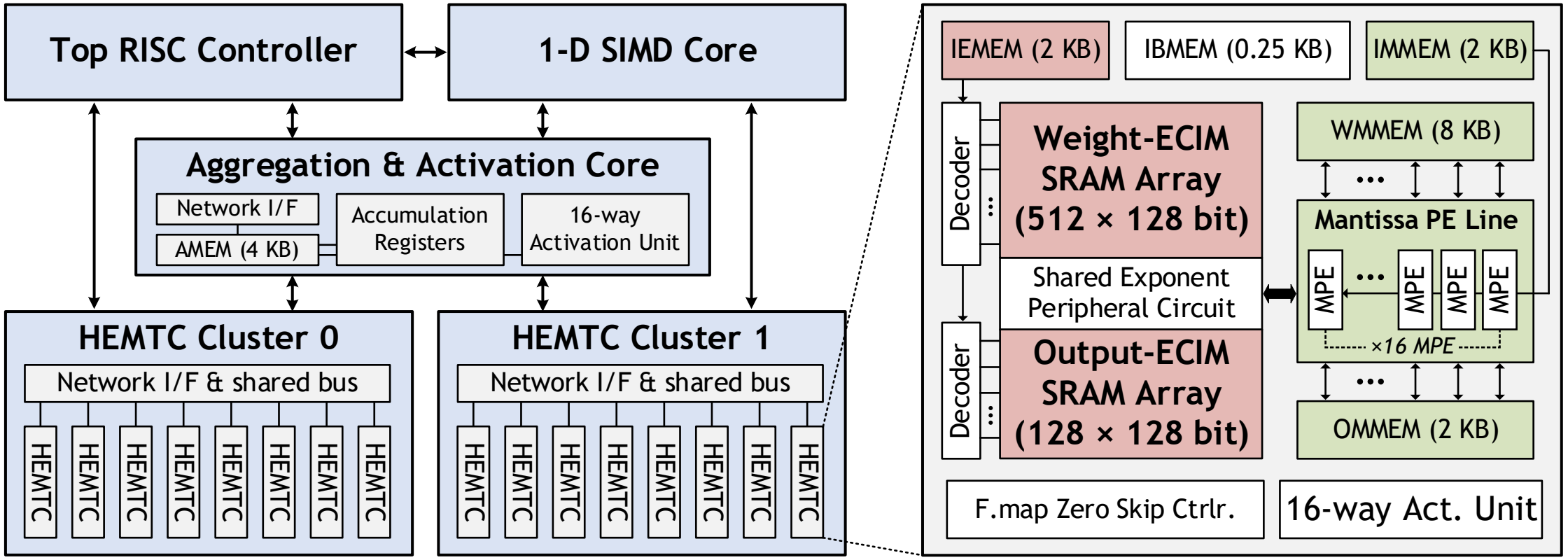
Overall Processor Architecture

- 16 Heterogeneous-exponent-mantissa-training-core (HEMTC)
- Aggregation & Activation Core, Top RISC Controller, etc …



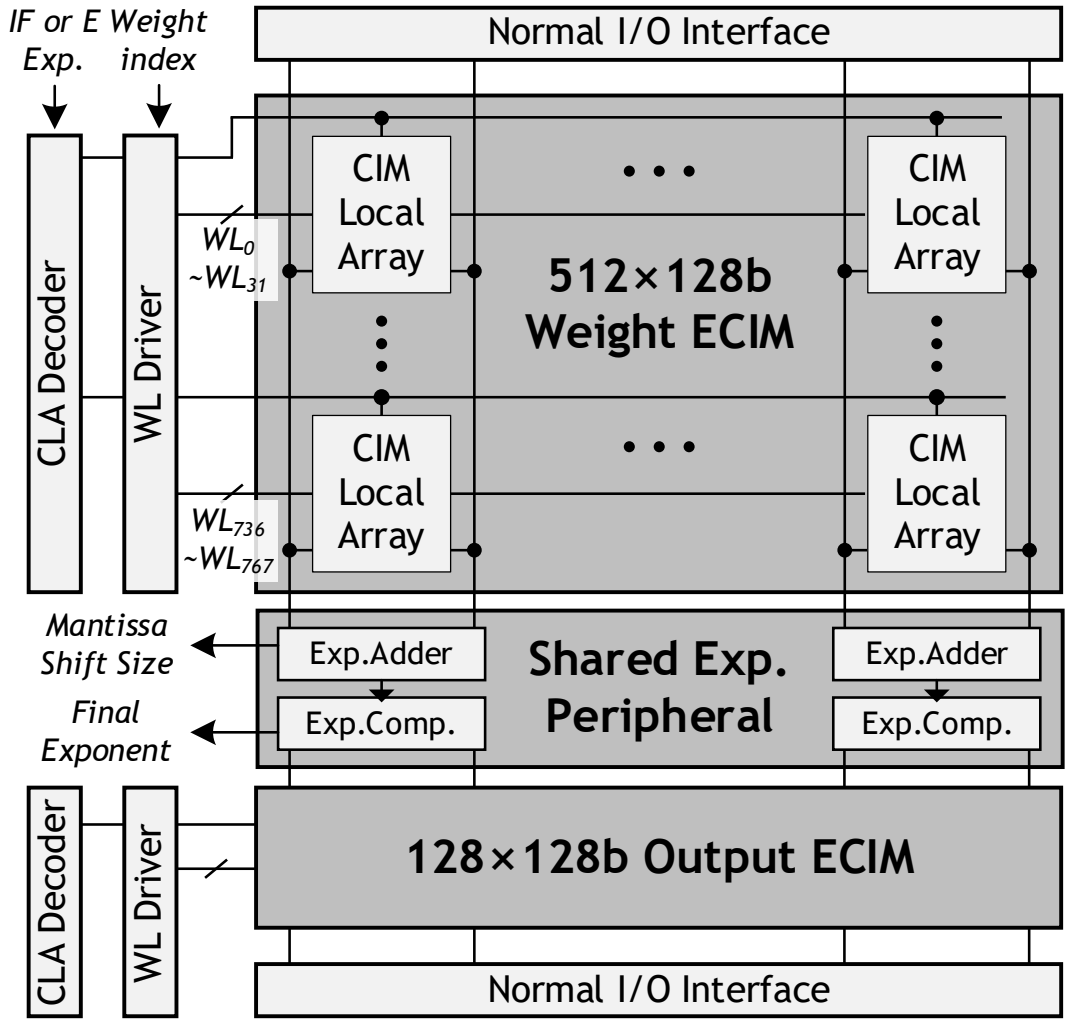
Key Features for Energy-Efficient FP CIM

- 1. Exponent-Computing-in-Memory (ECIM) ➔ Exponent
- 2. Mantissa-Free-Exponent-Calculation (MFEC) ➔ Mantissa



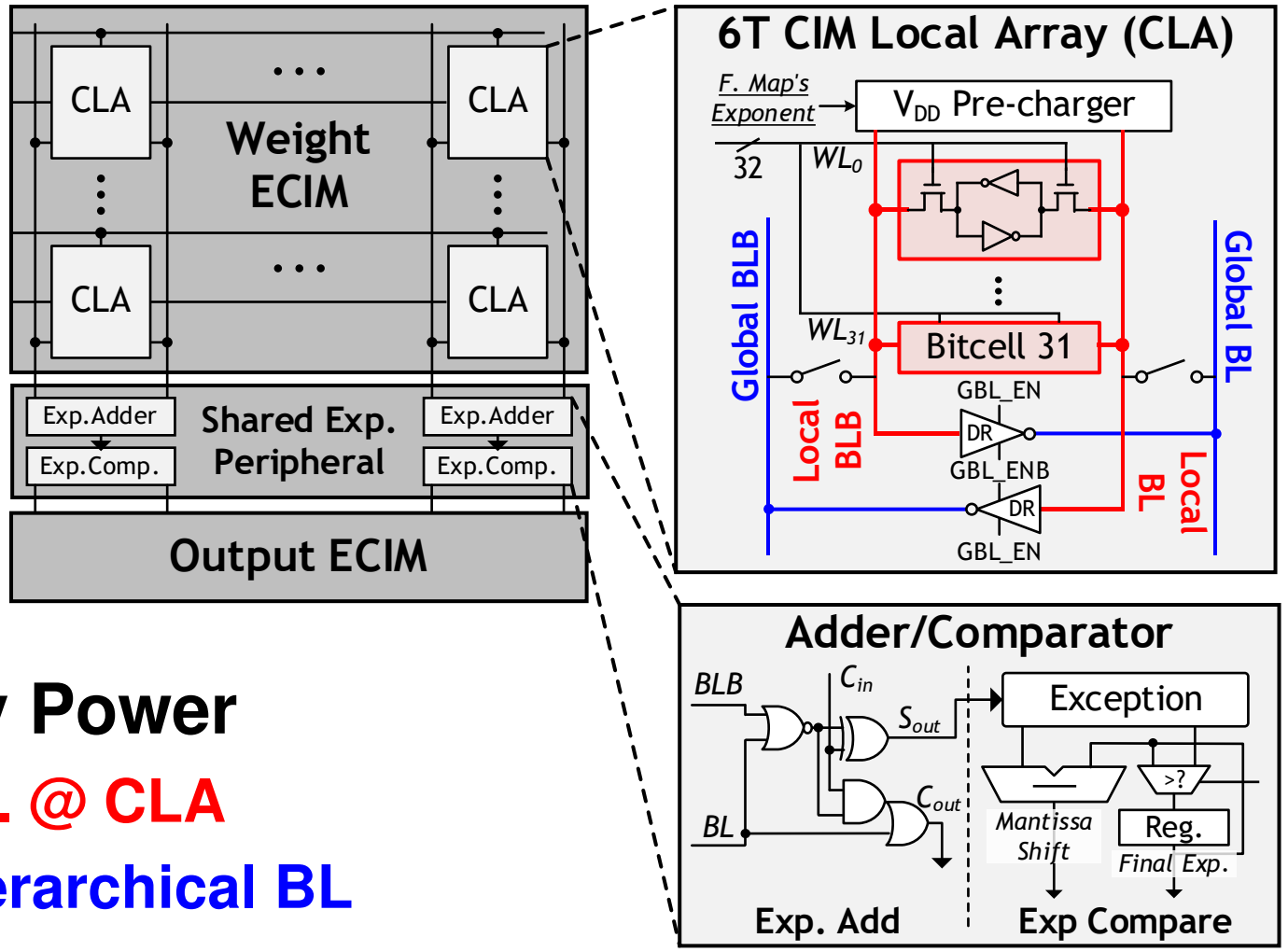
ECIM Architecture

- **512×128 Weight ECIM**
  - Inference, Back-propagation
- **128×128 Output ECIM**
  - Weight-gradient
- **Shared Exponent Peripheral**
  - Finalize exponent op.
- **CLA Decoder, WL Driver**
- **Normal I/O Interface**



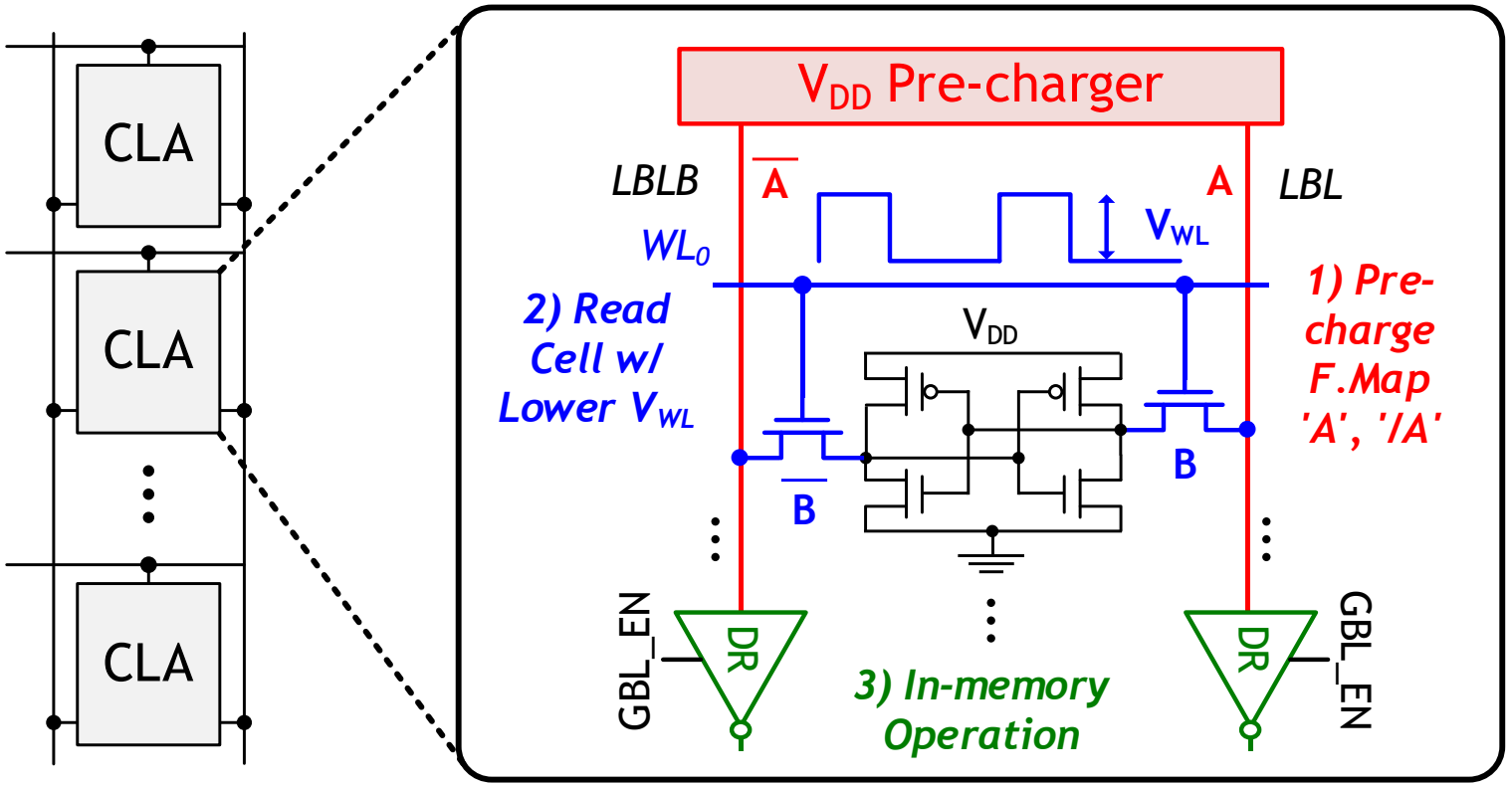
ECIM Architecture - Detail

- **CIM Local Array (CLA)**
  - 32 6T bitcells
  - V_{DD} Pre-charger
- **Near-Memory Logic**
  - Simplified exp. Adder
  - Exp. comparator
- **Features for Low Memory Power**
  - **1) In-memory AND/NOR op. @ CLA**
  - **2) BL charge reusing @ hierarchical BL**



In-memory AND/NOR Operation @ CLA

- **3 Stage 6T-SRAM In-memory AND/NOR Operations**
  - Preventing result corruption by lowering WL voltage



<In-memory Op. Truth Table>

Stored Bit	0	0	1	1
Precharge	0	1	0	1
LBL	0	0	0	1
LBLB	1	0	0	0

WL Op. Cond.

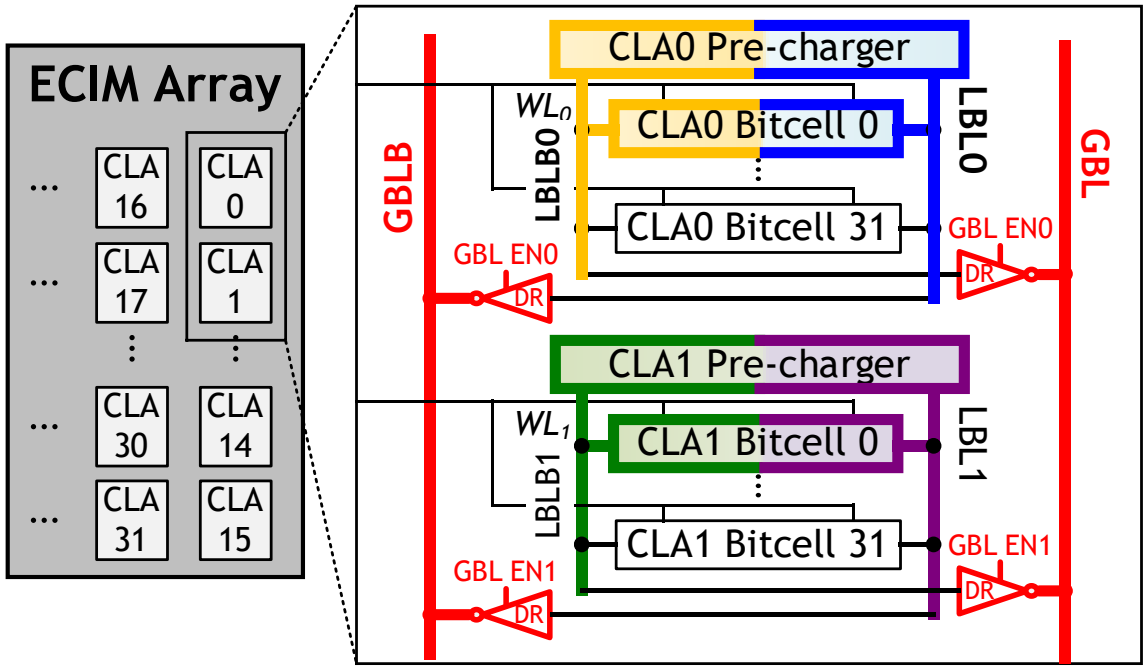
$$V_{th} \lessapprox V_{WL} \lessapprox V_{NML}^{*} + V_{th}$$

* V_{NML} = GBL driver's noise margin low

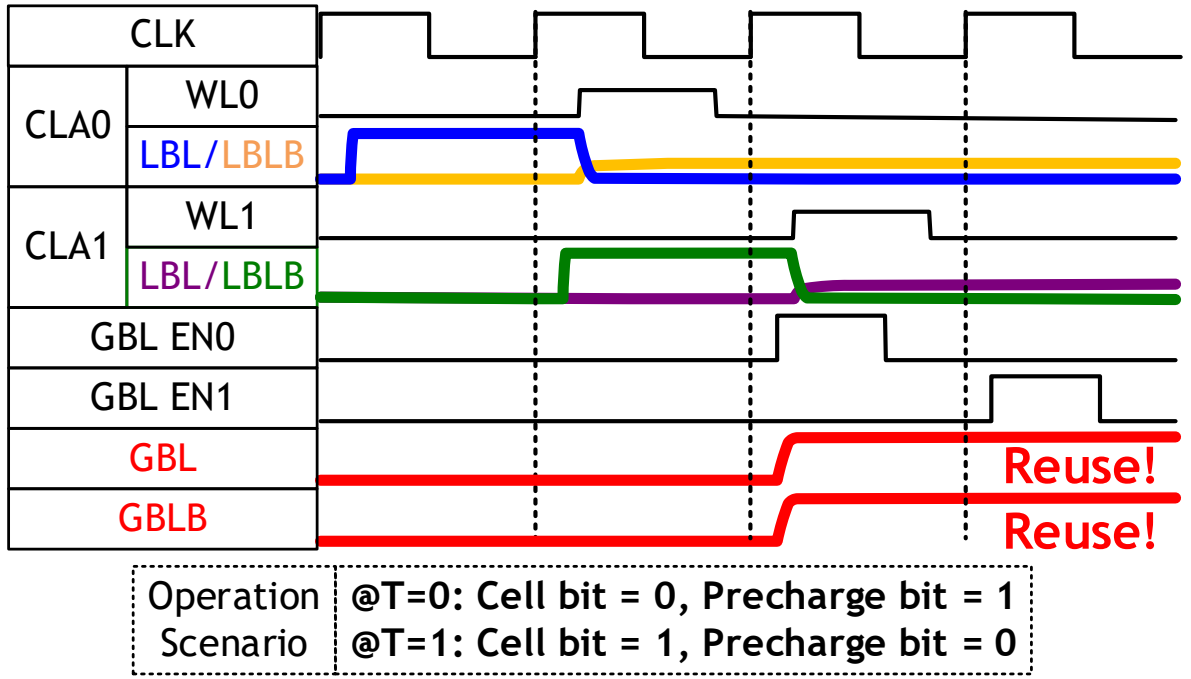
Operation w/ Bitline Charge Reusing

- **Charge Reusing w/ Hierarchical Bitline Structure**
  - Condition: Previous CLA's and/nor result = current CLA's and/nor result

<ECIM Status>



<Operational Waveforms>

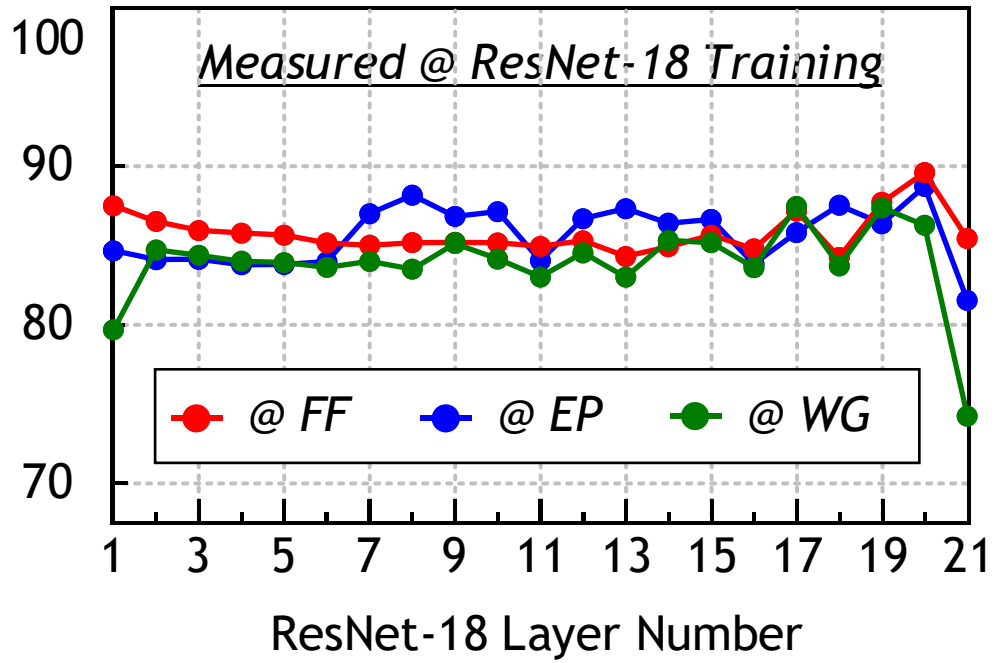


Operation Scenario	@T=0: Cell bit = 0, Precharge bit = 1 @T=1: Cell bit = 1, Precharge bit = 0

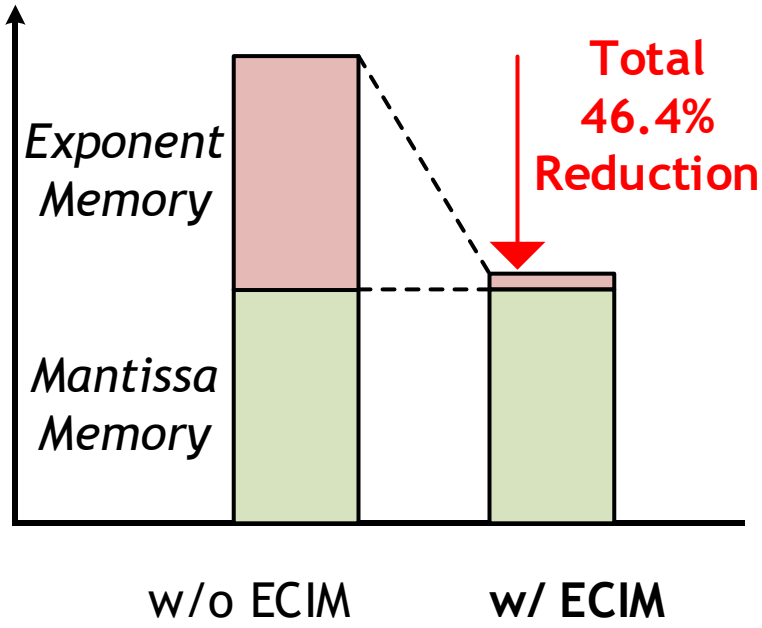
Result of ECIM

- **Reducing BL Pre-charge Power Consumption**
  - >84 % of GBL charge reused, 46.4 % of memory access power ↓ ☺

<GBL Charge Reuse Ratio (%)>



<Normalized MEM Power*>

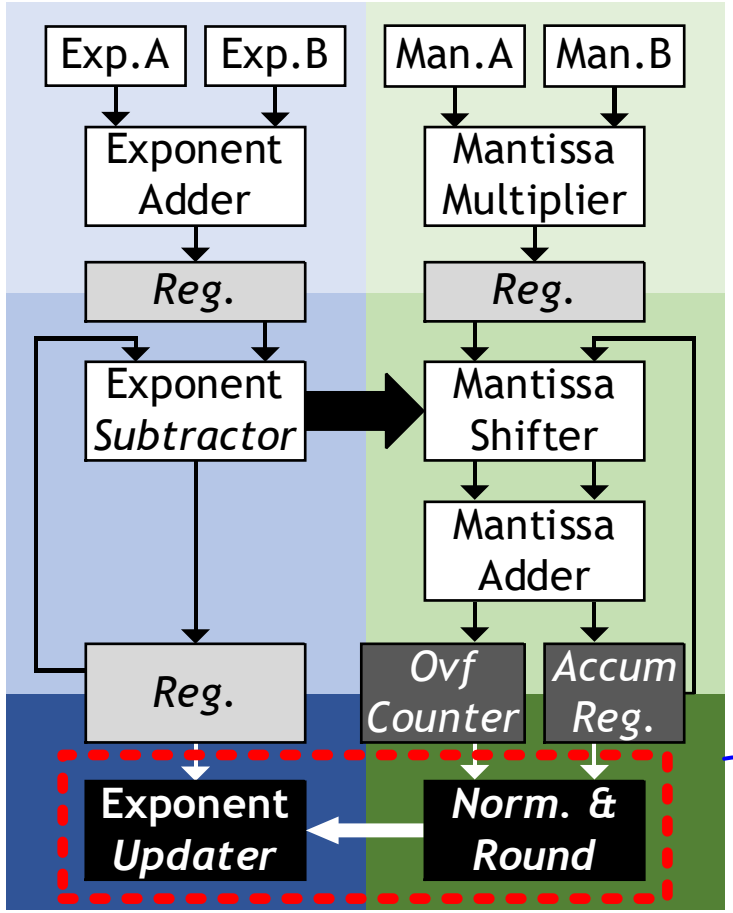


** Measured @ 512 Ch_{in}, 16 Ch_{out}, No Sparsity, 512×128 bit ECIM, 32 cells in CLA*

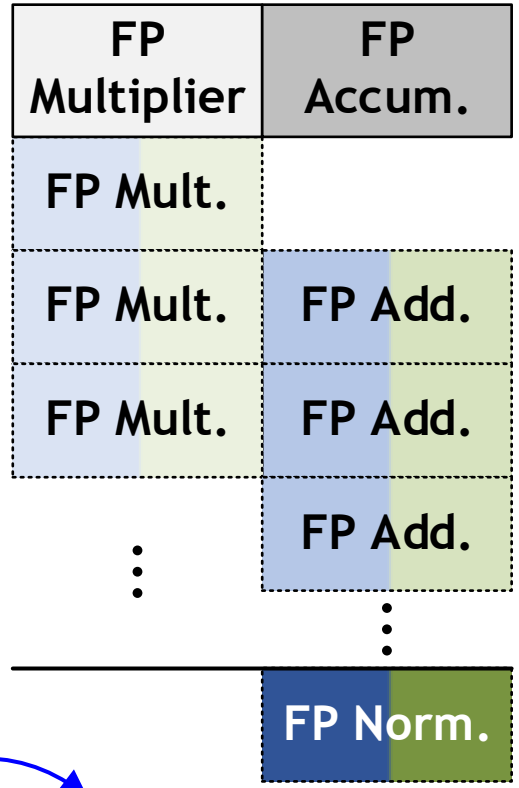
Mantissa-Free-Exponent-Calculation

- **Key Concept: Removing Redundant Norm. Process**
  - Normalize every MAC
    - ➔ Overflow counter + Accumulation register
  - Normalize once after accumulation finished
    - ➔ Pipelining btw/ ECIM & Mantissa PE available ☺
    - ➔ Shortening critical path ☺

<FP MAC w/ MFEC>



<FP Op. Pipeline>



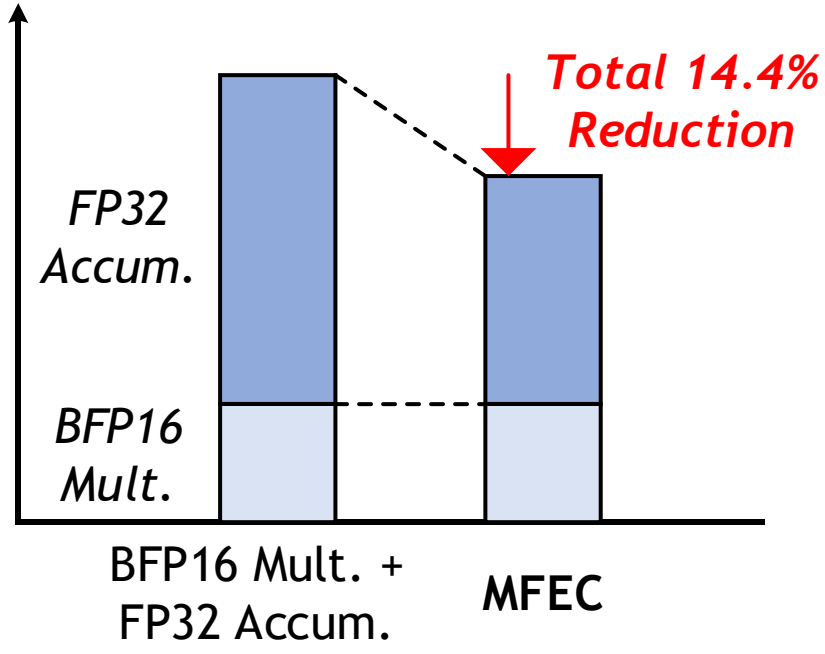
Normalize Once after Accum. Finished ☺

Result of MFEC

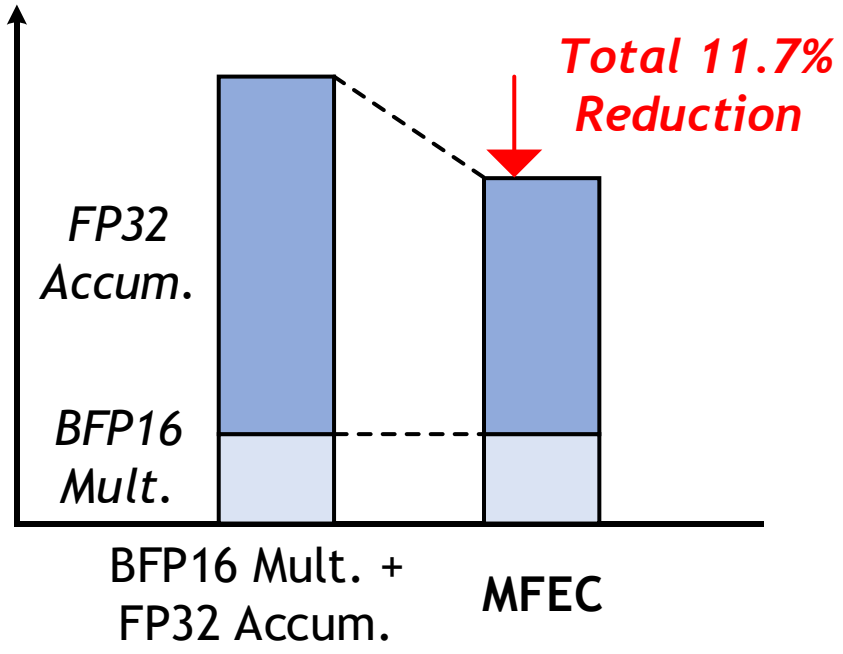
- **Removing Redundant Normalization Process**

➔ **Reduce Power and Area of Mantissa MAC ☺**

<Normalized MAC Power>

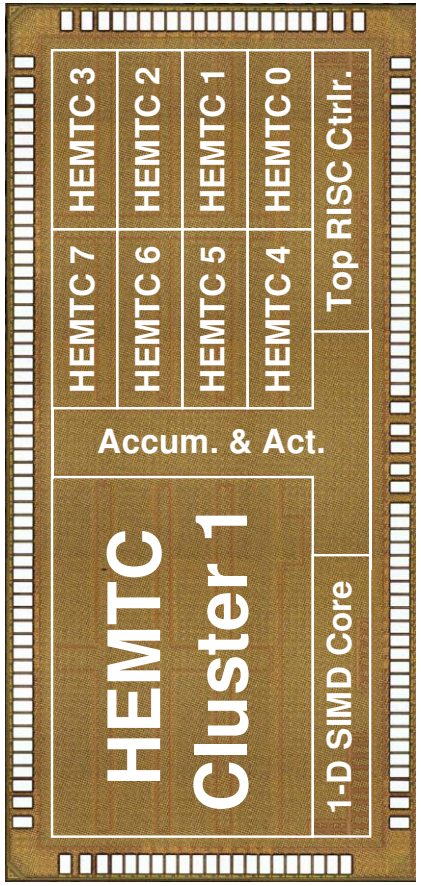


<Normalized MAC Area>



Chip Photograph and Summary

- ×28.6 ~ ×274 Higher Energy Efficiency than Previous FP CIM



	Chip Specifications
Technology	28nm Logic CMOS
Die Area	1.62 mm × 3.6 mm
Precision	BFloat16
Supply Voltage	0.76 V ~ 1.1 V
Frequency	~ 250 MHz
Peak Performance	119.4* – 662.0** GFLOPS @ 250MHz
Power Consumption [mW]	1.2** – 2.1* @ (5MHz, 0.76V)
	91.0** – 156.1* @ (250MHz, 1.1V)
Energy Efficiency [TFLOPS/W]	1.43* – 13.7** @ (40MHz, 0.76V)
	0.76* – 7.3** @ (250MHz, 1.1V)

* = Activation Sparisty 0%, ** = Activation Sparisty 90% (1 MAC = 2 OP)

Conclusion

- **An Energy-Efficient Floating-Point (FP) CIM Processor**

  - **For High Energy-Efficiency**
    - **Heterogenous FP computing arch. → MAC cycle: >5000 → 2 cycle**
    - **Mantissa-free-exponent-calculation → MAC power 14.4 % ↓**
    - **Exponent Computing-in-Memory → MEM power 46.4 % ↓**

A 13.7 TFLOPS/W Floating-point Computing-in-Memory Processor for DNN Training

Thank You!

- **Questions? Feel Free to Contact Me!**
  - E-mail: juhyoung@kaist.ac.kr
  - LinkedIn: https://www.linkedin.com/in/juhyounglee
  - Zoom Meeting: https://us02web.zoom.us/j/3466650389?pwd=c1RWaXFTWGIjaVU1MINtcDhKaGg0dz09 (Password: HC_JHLEE)

- **Acknowledgement**
  - This work was supported by the Samsung Electronics.