ASIA SOUTH PACIFIC DESIGN AUTOMATION CONFERENCE

Tsinghua University

MPICC: Multiple-Precision Inter-Combined MAC Unit with Stochastic Rounding for Ultra-Low-Precision Training

Leran Huang[1] (Speaker)

Yongpan Liu[2], Xinyuan Lin[2], Chenhan Wei[2], Wenyu Sun[2], Zengwei Wang[2]

Boran Cao[1], Chi Zhang[2], Xiaoxia Fu[2], Wentao Zhao[2], and **Sheng Zhang[1]***

[1] Tsinghua Shenzhen International Graduate School, Shenzhen, China

[2] Department of Electronic Engineering, Tsinghua University, Beijing, China

Self Introduction

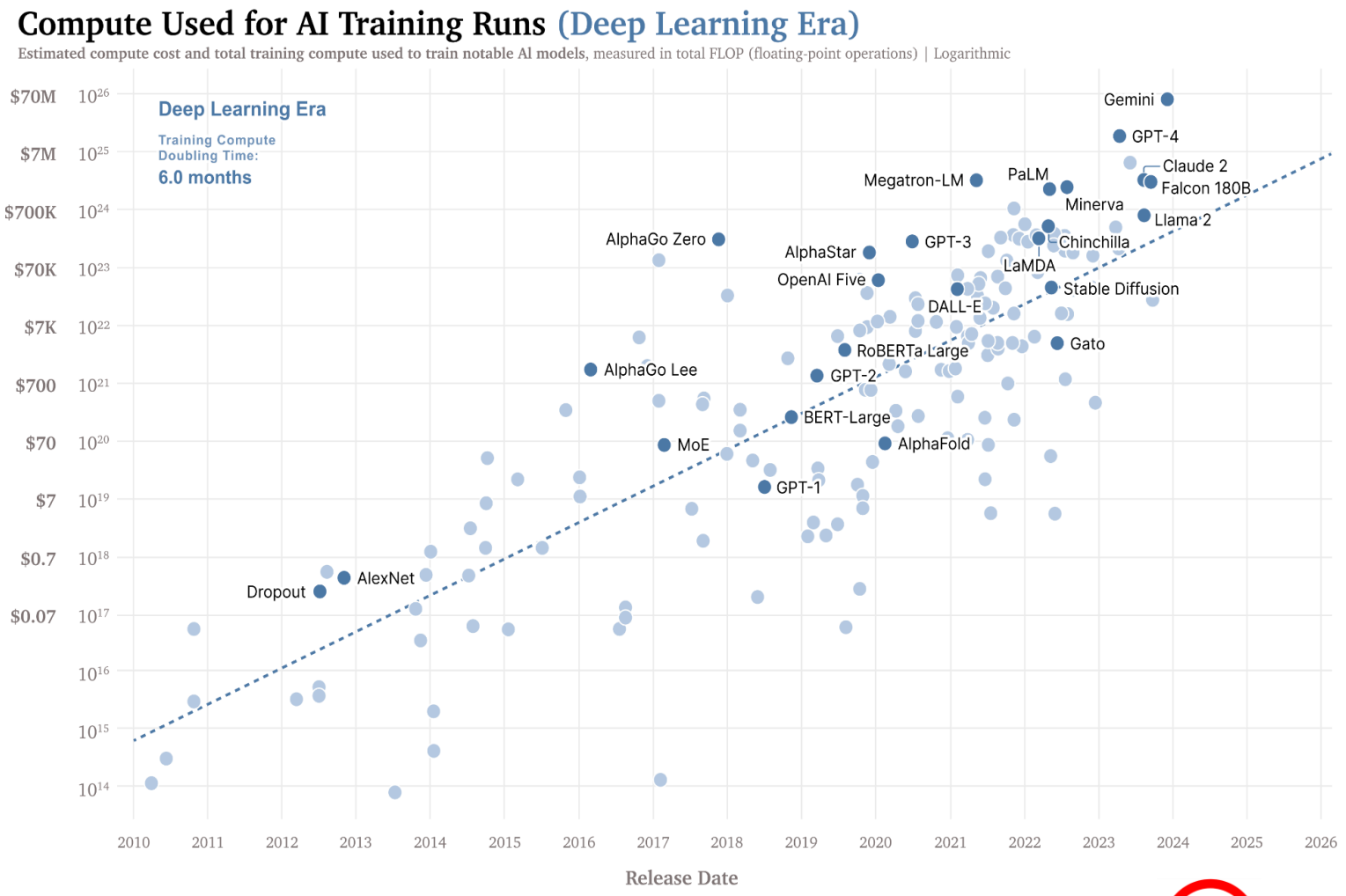
- B.Eng. degree from Harbin Engineering University, Harbin, China, in 2022.
- M.Eng. candidate at Tsinghua University, Beijing, China.
- My research interests include the design of energy-efficient AI accelerators and multi-precision processing elements.

Leran Huang

Outline

- **Background & Challenges**
  - Flexible Support for Ultra-Low-Precisions
  - Reduce Bit Width for Accumulation
  - Support High-Precision at Low Cost
- **Overall Introduction of MPICC**
- **Key Features of MPICC**
- **Experiment & Results**
- **Conclusion**

Cost Trend of Deep Learning Training



[L. Heim et al, arXiv 2024]

- **Model Complexity**
  - Larger model sizes
  - Greater compute load
- **Hardware Needs**
  - More processing units
  - Larger memory capacity
  - Wider bandwidth
  - Longer processing time

⇩

☹ **An exponential growth in training costs**

Quantization: Reduce Costs & Improve Performance

How Quantization Benefits AI

Lower power

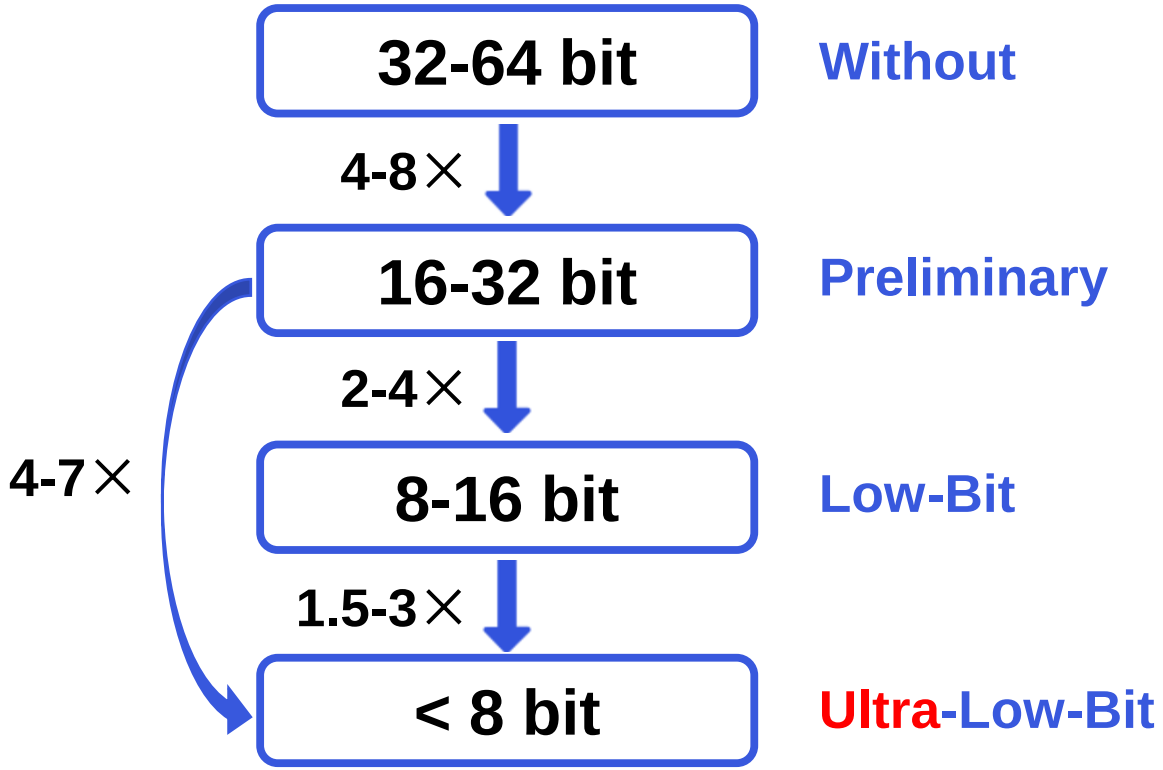
Lower memory bandwidth

Lower storage

Higher performance

[Jilei Hou, Qualcomm AI Research]

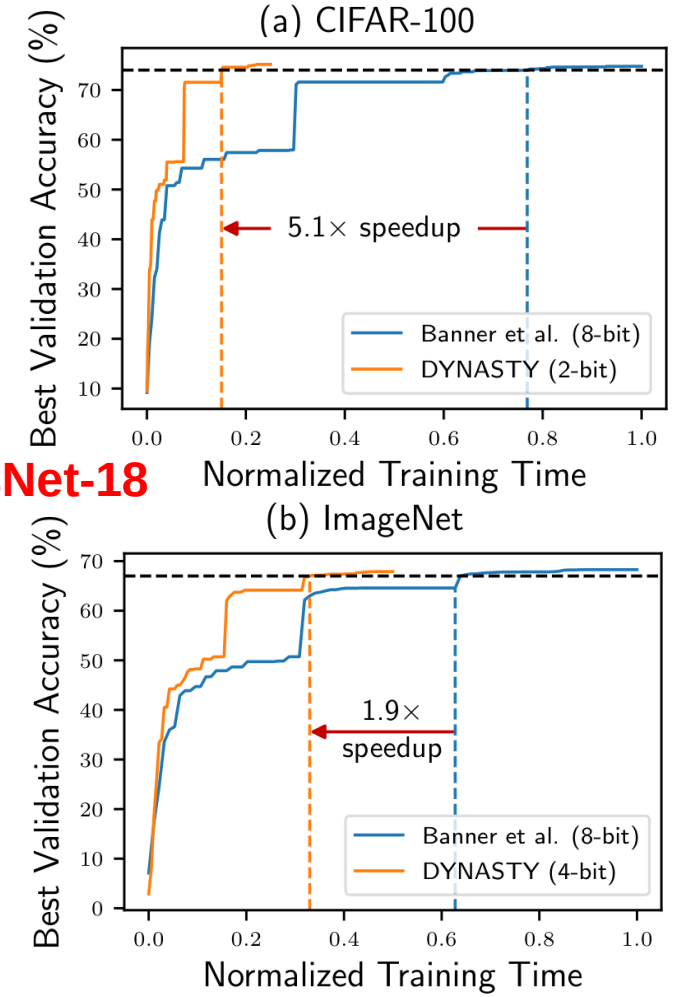
Evolution of Quantization



Estimated performance improvement

Research on Ultra-Low-Precison Training

Improve Training Speed

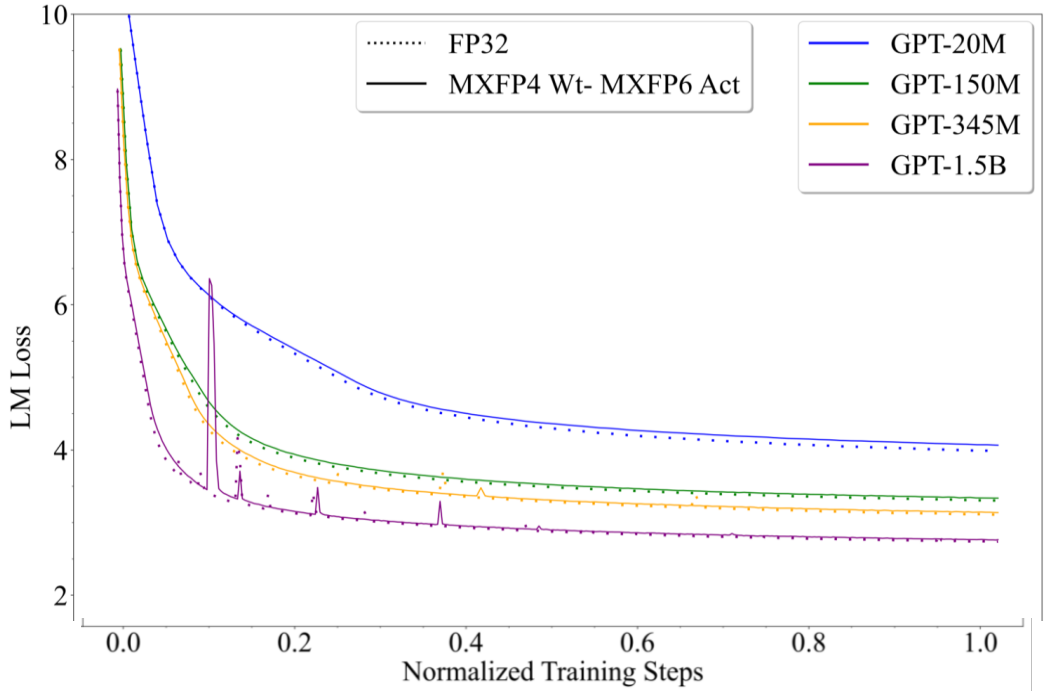


ResNet-18

[Ruoyang Liu et al, ASP-DAC 2023]

Maintain Language Model Loss

Model	FP32	MXFP4 Wt & MXFP6 Act
GPT-20M	3.98	4.04
GPT-300M	3.11	3.14
GPT-1.5B	2.74	2.76

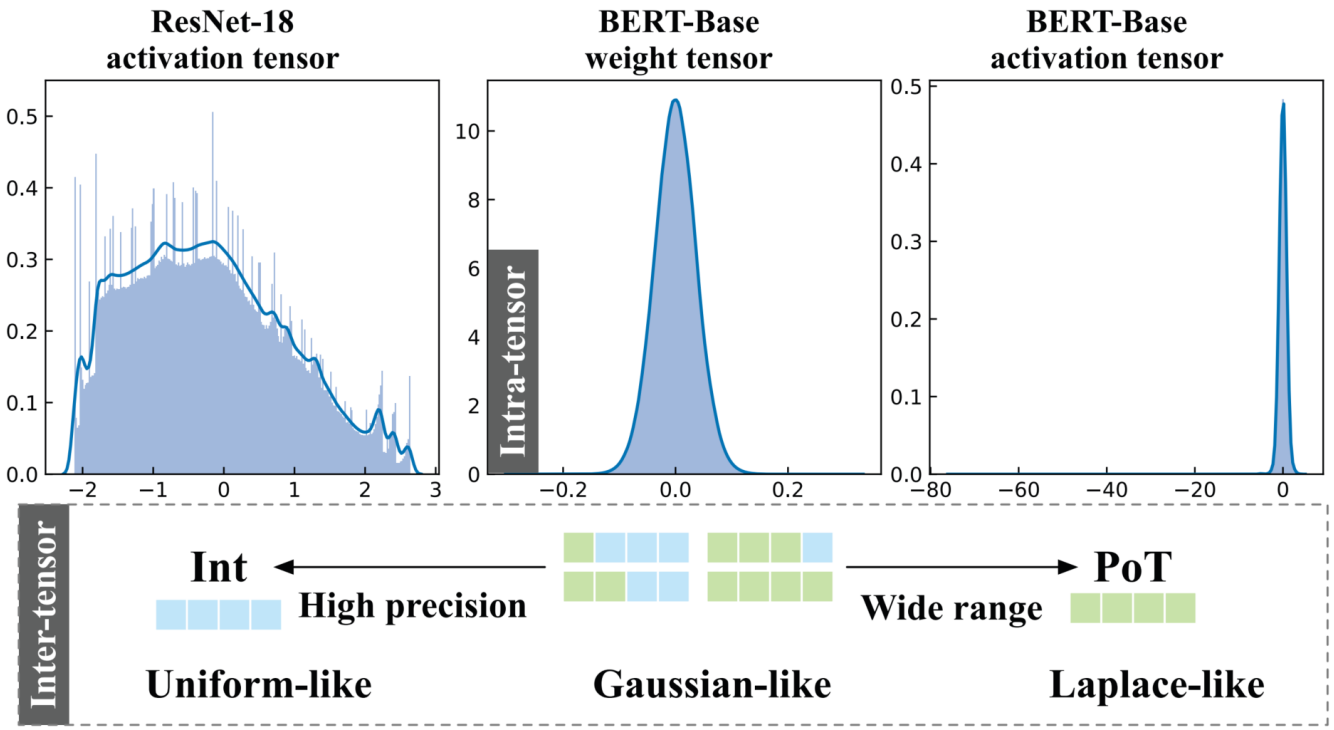


[Microsoft, AMD, Intel, NVIDIA, arXiv 2024]

Chal.1: Flexible Support for Ultra-Low-Precisions

Features of Ultra-Low-Bit Quantization

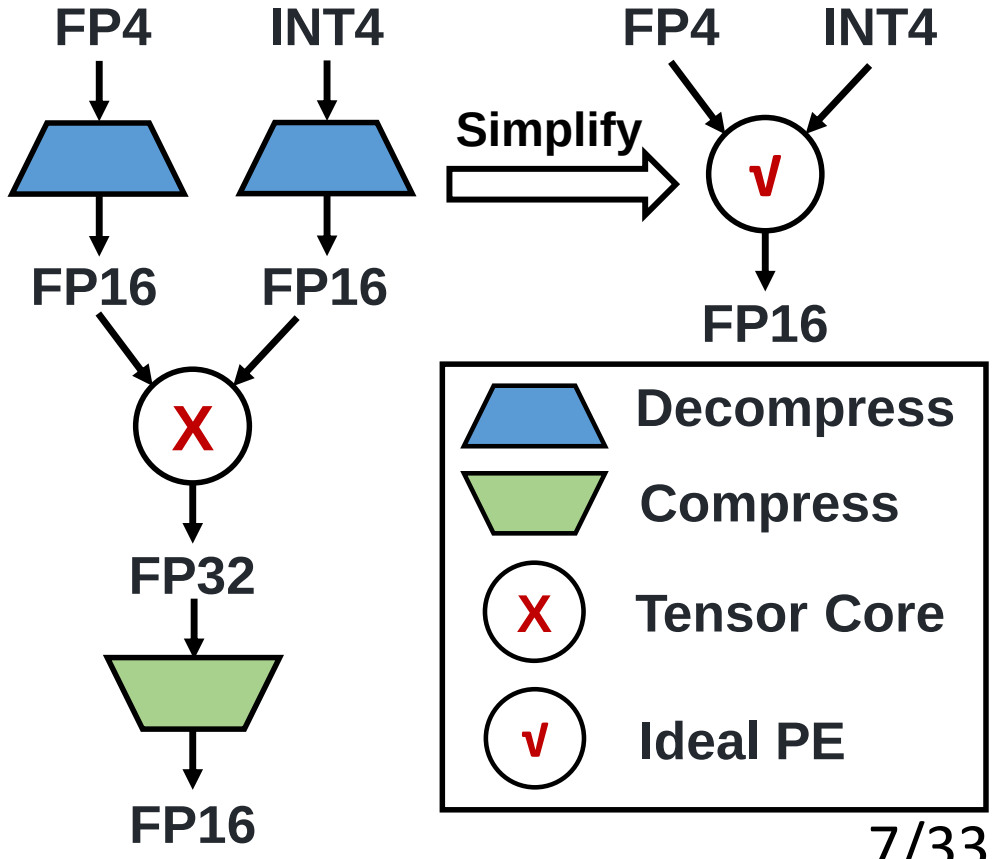
- Use multiple ultra-low-precisions (FP, INT, LOG)
- Determine precision based on **tensor distribution**



[Cong Guo et al, MICRO 2022]

Demands for Direct Computing

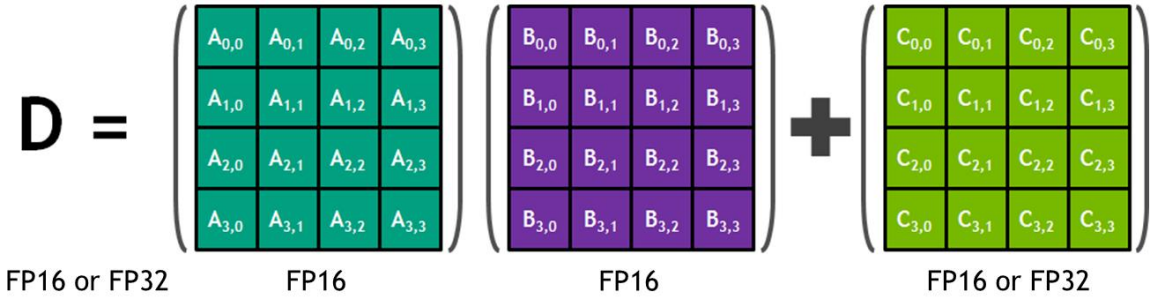
- Act & Wt often use **different precision**
- Direct computing saves resources



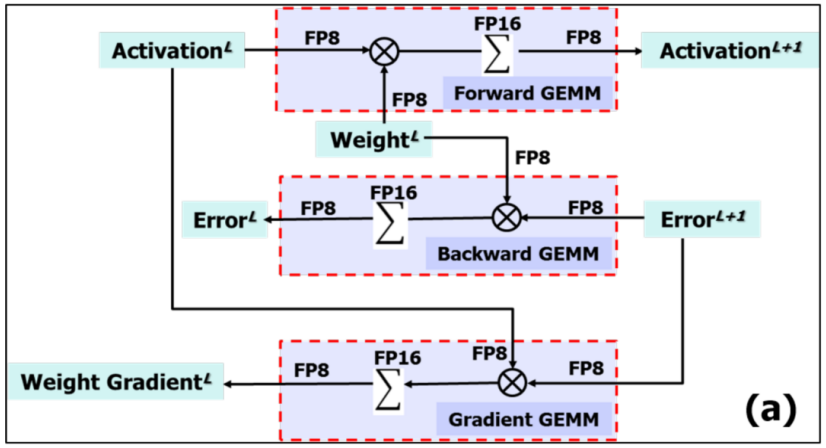
Chal.2: Reduce Bit Width for Accumulation

Why Bit Width Difficult to Reduce

- Ensure a wide range to **prevent overflow**
- Prevent training stagnation caused by **swamping**



[VoltaTensor Core, Nvdia 2017]

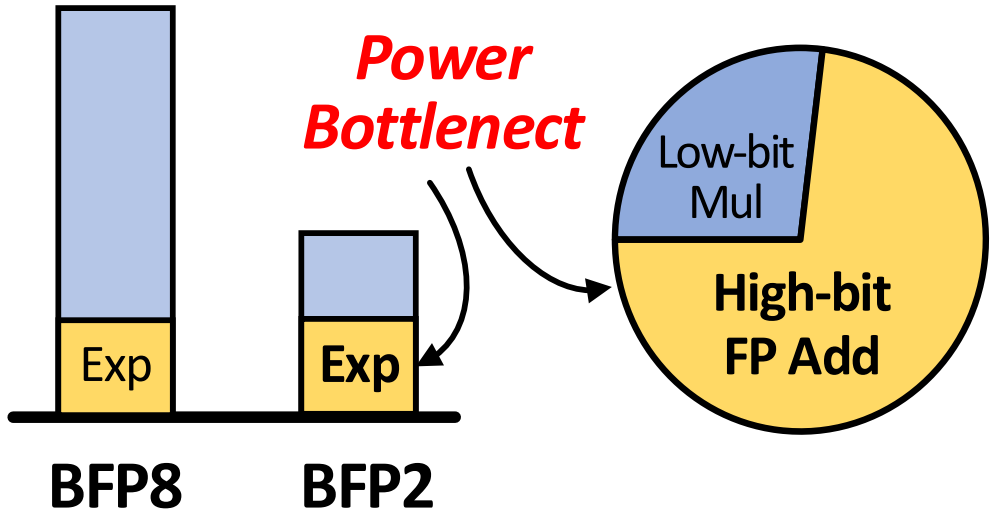


[Naigang Wang et al., NeurIPS 2018]

Necessity of Reducing Bit Width

- FP adders become power **bottleneck**
- Reduce the cost of result compression

FP Processing Power Breakdown



[Ruoyang Liu et al., CICC 2023]

Chal.3: Support High-Precision at Low Cost

Use High-Precision for Critical Layers

- Improve accuracy and support more networks
- Maintain performance ([3] throughput reduces **24%**)

Work	Model	Method	Degradation
[1]	AlexNet	Upgrade the final layer precision from FP8 to FP16	0.07%
[2]	Mobile NetV2	Upgrade Conv1x1 layers precision from FP4 to FP8	0.60%
[3]	ResNet -50	Add 3 FP16 fine-tuning epochs (originally 4bit)	0.32%

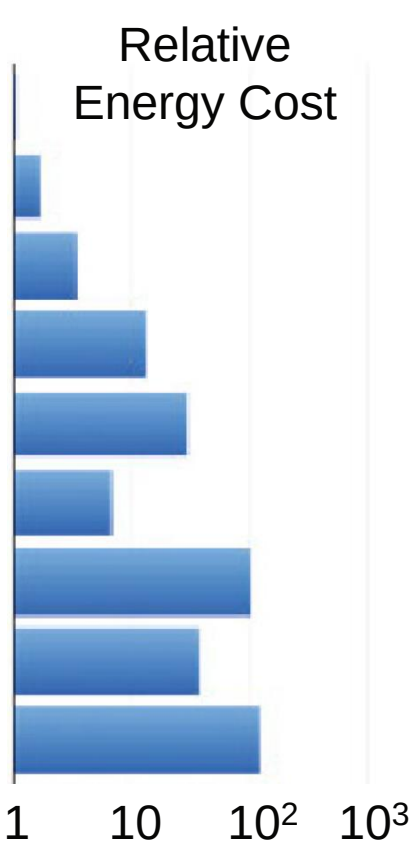
- **Reference List**

[1] *[Naigang Wang et al., NeurIPS 2018]*
[2] *[Xiao Sun et al., NeurIPS 2020]*
[3] *[Brian Chmiel et al., ICLR 2023]*

Cost Versus Bit Width

- Cost increases rapidly with bit width
- Support high-precision is **expensive**

Operation	Energy (pJ)
8b INT Add	0.03
16b INT Add	0.05
32b INT Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b INT Mult	0.2
32b INT Mult	3.1
8b FP Mult	1.1
32b FP Mult	3.7



[M. Horowitz, ISSCC 2014]

Outline

- Background & Challenges
- **Overall Introduction of MPICC**
  - **Multiple-Precision Inter-Combined Computing**
- **Key Features of MPICC**
- **Experiment & Results**
- **Conclusion**

Overall Introduction of MPICC

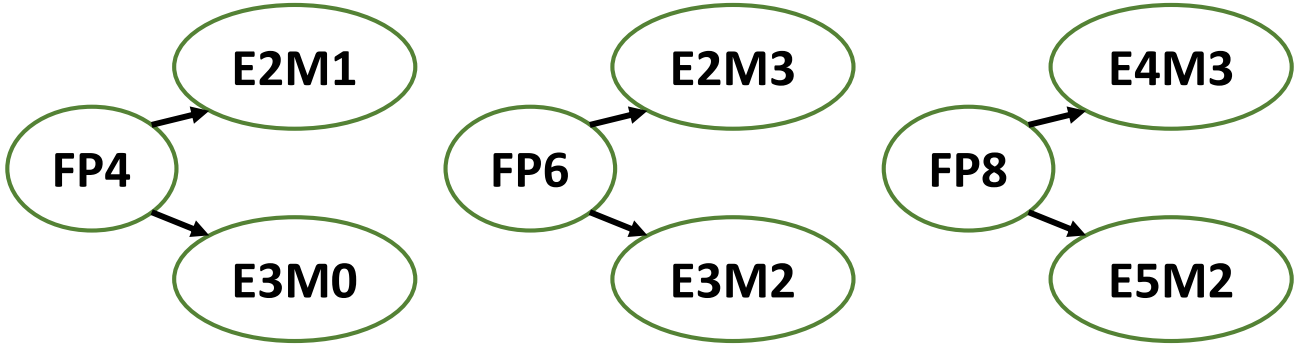
Function of MPICC

- Performs 1/2/4 (**K**) dot products per operation
- Accumulates the results in FP12 & INT8

$$F = \begin{cases} A_1 \times B_1 + C & K \le 1 \\ A_1 \times B_1 + A_2 \times B_2 + C & K = 2 \\ A_1 \times B_1 + A_2 \times B_2 + A_3 \times B_3 + A_4 \times B_4 + C & K = 4 \end{cases}$$

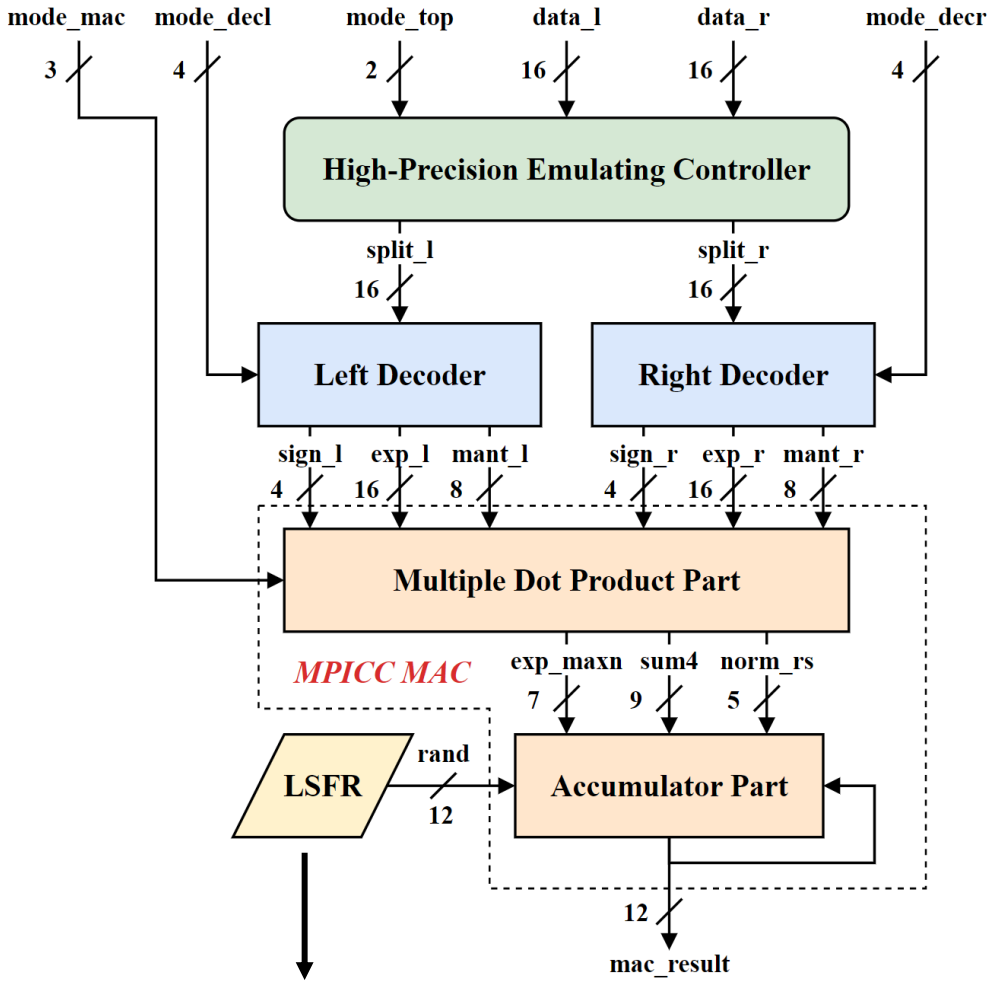
Supported Precision Modes

- LOG4 means radix-4 FP4 (S1E3M0)



All variants are supported

MPICC Overall Architecture



Linear Feedback Shift Register

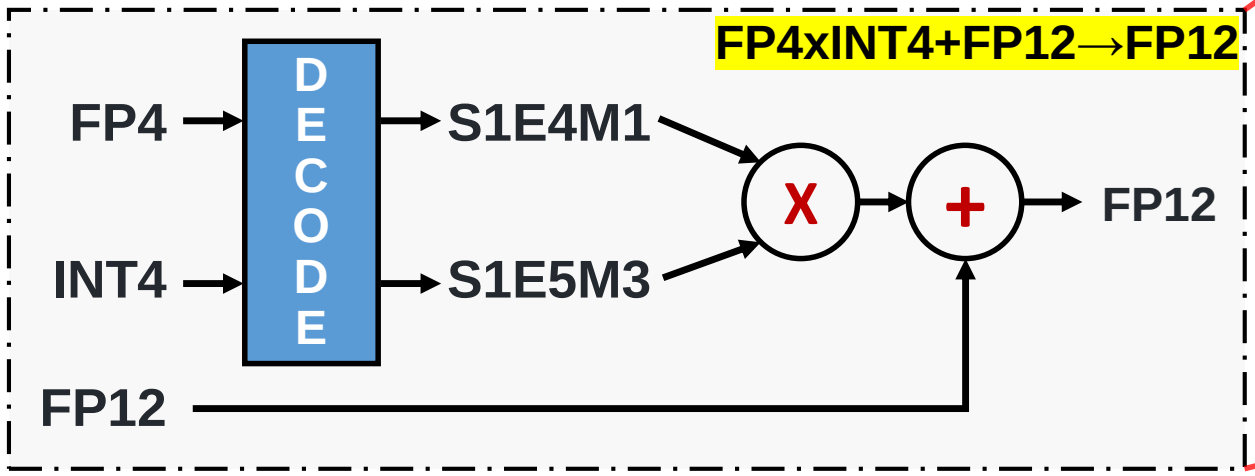
Feature 1 of MPICC

Feature1: MPICC Architecture

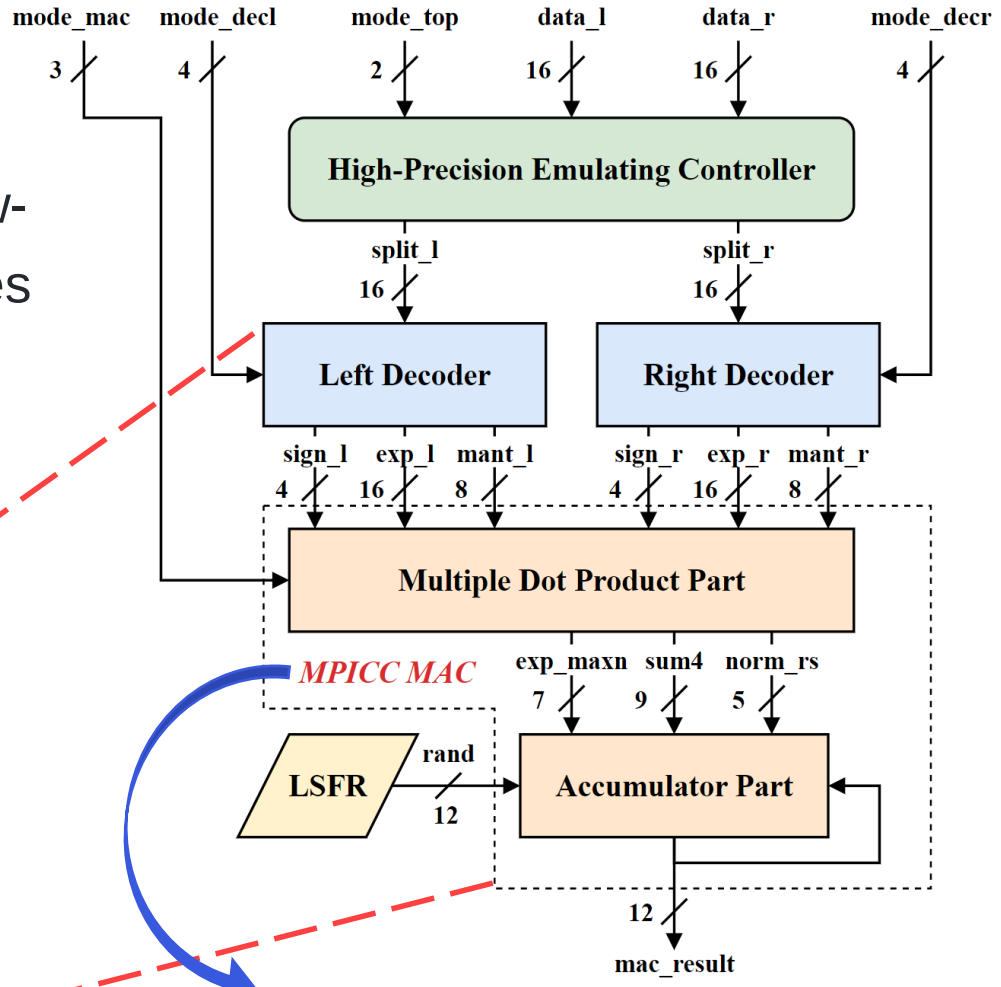
- ***Aim:*** Supports inter-computations among multiple FP, INT, and LOG formats
- ***Chal. 1:*** Solve need for multiple precision in ultra-low-precision training & Direct computing saves resources

Principle of MPICC

- Compute after decoding into a **unified format**



MPICC Overall Architecture



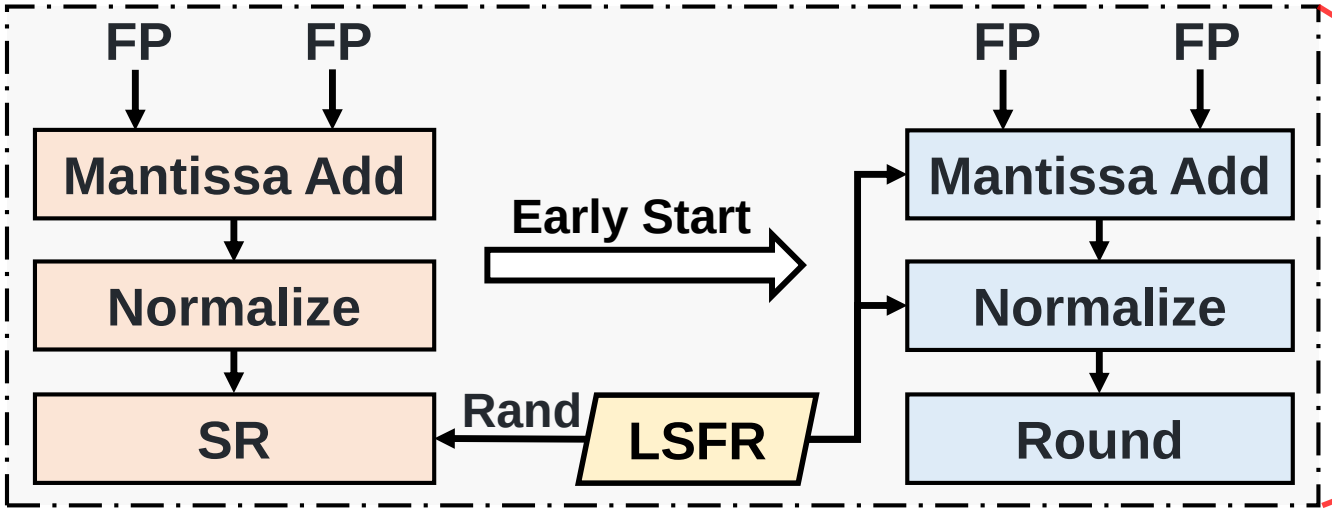
Feature 2 of MPICC

Feature2: Optimized Stochastic Rounding

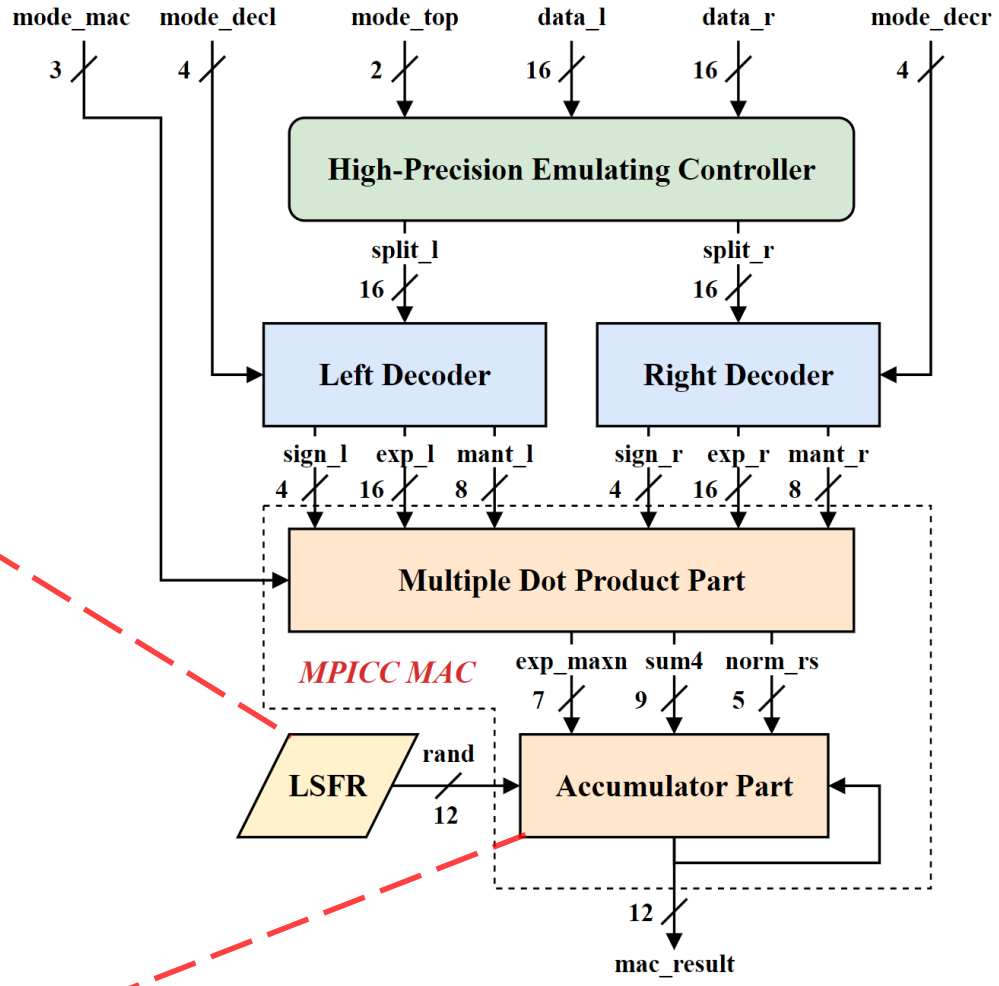
- ***Aim:*** Perform SR **in advance** to maintain accuracy and reduce cumulative bit width
- ***Chal. 2:*** Prevent **swamping** and training stagnation & Save accumulator resource consumption

Principle of Variable-SR

- Unlike [1], Variable-SR can be used on **any MAC**



MPICC Overall Architecture



[1] *[Sami Ben Ali et al., DATE 2024]*

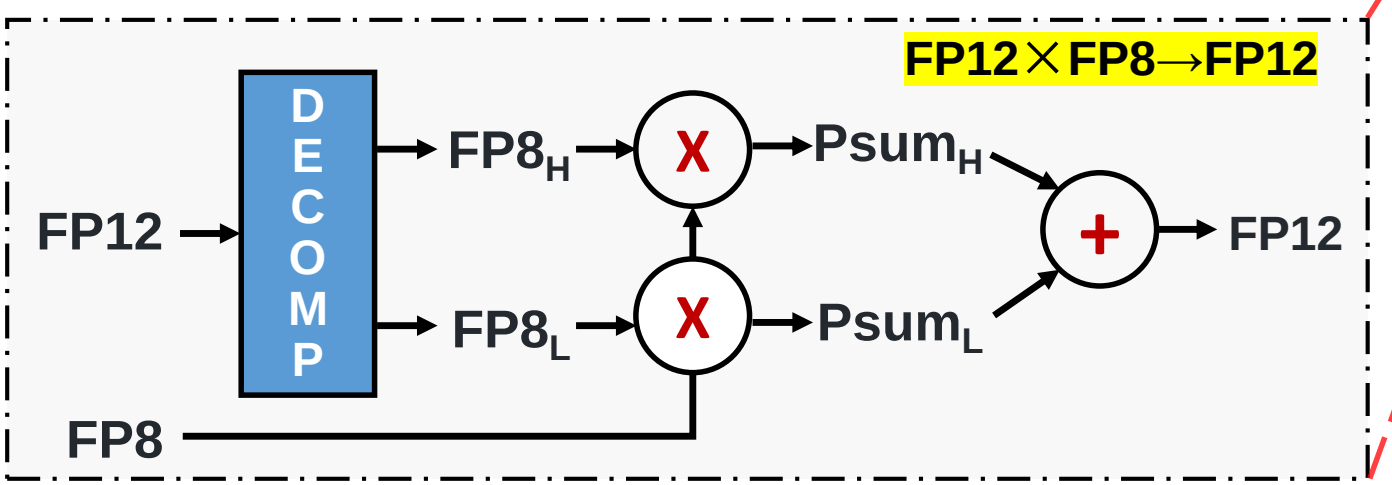
Feature 3 of MPICC

Feature3: High-Precision Emulating

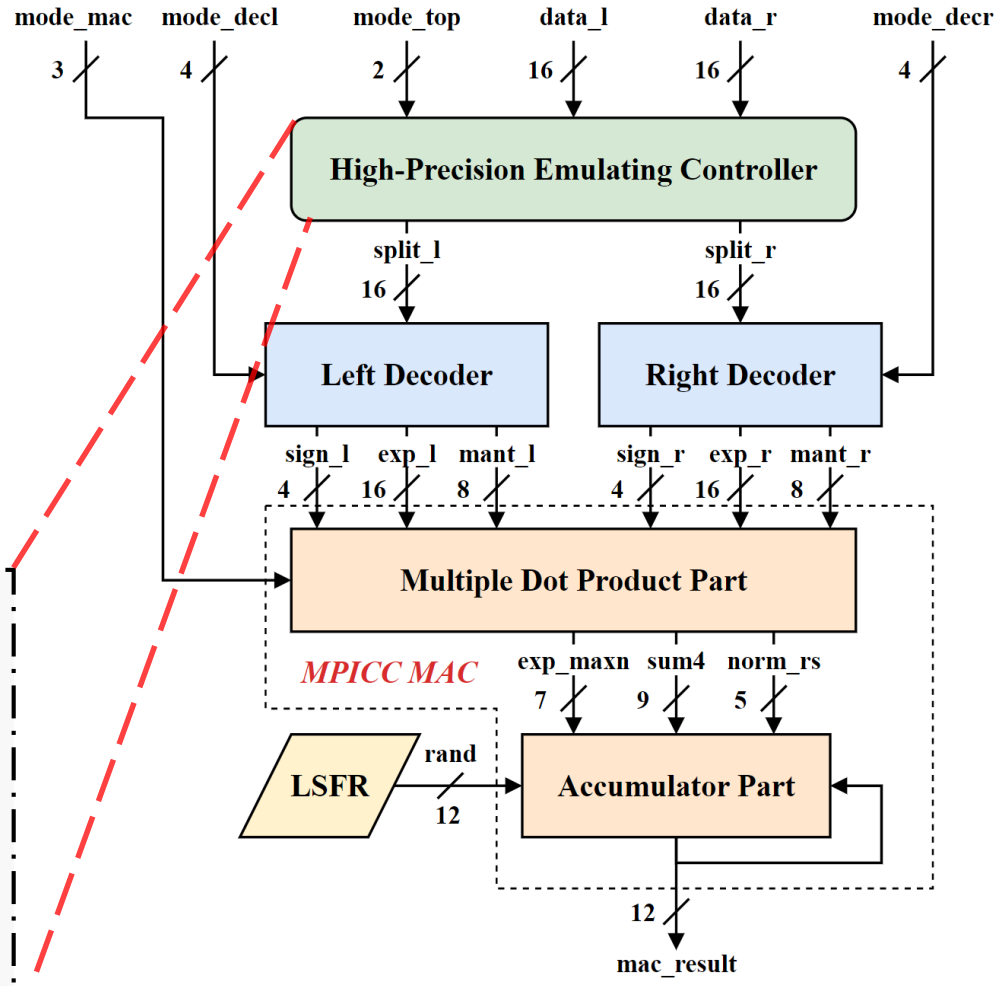
- ***Aim:*** **Decompose** FP12 computation into multiple FP8 computations in low-precision hardware
- ***Chal. 3:*** **Avoid adding** high-precision hardware & Only **5.7%** increase in area cost for control

Principle of High-Precision Emulating

- Decompose 1 FP12 into the sum of 2 FP8s



MPICC Overall Architecture



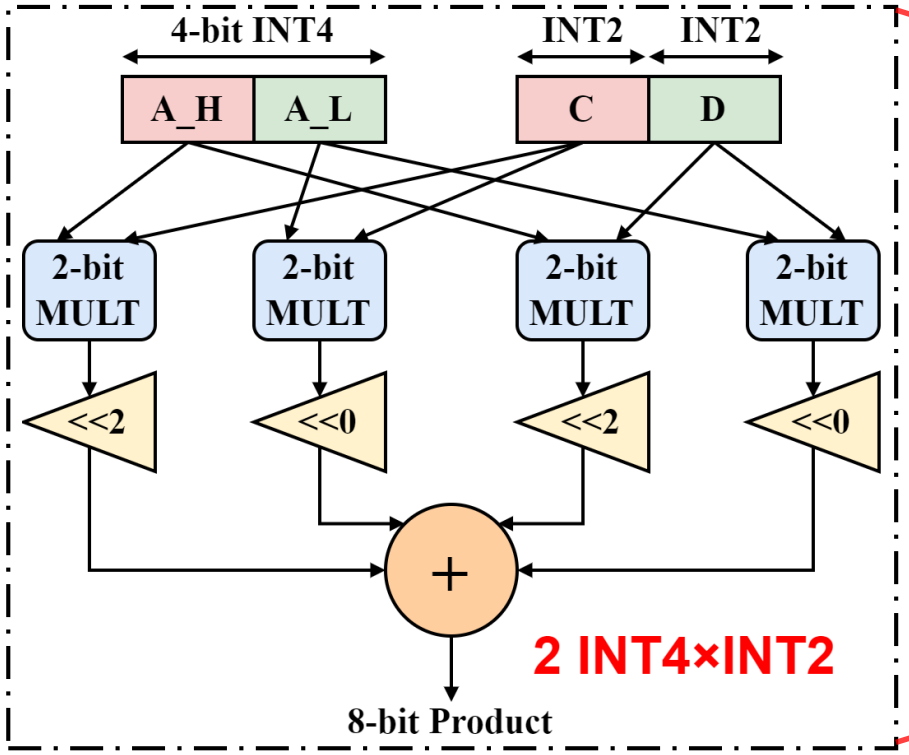
Outline

- Background & Challenges
- Overall Introduction of MPICC
- **Key Features of MPICC**
  - Feature 1: MPICC Architecture
  - Feature 2: Optimized Stochastic Rounding Strategy
  - Feature 3: High-Precision Emulating Controller
- **Experiment & Results**
- **Conclusion**

Feature 1 — MPICC Implementation Principle

Realize Multi-Shape Multiplication (INT)

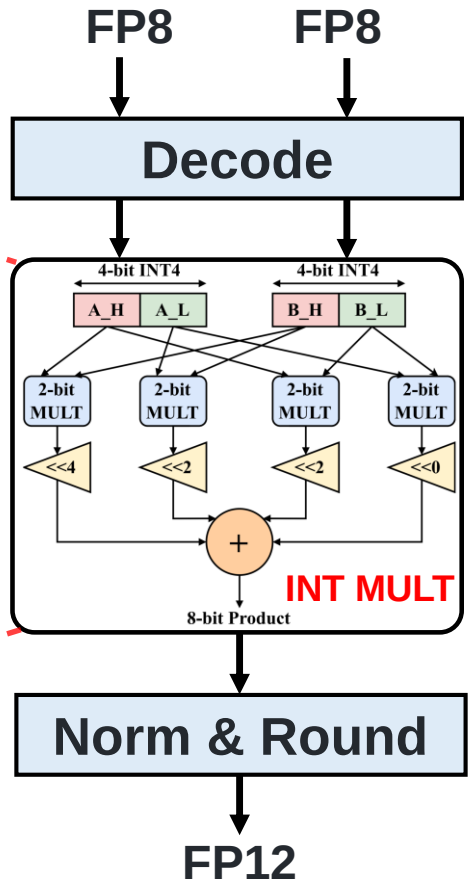
- Decompose 2^n bit multiplication into the sum of multiple 2 bit multiplications after shifting
- Adjust shift amounts to perform **multi-shape**



[Hardik Sharma et al., ISCA 2018]

Migration of INT Scheme to FP

- Add **FP-specific processes** such as decoding, normalization, rounding …



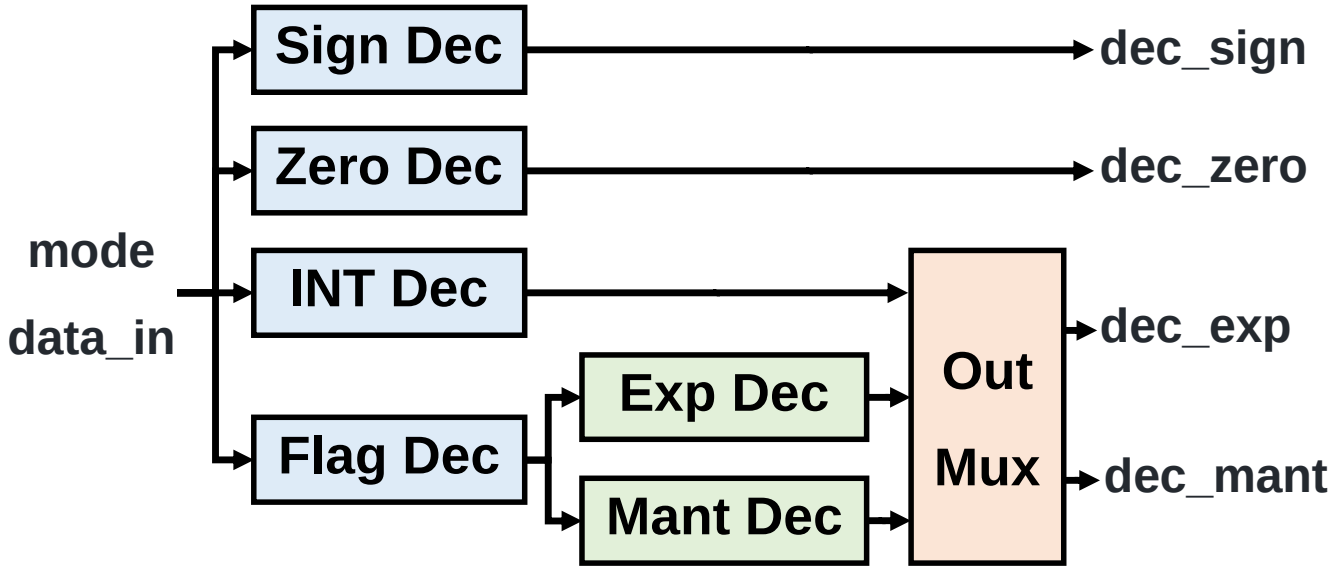
Feature 1 — Logic of Decoding

Decode to Uniform Type

- Decode to sign, exponent, and mantissa

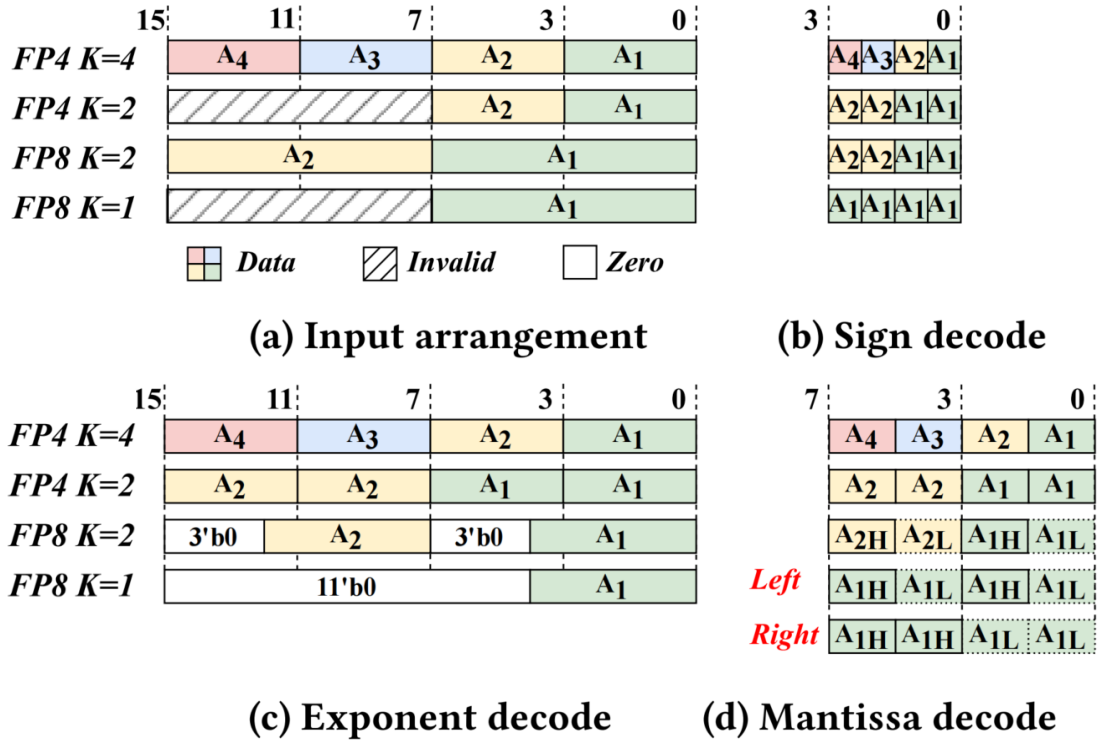
Result	Input
S1E5M3	FP8, FP6, INT4
S1E4M1	FP4, LOG4

- Decoder overall architecture



Input & Output Arrangement

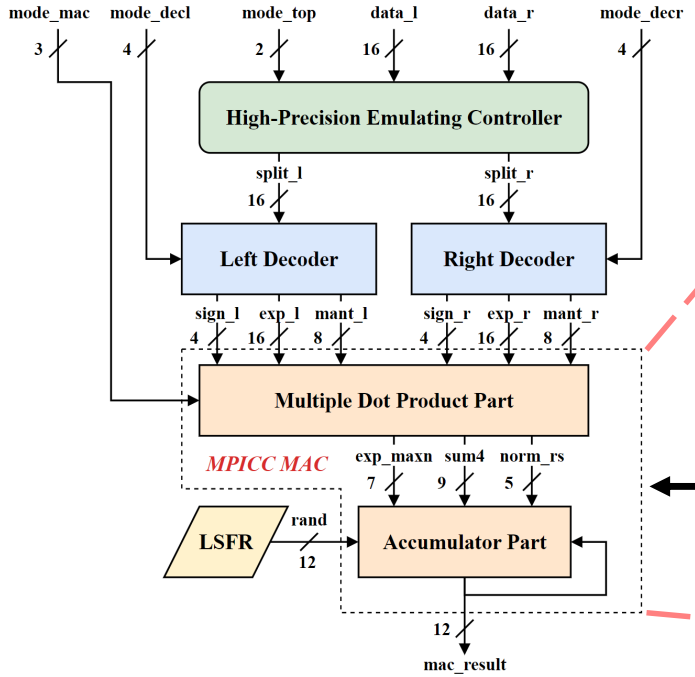
- Input comprises 2 8b data or 4 4b data
- Adjust bias when decoding exponents
- Mant. decoding result contain leading bit



Feature 1 — Multiple Dot Product Unit

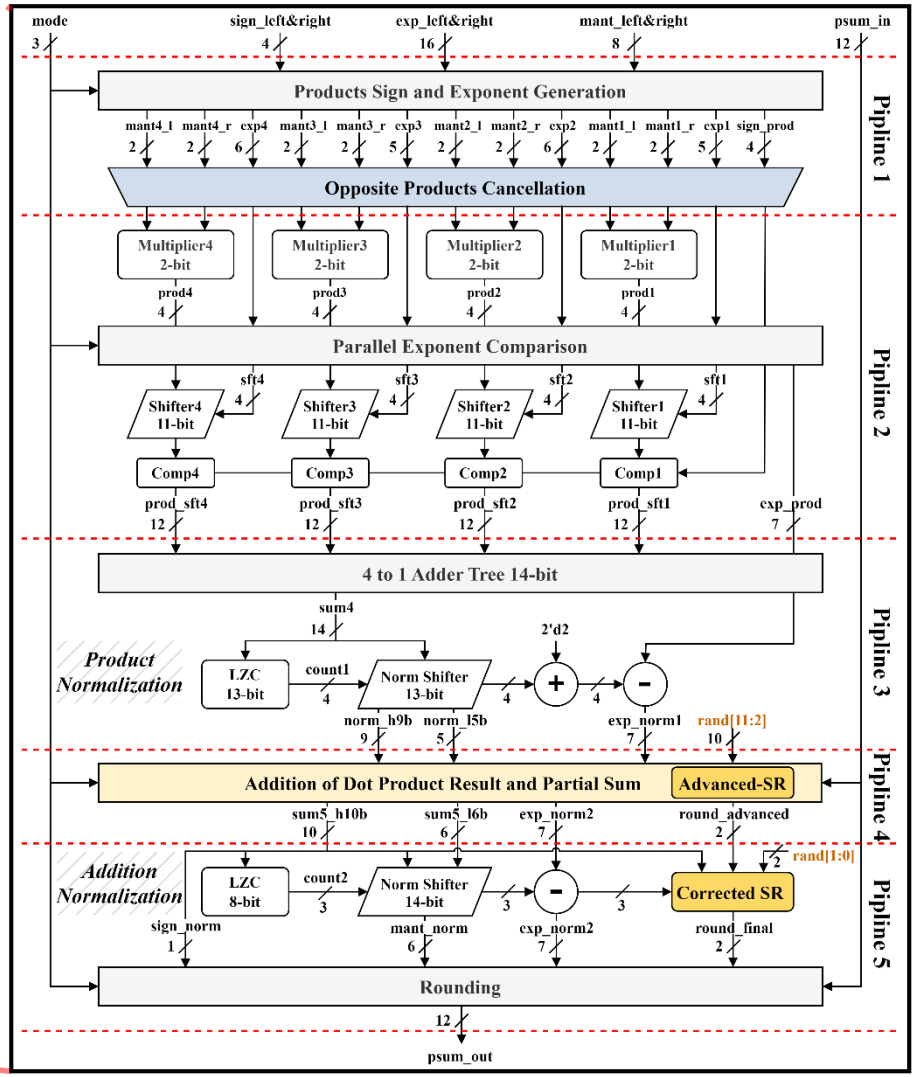
Functions of the 5 Pipelines

- ***1:*** Product unpack and cancellation
- ***2:*** Multiplication, exponent comparison and align
- ***3:*** Reduction and product normalization
- ***4:*** Addition of dot product result and partial sum
- ***5:*** Normalization and rounding



← MPICC Architecture

Multiple Dot Product Unit Architecture

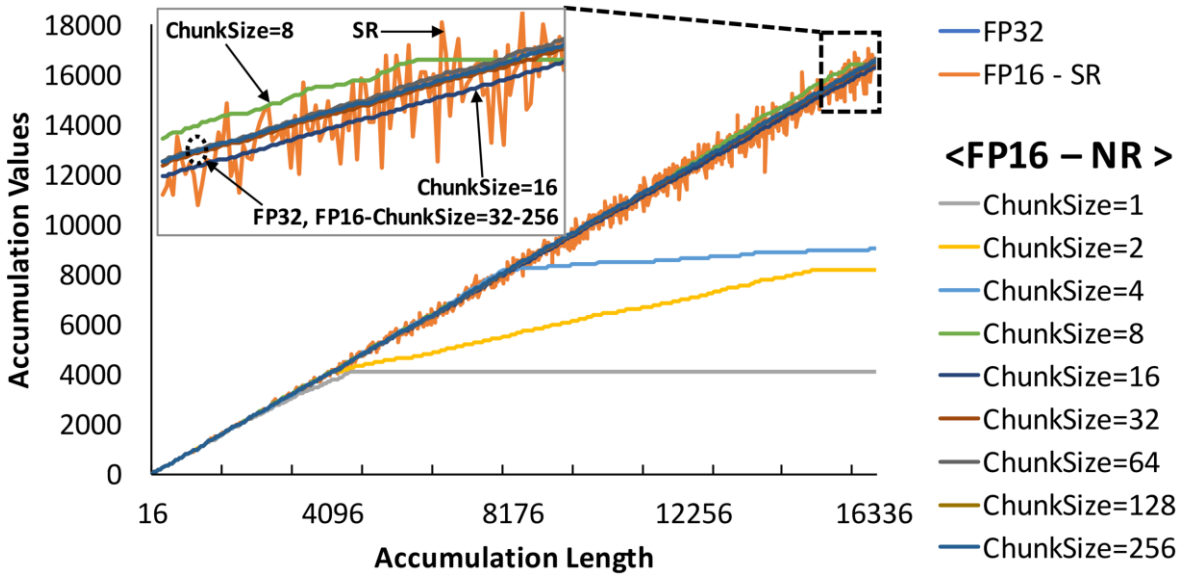


Feature 2 — Introduction of Stochastic Rounding

Principle & Advantages

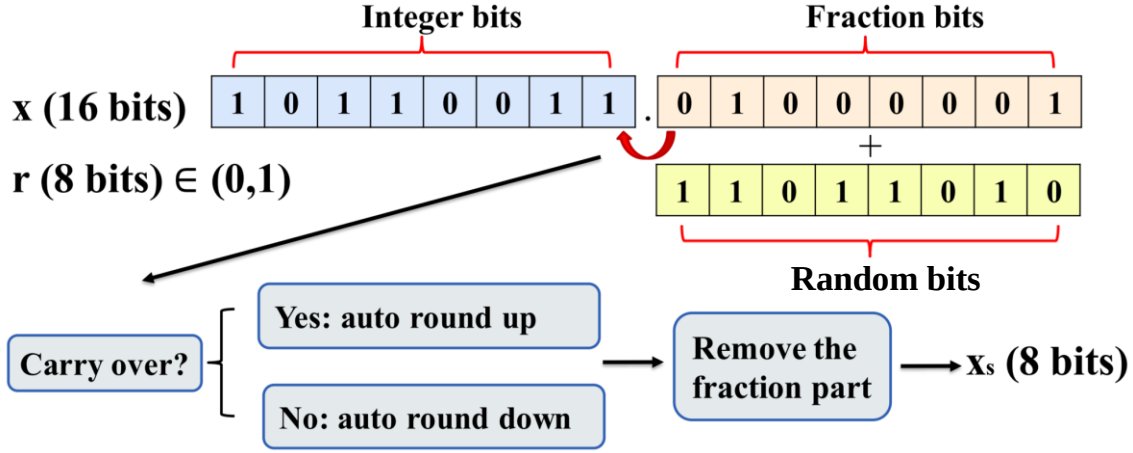
- $SR(x) = \begin{cases} \lfloor x \rfloor, & probability: 1 - (x - \lfloor x \rfloor) \\ \lceil x \rceil, & probability: x - \lfloor x \rfloor \end{cases}$
- Prevent **swamping** and training stagnation
- Reduce bit width and cost for accumulation

Comparison of FP16 and FP32



[Naigang Wang et al., NeurIPS 2018]

Hardware Implementation of SR



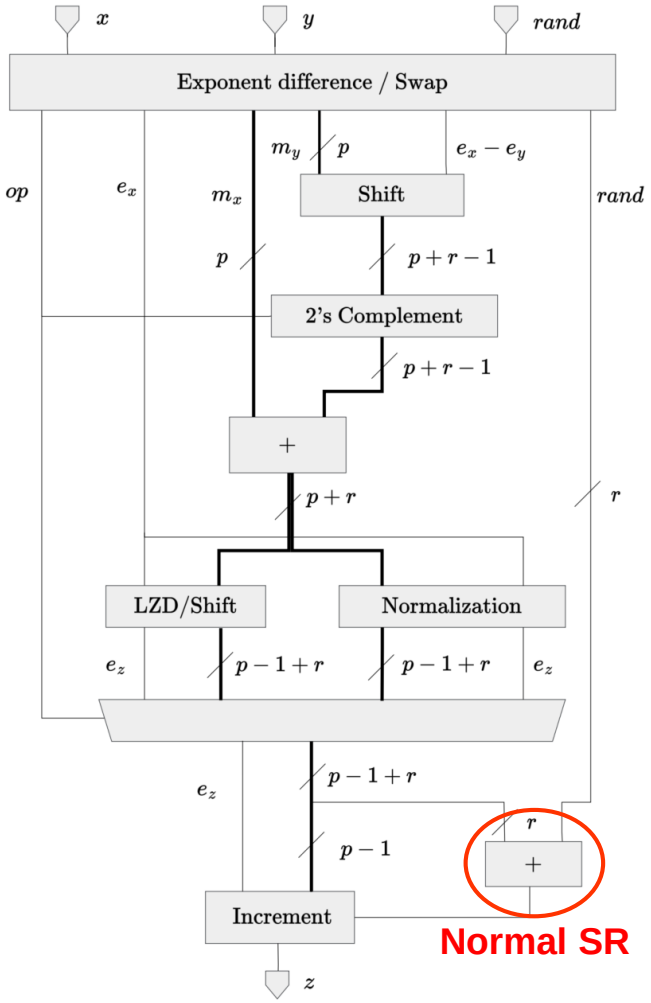
Example of 2 Fraction Bits SR

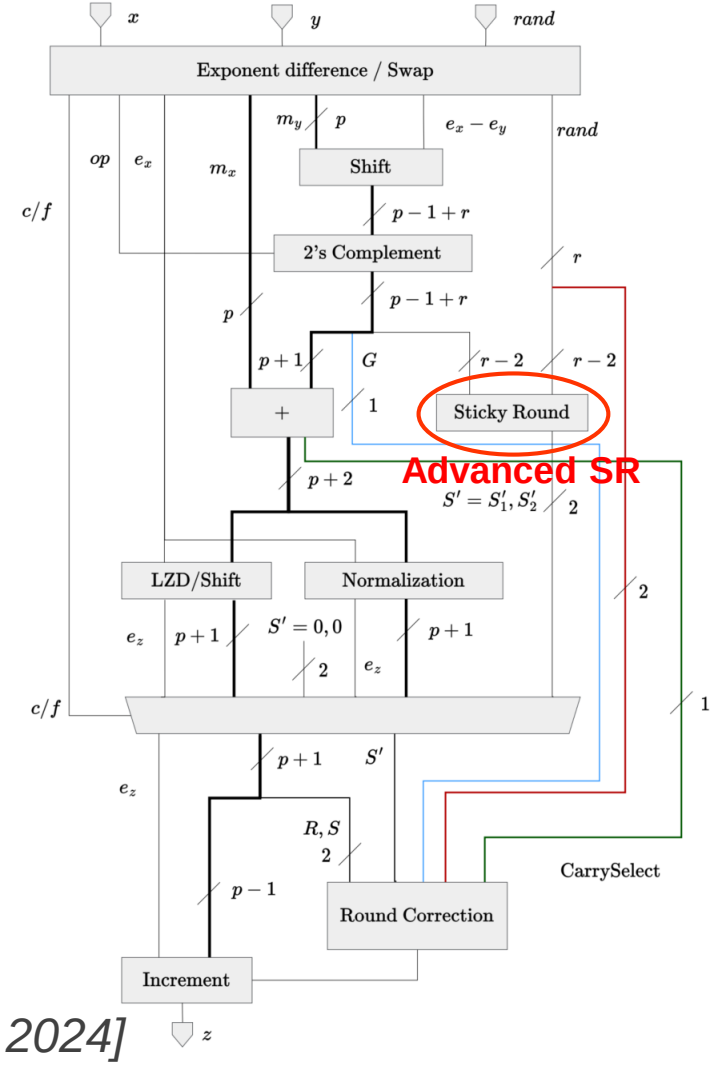
Fraction Bits	SR Result
00	0% round up
01	25% round up
10	50% round up
11	75% round up

[Sun Chang et al., DATE 2023]

Feature 2 — Introduction of Advanced SR

Comparison of 2 FP Adders with SR





[Sami Ben Ali et al., DATE 2024]

Motivation of Adv. SR

- SR **starts early** on addition
- **Correct** SR results at rounding
- Reduce bit with of LZD & Norm.

Advantage of Adv. SR

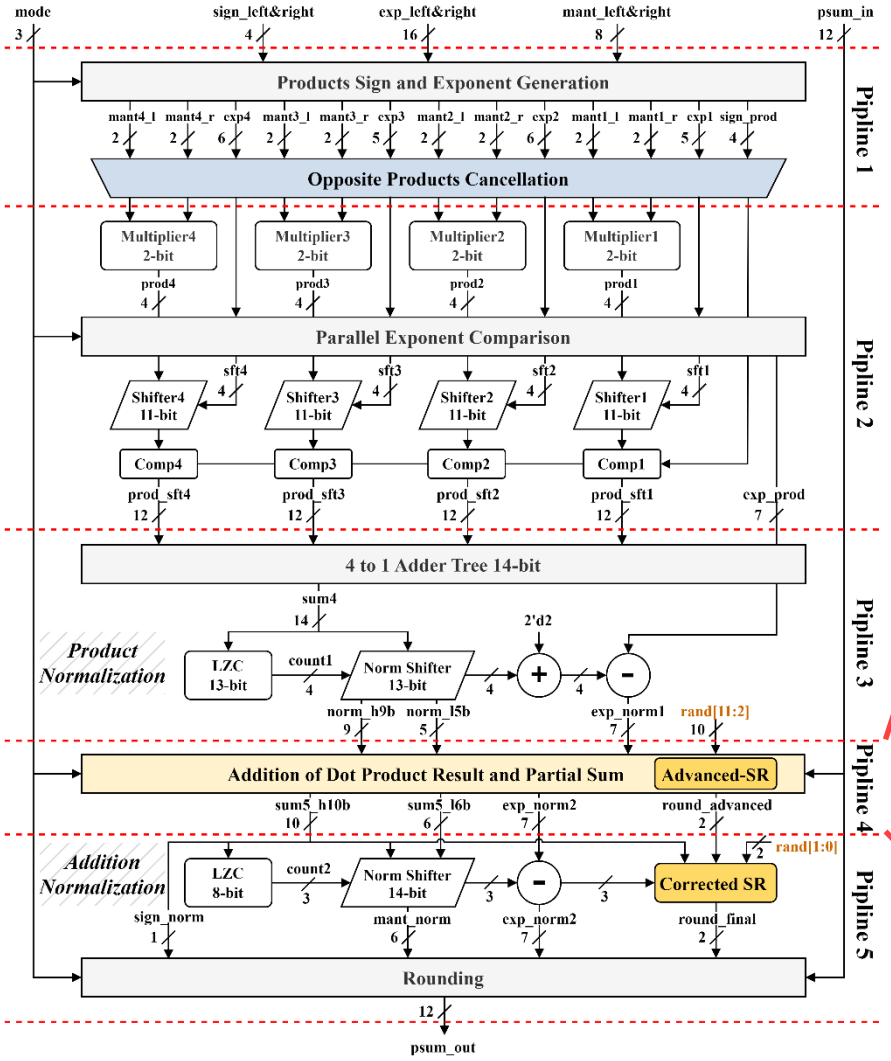
- Lower logic latency
- Smaller area & power usage
- Maintaining the accuracy

Existing Problems

1. Support for **FP addition only**
2. **Cannot be applied to any MAC**

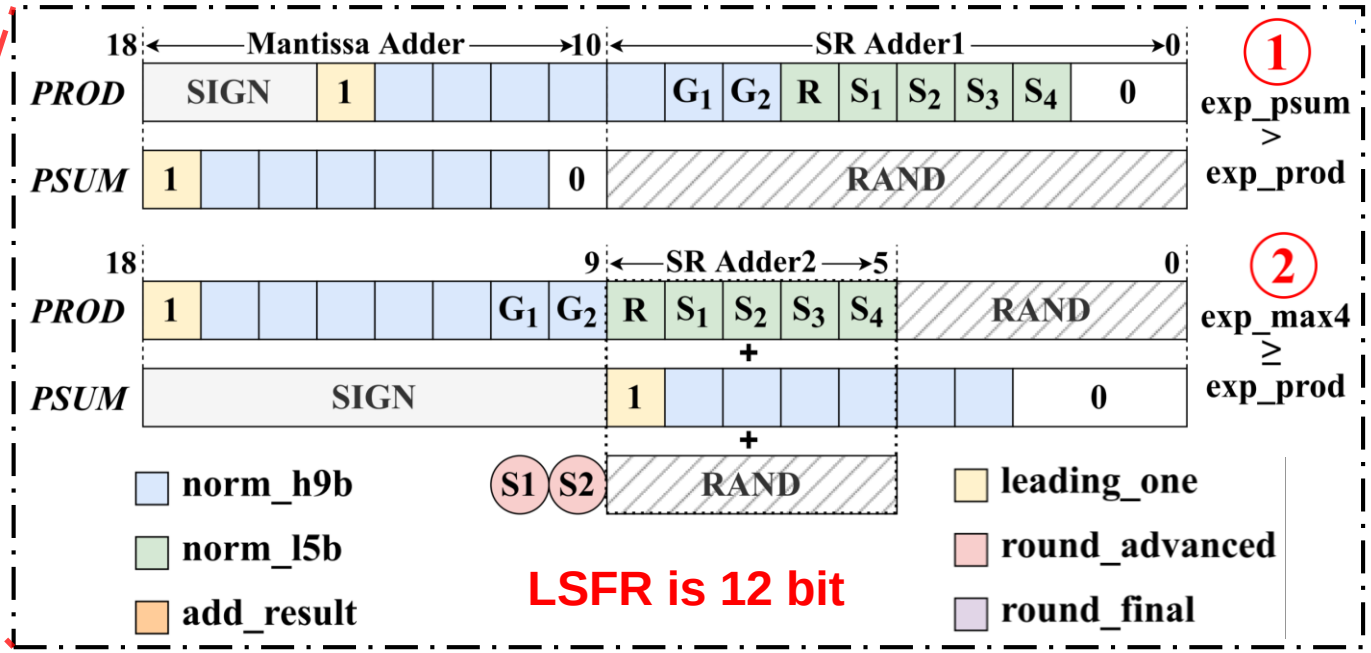
Feature 2 — Principle of Variable-SR

Multiple Dot Product Unit



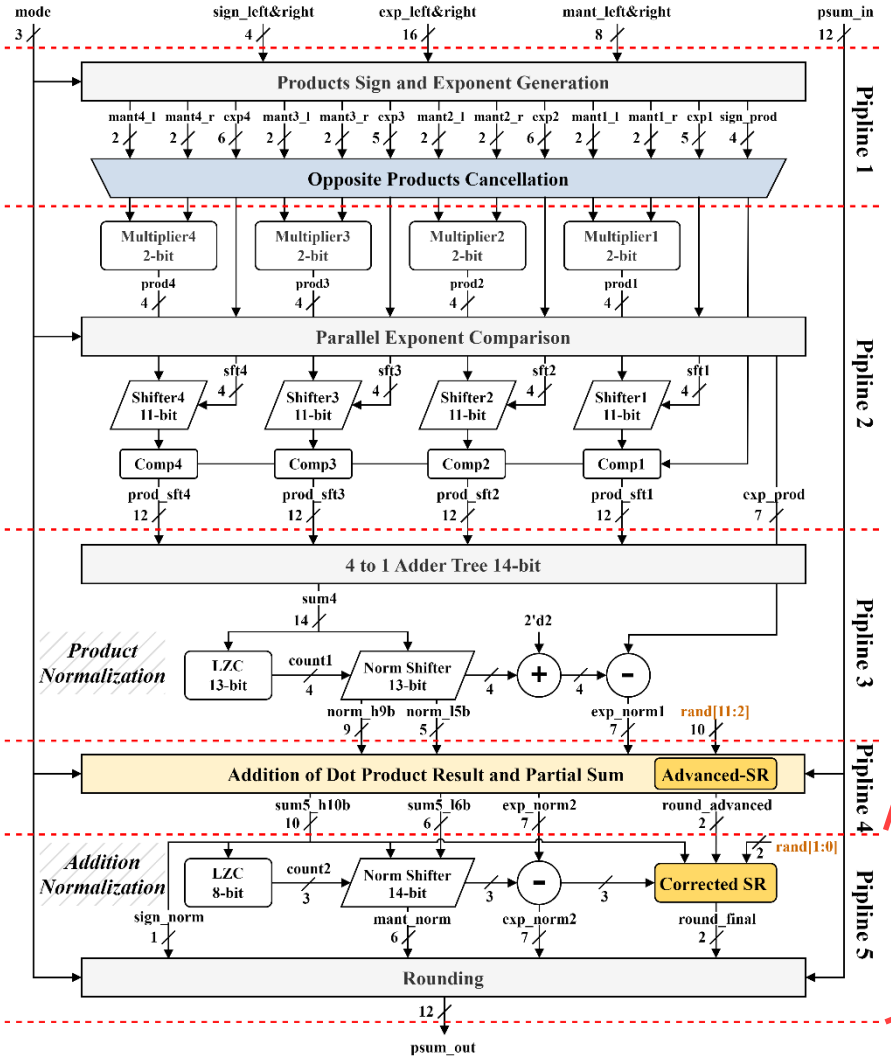
Variable-SR During Mantissa Addition

- Product's higher for mantissa ADD, lower for SR ADD
- Insert random number in exponent alignment by case
- ***Case 1:*** Random number inserted to PSUM's right
- ***Case 2:*** Inserted to PROD's right and SR Adder2



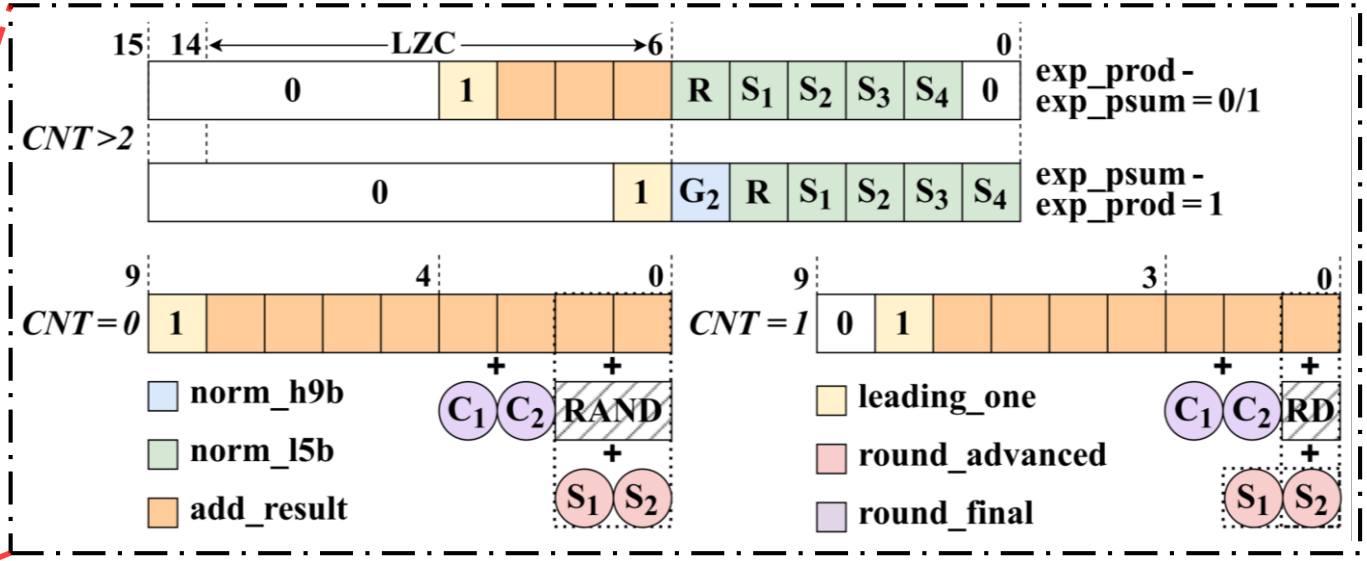
Feature 2 — Principle of Variable-SR

Multiple Dot Product Unit



SR Correction During Normalization

- Based on the count of leading zeros divided 4 cases
- ***CNT=0:*** Carry-over; Perform 2 bit addition to correct
- ***CNT=1:*** Hold; Perform 1 bit addition to correct
- ***CNT=2:*** Carry-back; No need to correct
- ***CNT≥3:*** Cancellation; No need to SR

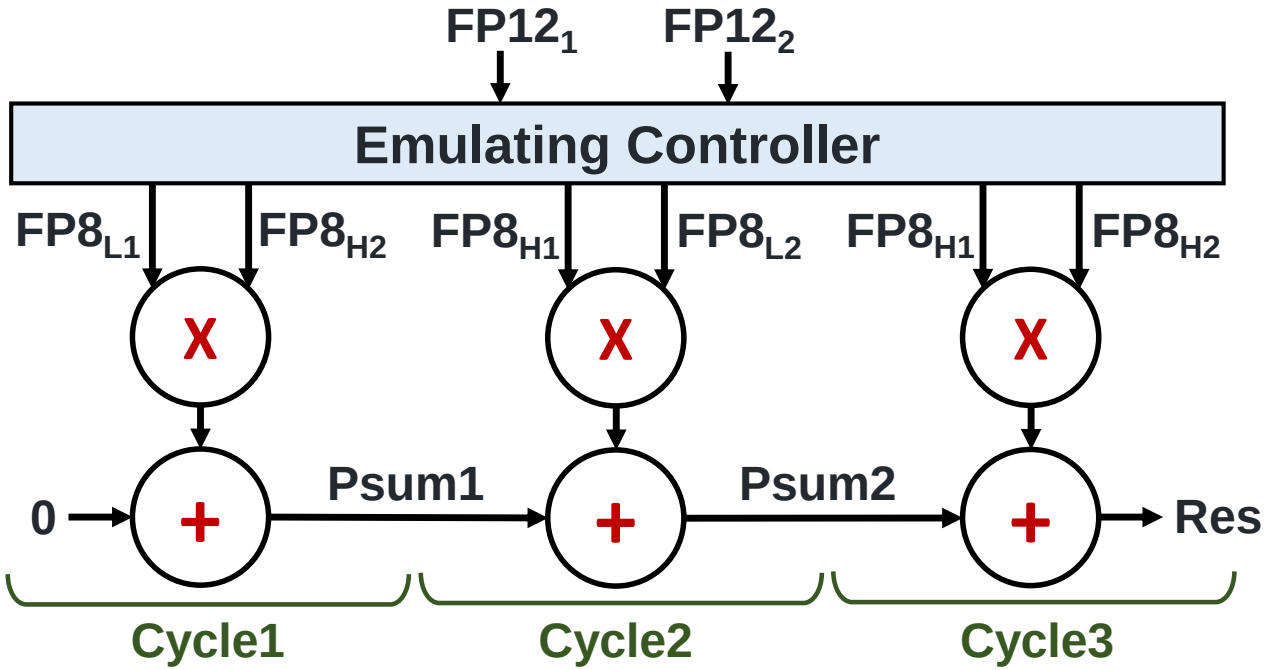


Feature 3 — High-Precision Emulating Controller

Realization of High-Precision Emulating

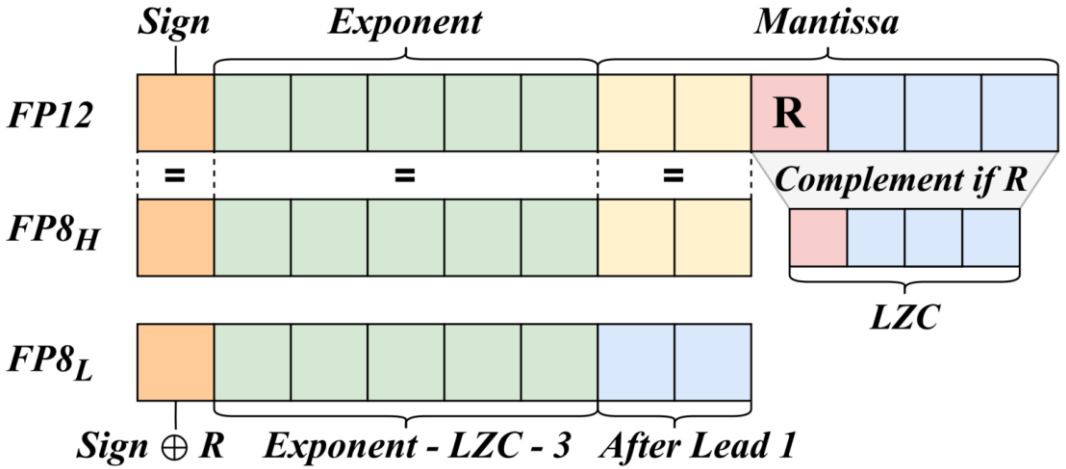
- $FP12_1 \times FP12_2 = FP8_{H1} \times FP8_{H2} + FP8_{H1} \times FP8_{L2} + FP8_{L1} \times FP8_{H2} + FP8_{L1} \times FP8_{L2}$ (Omitted)
- $FP12 \times FP8 = FP12_H \times FP8 + FP12_L \times FP8$

Multiplex Low-Precision MAC



Specific Decompose Process

- $FP8_H$ is FP12 truncated and rounded
- $FP8_L$ is smaller, indicating remainder



Accuracy Analysis

1. Accurate when FP12 exponent ≥ 7
2. Support subnormal to reduce 7 to 5

Outline

- Background & Challenges
- Overall Introduction of MPICC
- Key Features of MPICC
- **Experiment & Results**
  - Analysis of Different Accumulator Modes
  - Practical Results of Network Training
  - Accuracy for High-Precision Emulation
  - Performance Comparison
- **Conclusion**

Experiment Conditions

Hardware Experiment Conditions

- TSMC 28nm technology
- Under typical corner (TT, 0.9V, 25°C)
- Area from Synopsys Design Compiler 2021.09
- Power consumption from Synopsys PrimeTime PX 2021.06

Software Experiment Conditions

- ResNet-20 model on CIFAR-100 dataset
- SGD optimizer with an initial learning rate of 0.1
- Use gradient scaling, gradient clipping, and a momentum parameter of 0.9
- Train with a batch size of 128 for 200 epochs

Analysis of Different Accumulator Modes

Accuracy Analysis

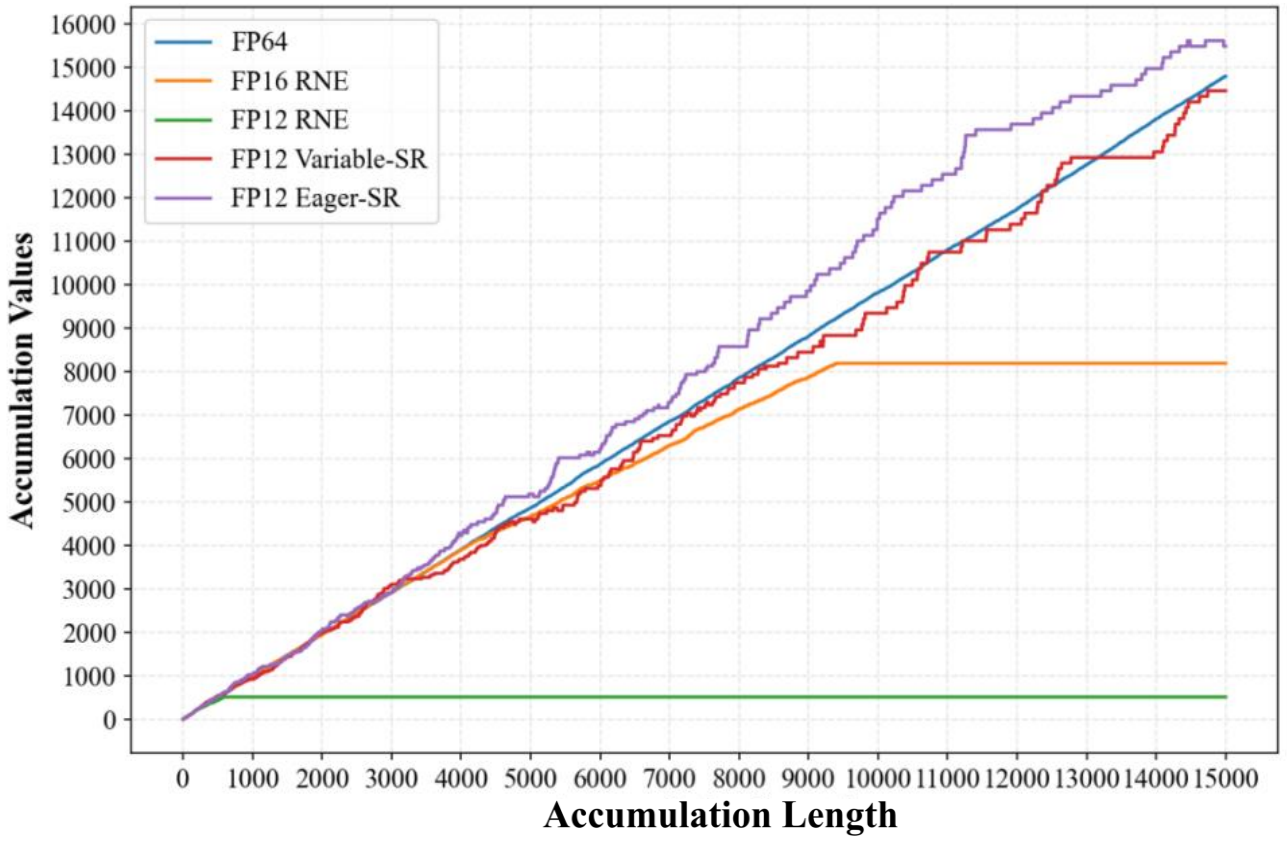


Figure 7: Accumulation results for different accumulators

Experiment Setup

- The operation mode is FP8×FP8
- One input set to 1
- The other following a uniform distribution (mean=1, standard deviation=1)
- Accumulated over 15,000 times

Experiment Result

- Nearest Even (RNE) leads to **swamping**
- Variable-SR is **accurate than** Eager-SR
- Former less rounded **1 time** than latter

[Eager-SR] *[Sami Ben Ali et al., DATE 2024]*

Analysis of Different Accumulator Modes

Experiment Setup

- Select accumulators in MPICC MACs
- Eager-SR performs SR on product to ensure that subsequent addition bit widths are equal

[10] [Sami Ben Ali et al., DATE 2024]

Experiment Result

- Compared with Conventional-SR, area/power reduced by **6.6%/11.3%**
- Compared with E5M10 RNE, delay/area/power reduced by **16.7%/15.7%/14.9%**

Table 3: Performance comparison of different accumulators

Configuration	Area(μm²)	Power(μW/MHz)	Delay(ns)	Accuracy
E5M6 RNE	320.81	0.282	5	Low
E5M6 Variable-SR	390.94	0.337	5	High
E5M6 Eager-SR[10]	401.16	0.342	5	Middle
E5M6 Conventional-SR	418.52	0.375	5	High
E5M10 RNE	463.64	0.396	6	High

Practical Results of Network Training

- **Experiment Result**
  - E8M6 RNE shows a **2.17%** degradation
  - E8M6 SR12 shows a **0.64%** degradation
  - E8M6 SR12 than E8M10 RNE **0.38%** accurate
- **Existing Improvements**
  - SR16 accuracy reduced due to small model
  - Not employing SOTA quantization scheme to compress 8 bit exponent to 5 bit

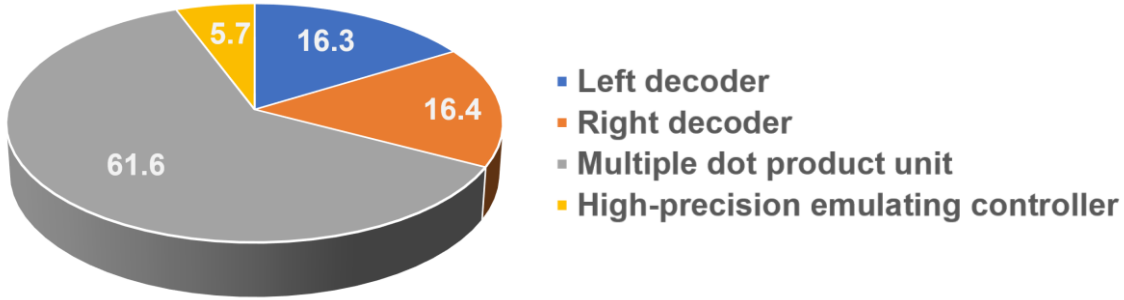
Table 4: Training accuracy with ResNet-20 on CIFAR-100

Precision	Rounding Mode	SR bit width	Accuracy (%)
E8M23	RNE	-	75.05
E8M10	RNE	-	74.03
E8M6	RNE	-	72.88
E8M6	SR	16	74.39
E8M6	SR	12	74.41
E8M6	SR	8	74.23
E8M6	SR	4	73.94

Act & Wt: FP8

Accuracy for High-Precision Emulation

Area Decomposition of MPICC MAC



Accuracy Result for High-Precision

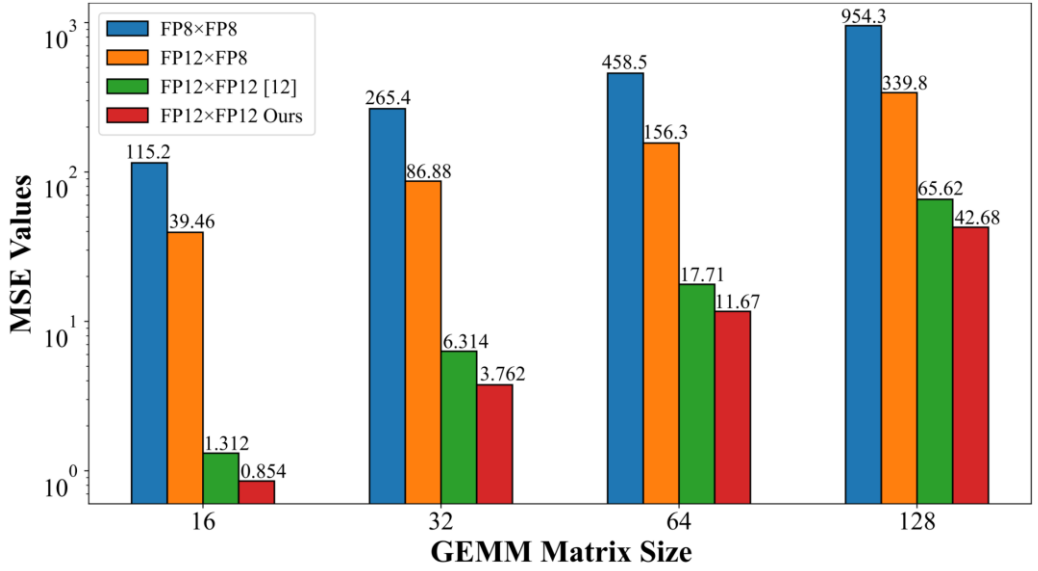


Figure 8: MSE for different sizes of GEMM computations

Experiment Setup

- Perform GEMM with random matrices of different sizes on MPICC MACs
- Direct precision decomposition in [12] means not using rounding or subnormal

Experiment Result

- FP8×FP8 MSE **3×** larger than FP12×FP8, **22-135×** larger than FP12×FP12 Ours
- Small accuracy improvement over [12]

[12] [*Stefano Markidis et al., IPDPSW 2023*]

Performance Comparison

Comparison Table & Figure

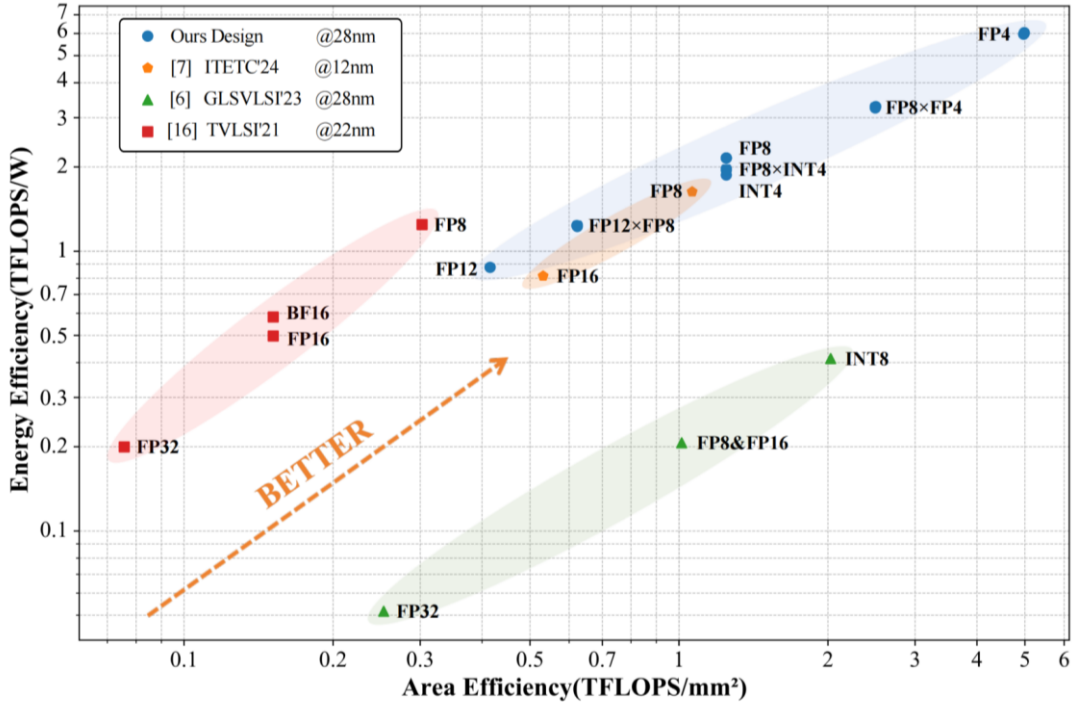


Figure 9: Comparison of area/energy efficiencies among PEs

Reference List

[6] *[Jing Zhang et al., GLSVLSI 2023]*
[7] *[Luca Bertaccini et al., NeurIPS 2024]*
[16] *[Zürich Switzerland et al., TVLSI 2021]*

Summary of Improvement

- Support widest (**18**) precision combinations
- FP8 area/energy efficiencies up **1.17×/1.19×**
- FP4 area/energy efficiencies up **4.69×/3.64×**

Improvement Analysis

- ***Focused Design:*** Target ultra-low-precision, **eliminate support** for redundant precisions
- ***Integrated Multiple-Precision:*** **Not combine** various single-precision PEs for efficiency
- ***MPICC:*** Use **direct computation**, eliminate precision conversion overhead

Outline

- Background & Challenges
- Overall Introduction of MPICC
- Key Features of MPICC
- Experiment & Results
- **Conclusion**

Conclusion

- We present a MPICC MAC unit to **maximize the hardware performance** brought by the ultra-low-precision training.

Table 1: Characteristics comparison of PEs for deep learning

Design	Low-Precision Supported					Stochastic Rounding	MPICC	High-Precision Emulation
	FP16	FP8	FP4	LOG4	INT4			
Ours	○	✓	✓	✓	✓	✓	✓	✓
[16]	✓	✓	×	×	×	×	×	×
[7]	✓	✓	×	×	×	✓	×	×
[10]	×	×	✓	×	✓	×	✓	×

○ *Supports FP12 with the same accumulation accuracy as FP16*

- Three key features are proposed:
  - **A MPICC architecture** that supports inter-computations among multiple precisions;
  - **A Variable-SR strategy** to reduce accumulator bit width while maintaining accuracy;
  - **A high-precision emulating controller** to support high-precision at low cost.
- Compared to SOTA designs, our design supports the **widest precisions** and improves area/energy efficiencies by **1.17×/1.19×**(FP8) & **4.69×/3.64×**(FP4).

Thanks for Attention!

Q&A