A 12nm 18.1TFLOPs/W Sparse Transformer Processor with Entropy-Based Early Exit, Mixed-Precision Predication, and Fine-Grained Power Management

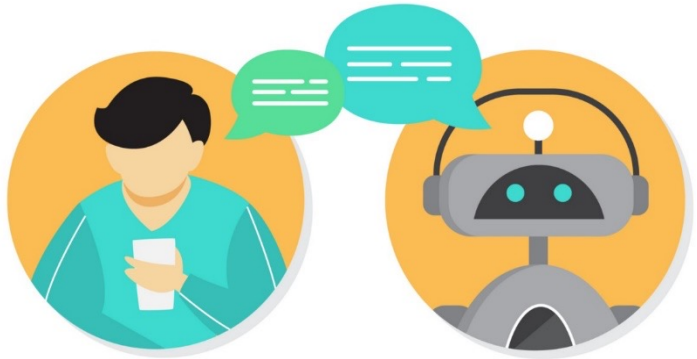
Thierry Tambe[1], Jeff Zhang[1], Coleman Hooper[1], Tianyu Jia[2], Paul N. Whatmough[1,3], Joseph Zuckerman[4], Maico Cassel[4], Erik Loscalzo[4], Davide Giri[4], Kenneth Shepard[4], Luca Carloni[4], Alexander M. Rush[5], David Brooks[1], Gu-Yeon Wei[1]

[1]Harvard University, Cambridge, MA, [2]Peking University, Beijing, China,
[3]ARM, Boston, MA, [4]Columbia University, New York, NY,
[5]Cornell University, New York, NY

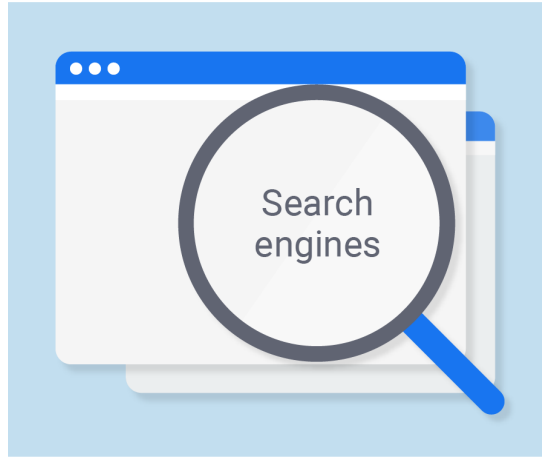
Harvard John A. Paulson School of Engineering and Applied Sciences

ML-based NLP is applied widely

Language Modeling & Understanding



Chat Bots (e.g., ChatGPT)



Search Engines

Understanding searches better than ever before

Oct 25, 2019 · 5 min read

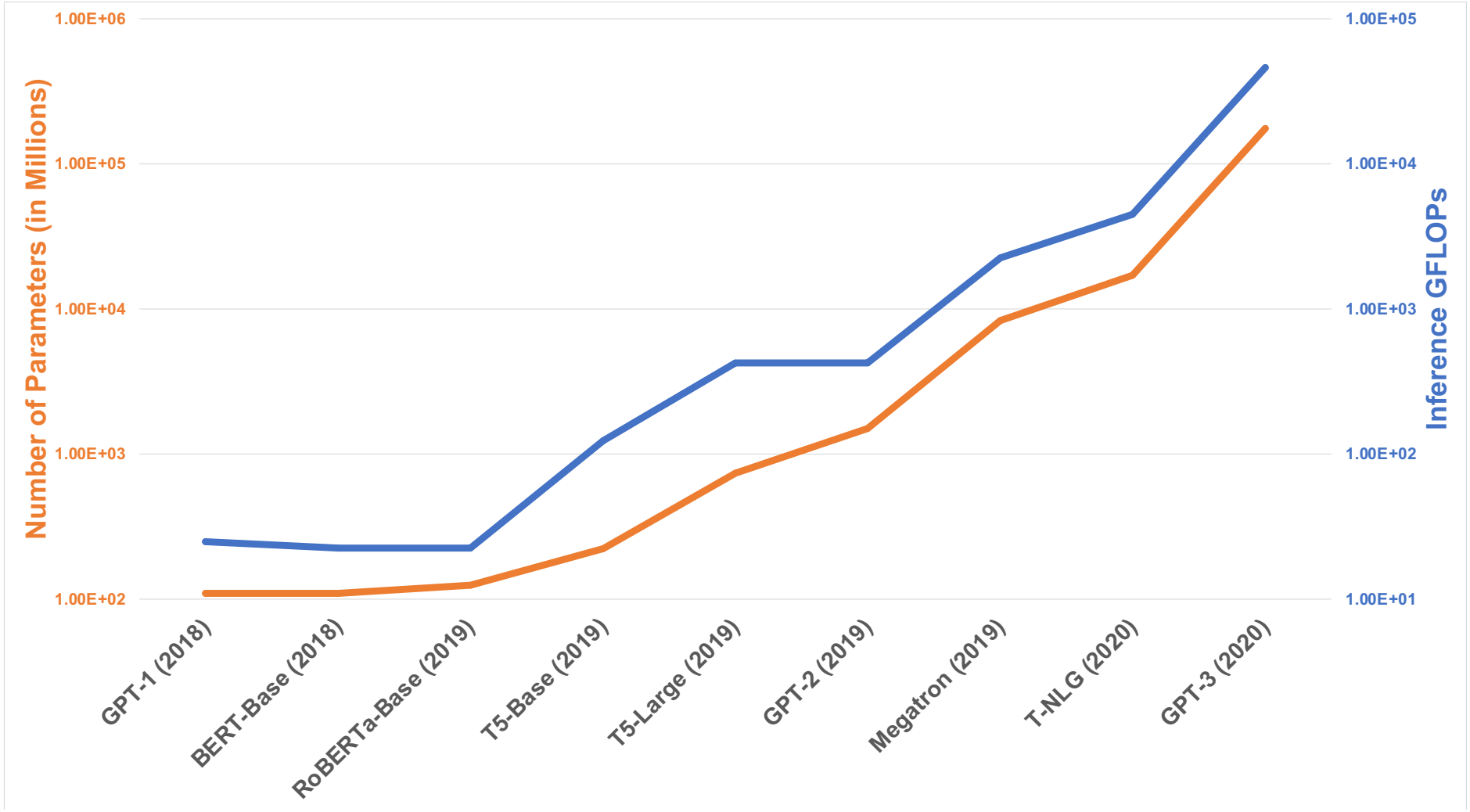
https://blog.google/products/search/search-language-understanding-bert/

Bing delivers its largest improvement in search experience using Azure GPUs

Posted on November 18, 2019

https://azure.microsoft.com/en-us/blog/bing-delivers-its-largest-improvement-in-search-experience-using-azure-gpus/

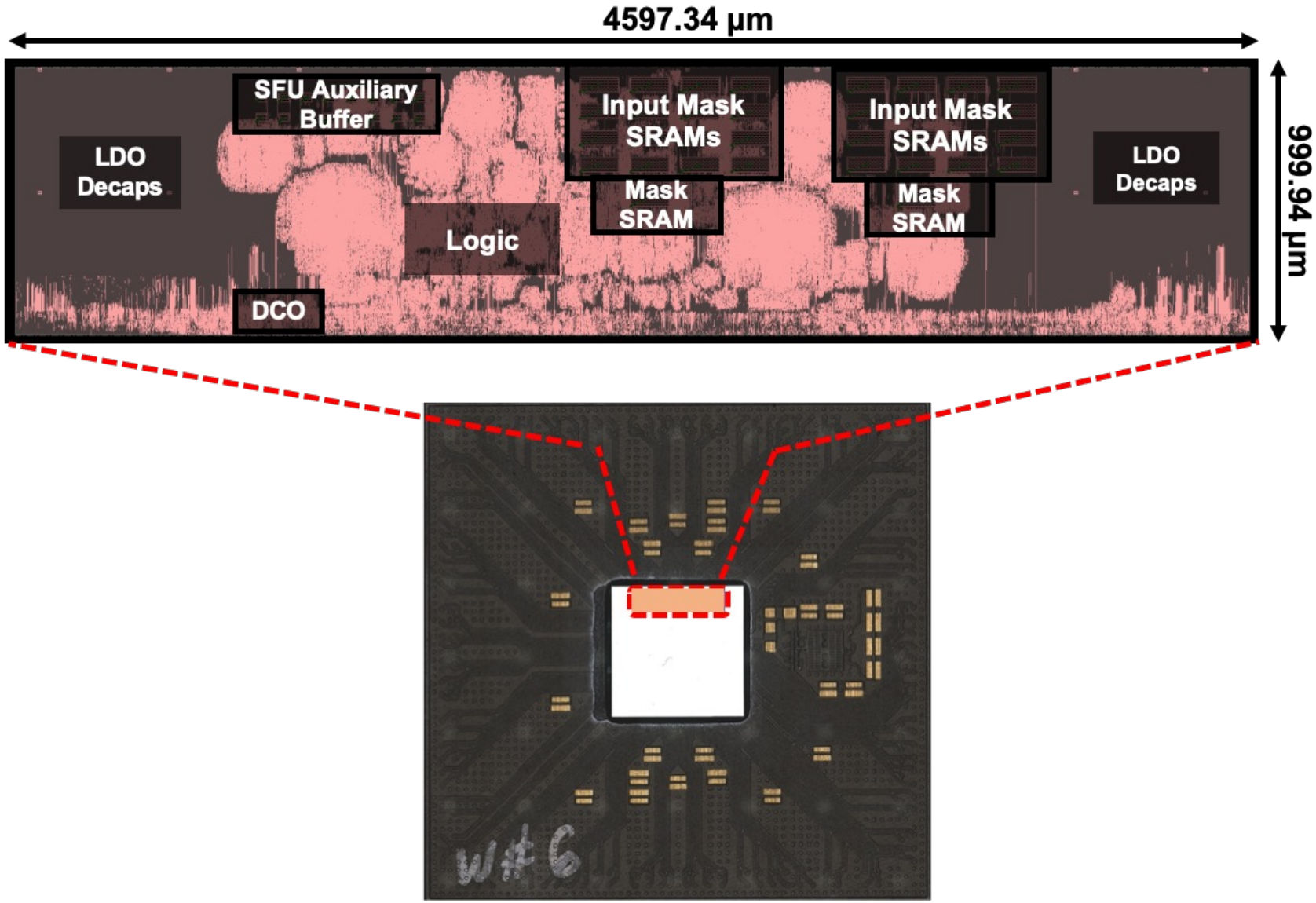
NLP Growing Overhead

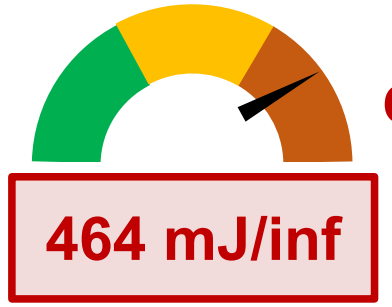


Source:
https://amatriain.net/blog/transformer-models-an-introduction-and-catalog-2d1e9039f376/

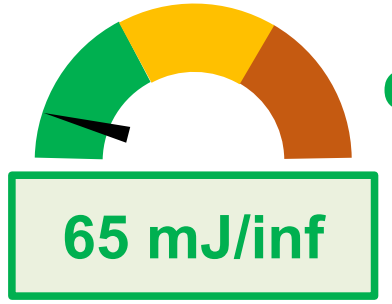
- **Opportunities to achieve higher energy efficiency on edge devices via careful algorithm-hardware co-design**

Processor for Efficient Transformer Computation





7x energy reduction

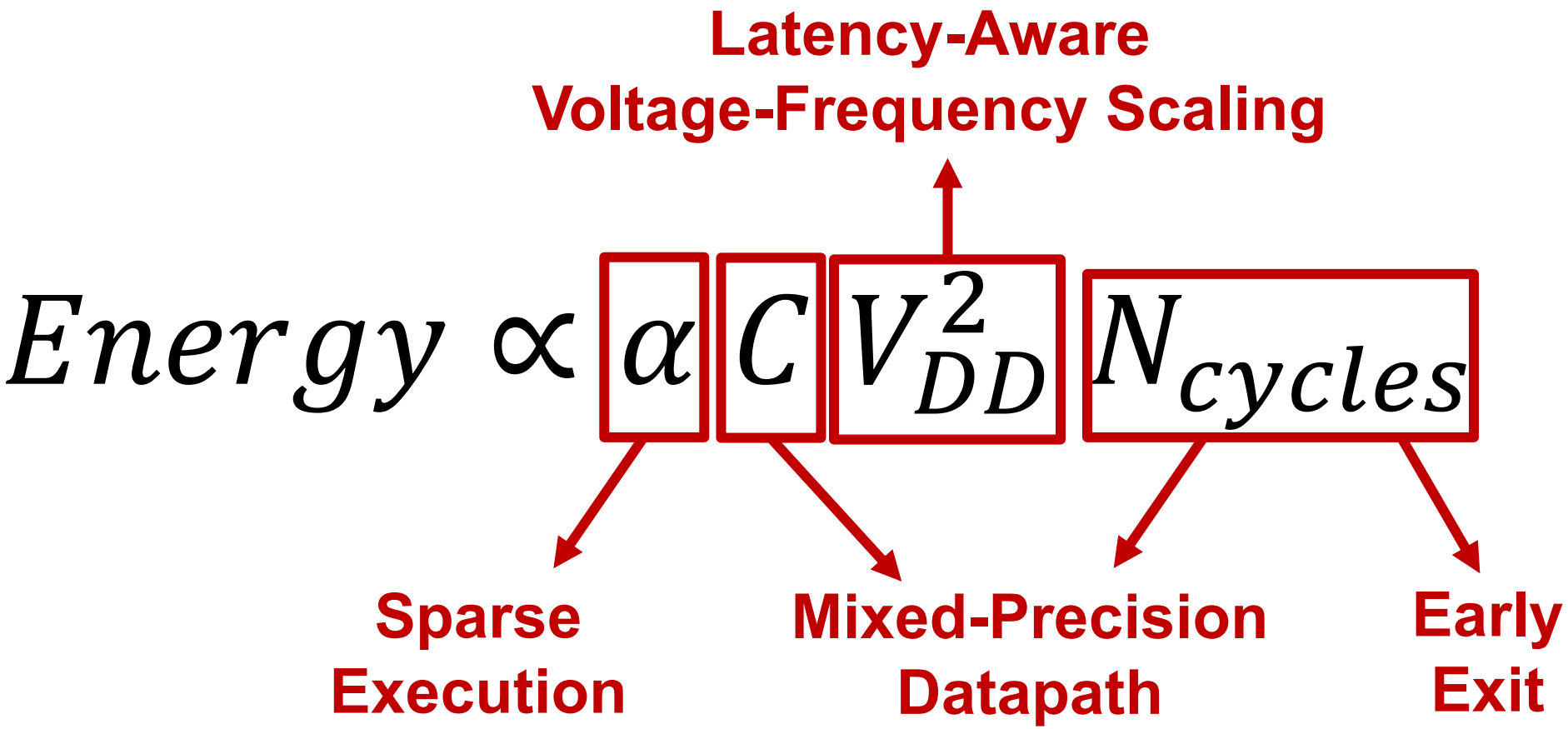


Abstracting Energy Consumption

$$Energy \propto \alpha\ C\ V_{DD}^2\ N_{cycles}$$

- α – switching activity factor
- C – wire and device capacitance
- V_{DD}^2 – supply voltage
- N_{cycles} – # of inference clock cycles

Proposed Optimization Schemes



Proposed Optimization Schemes

Entropy

Fine-Grained Sentence-Level Latency and Energy Optimization

Latency-Aware Voltage-Frequency Scaling

$$\propto \alpha \, C \, V_{DD}^2 \, N_{cycles}$$

Sparse Execution

Mixed-Precision Datapath

Early Exit

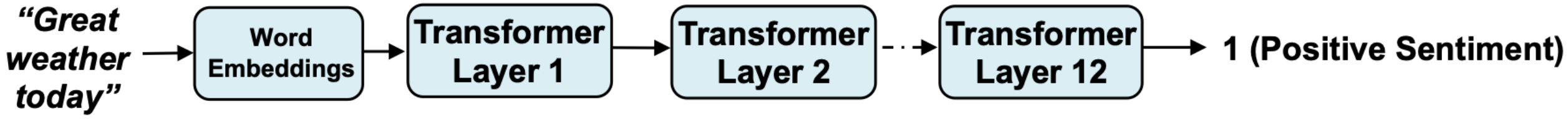
Outline

- Motivation
- Entropy-Driven Optimizations
  - Early Exit
  - Latency-Aware Voltage-Frequency Scaling
- 12nm Transformer Accelerator Architecture
  - Mixed-Precision FP4/FP8 Datapath
- Chip Measurement Results
- Summary

Outline

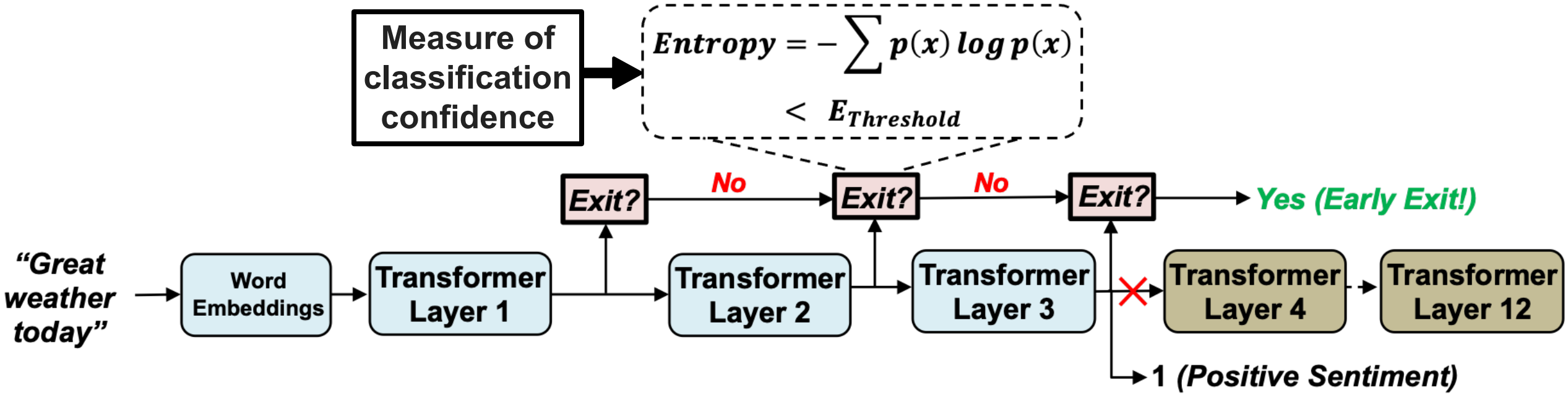
- Motivation
- **Entropy-Driven Optimizations**
  - **Early Exit**
  - **Latency-Aware Voltage-Frequency Scaling**
- 12nm Transformer Accelerator Architecture
  - Mixed-Precision FP4/FP8 Datapath
- Chip Measurement Results
- Summary

Conventional BERT Inference



Source: Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018)

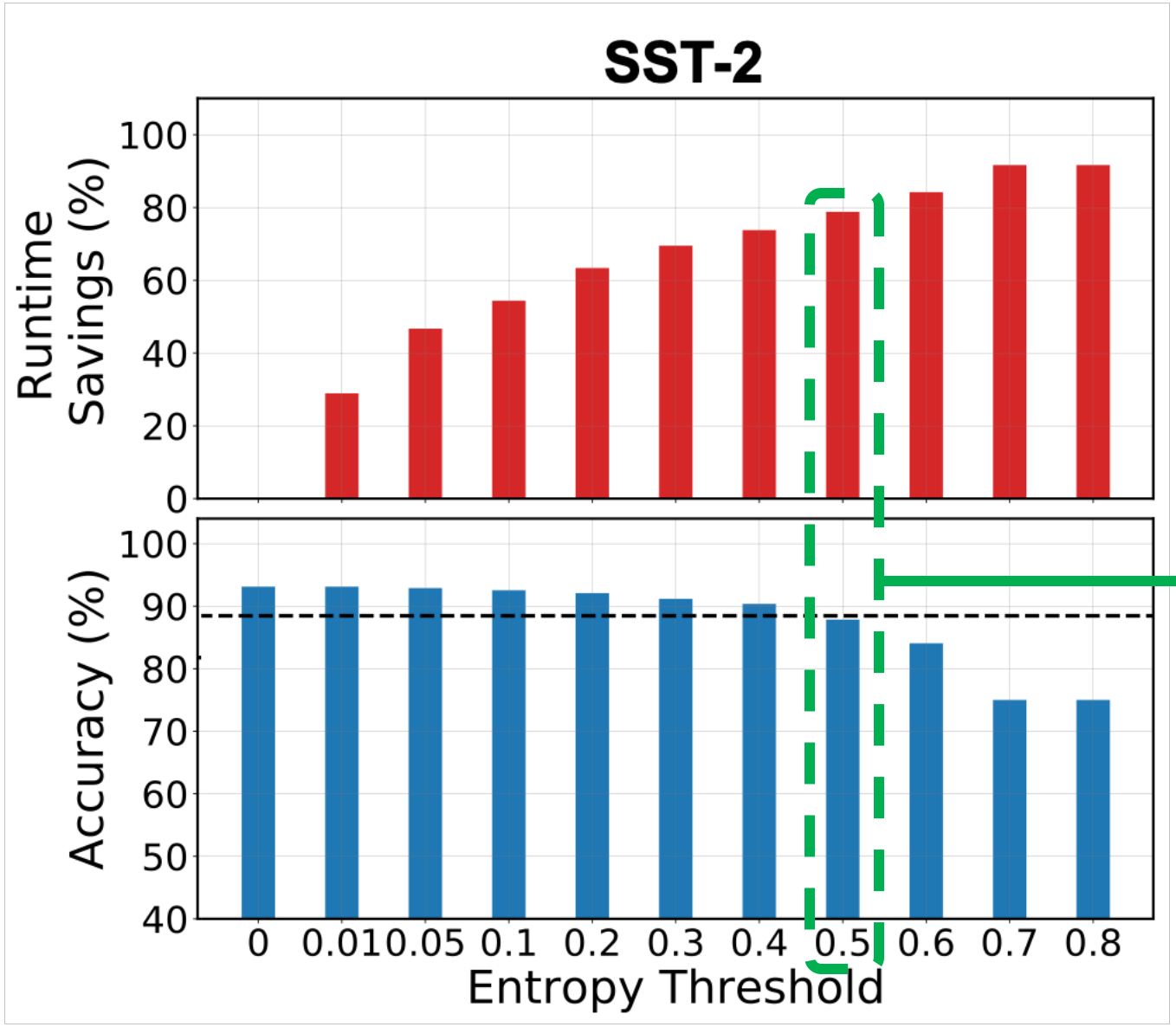
BERT Inference with Entropy-based Early Exit



- Inference exits early if the entropy is smaller than a user-given threshold

Source: Xin, Ji, et al. "DeeBERT: Dynamic early exiting for accelerating BERT inference." *arXiv preprint arXiv:2004.12993* (2020).

Significant Latency Savings

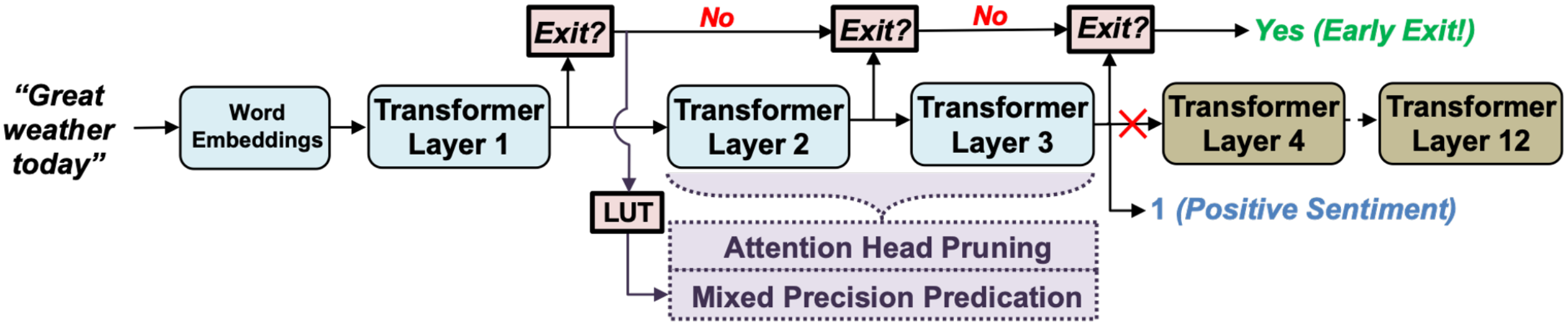


80% of BERT computations eliminated while maintaining 95% of the original accuracy!

Two Optimization Directions

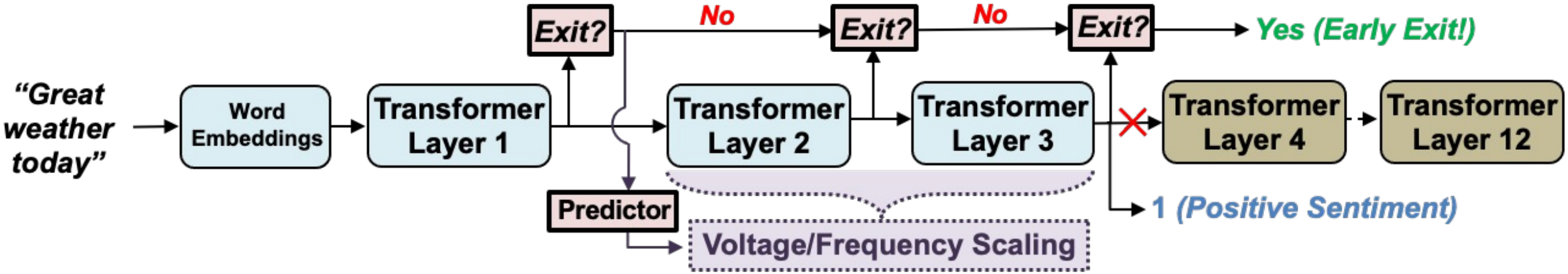
1. Latency minimization
2. Energy minimization

Latency Minimization



- Accelerator operates at max frequency with early exit
- Attention head pruning and mixed-precision FP4/FP8 datapath further cut inference latency

Energy Minimization

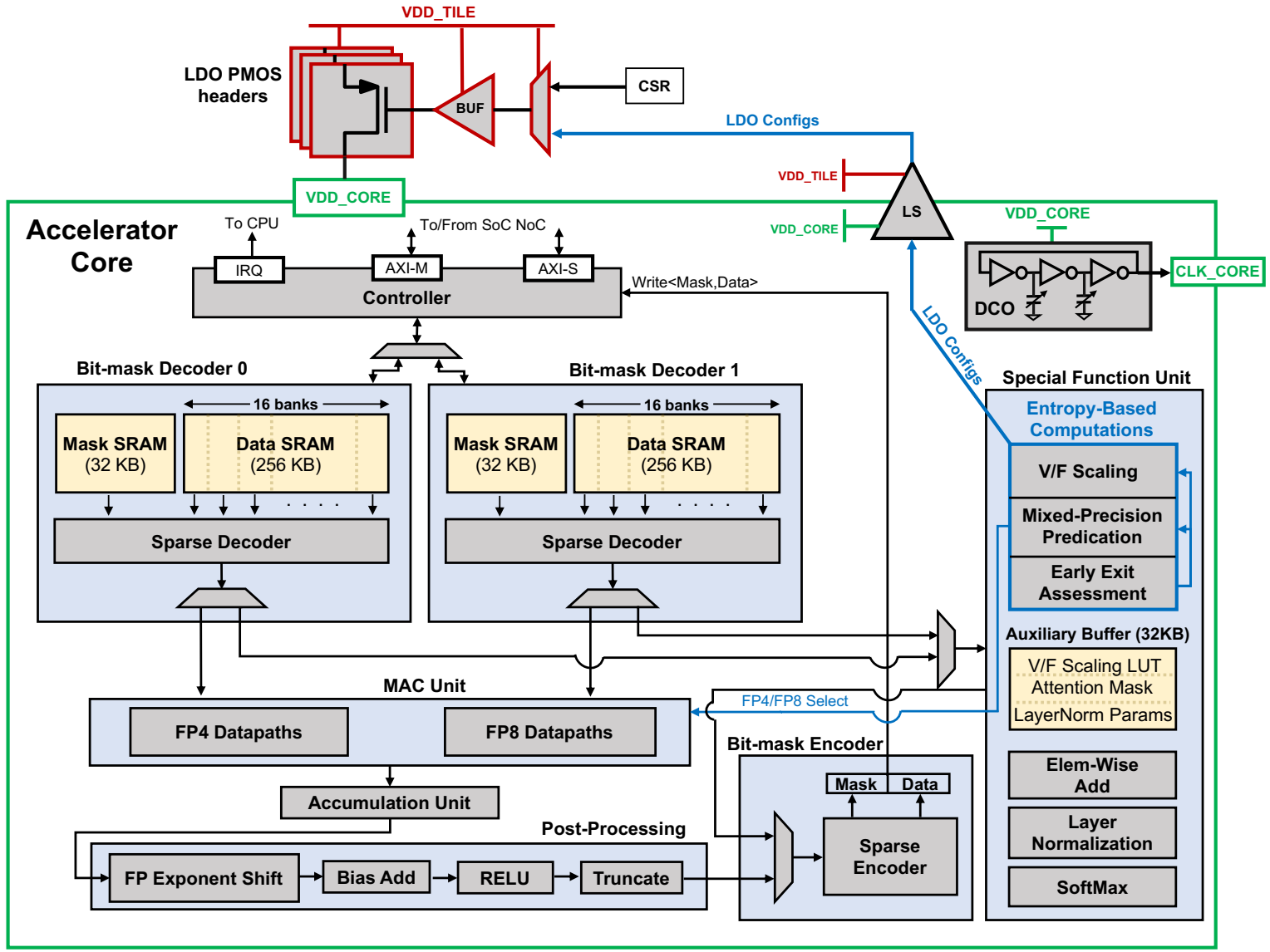


- Accelerator uses entropy statistics to derate its voltage and frequency while adhering to a prescribed latency target.

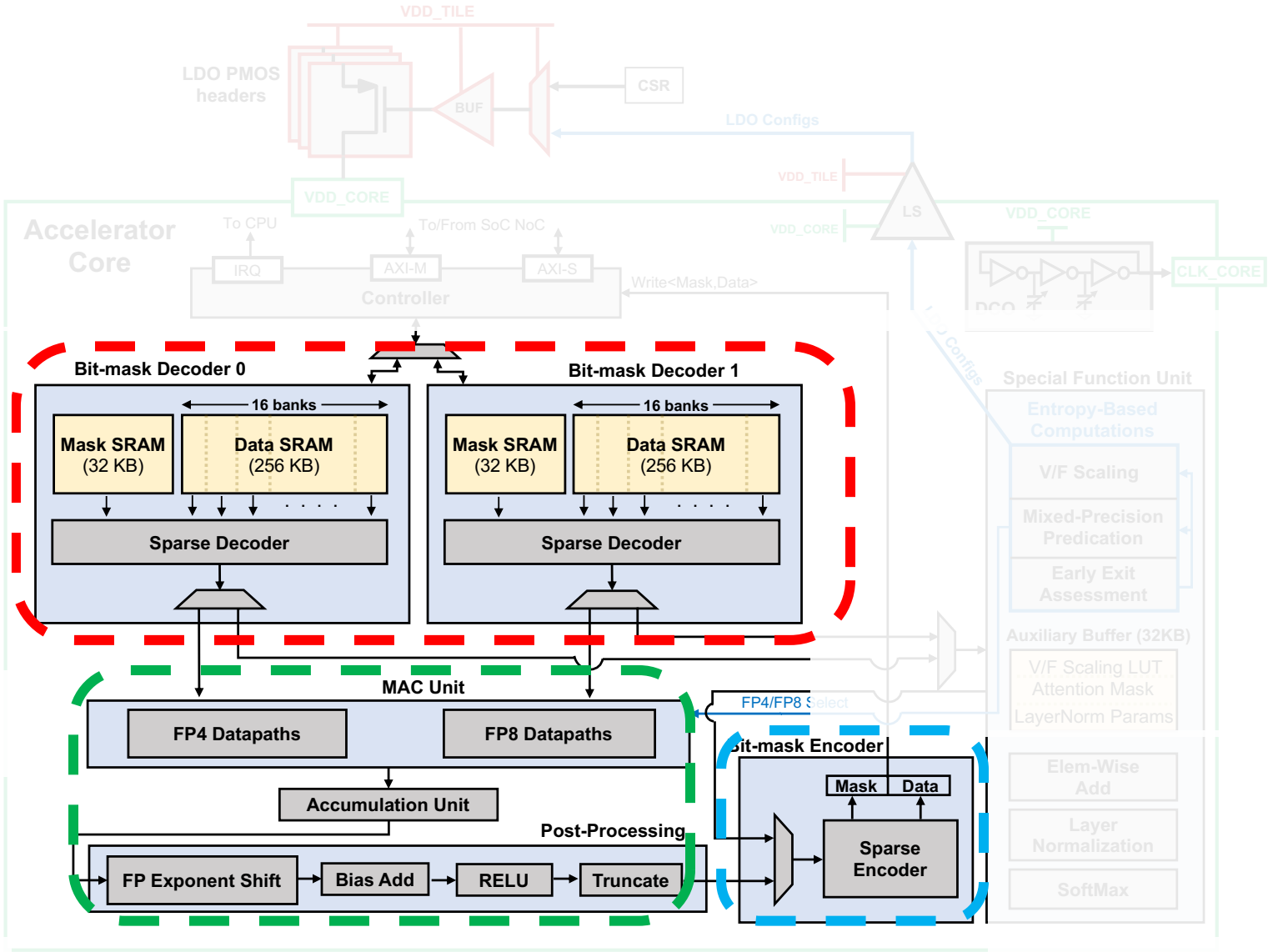
Outline

- Motivation
- Entropy-Driven Optimizations
  - Early Exit
  - Latency-Aware Voltage-Frequency Scaling
- **12nm Transformer Accelerator Architecture**
  - Mixed-Precision FP4/FP8 Datapath
- Chip Measurement Results
- Summary

Proposed Sparse Transformer Processor

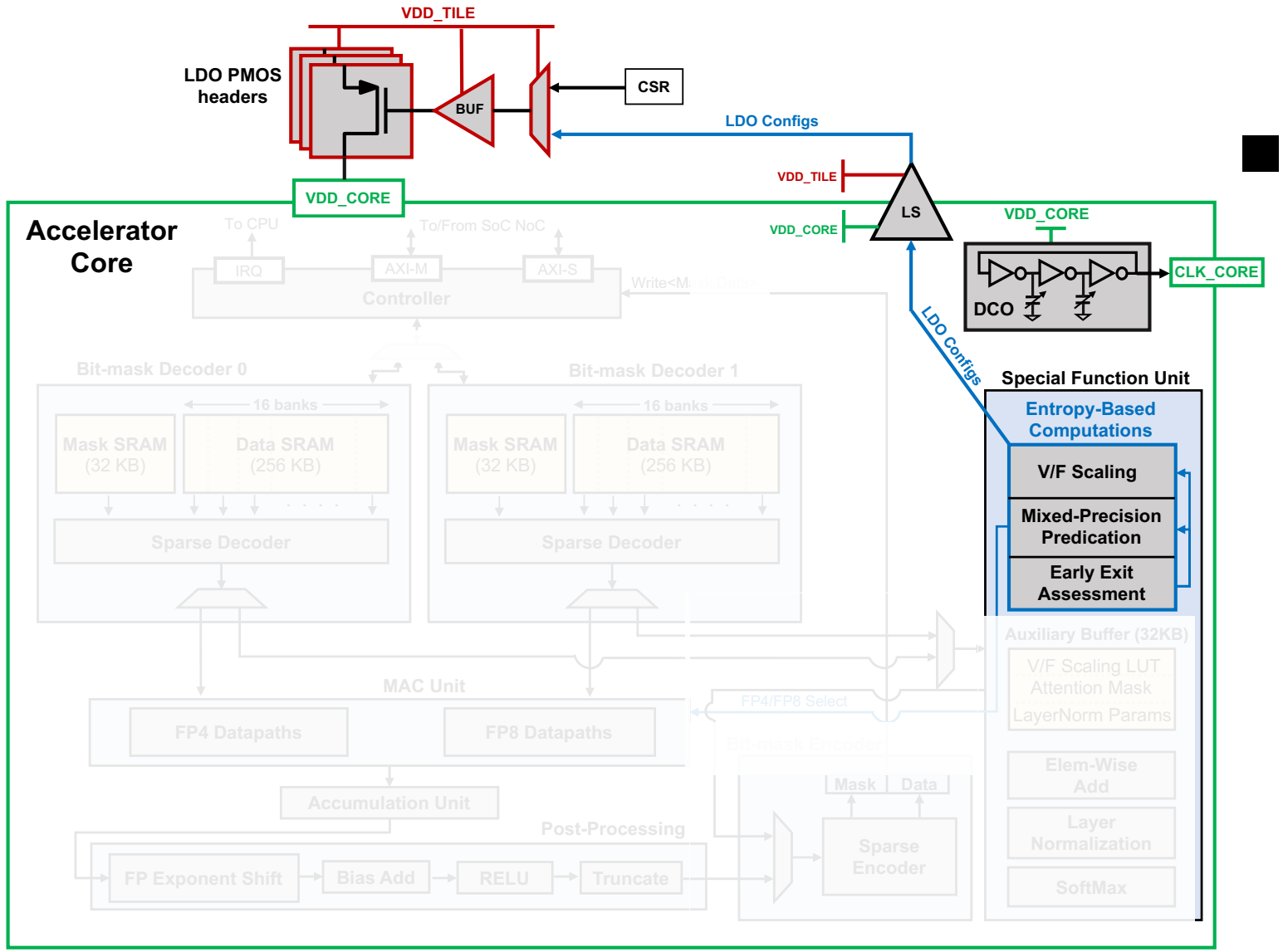


Compressed Sparse Mixed-Precision Execution



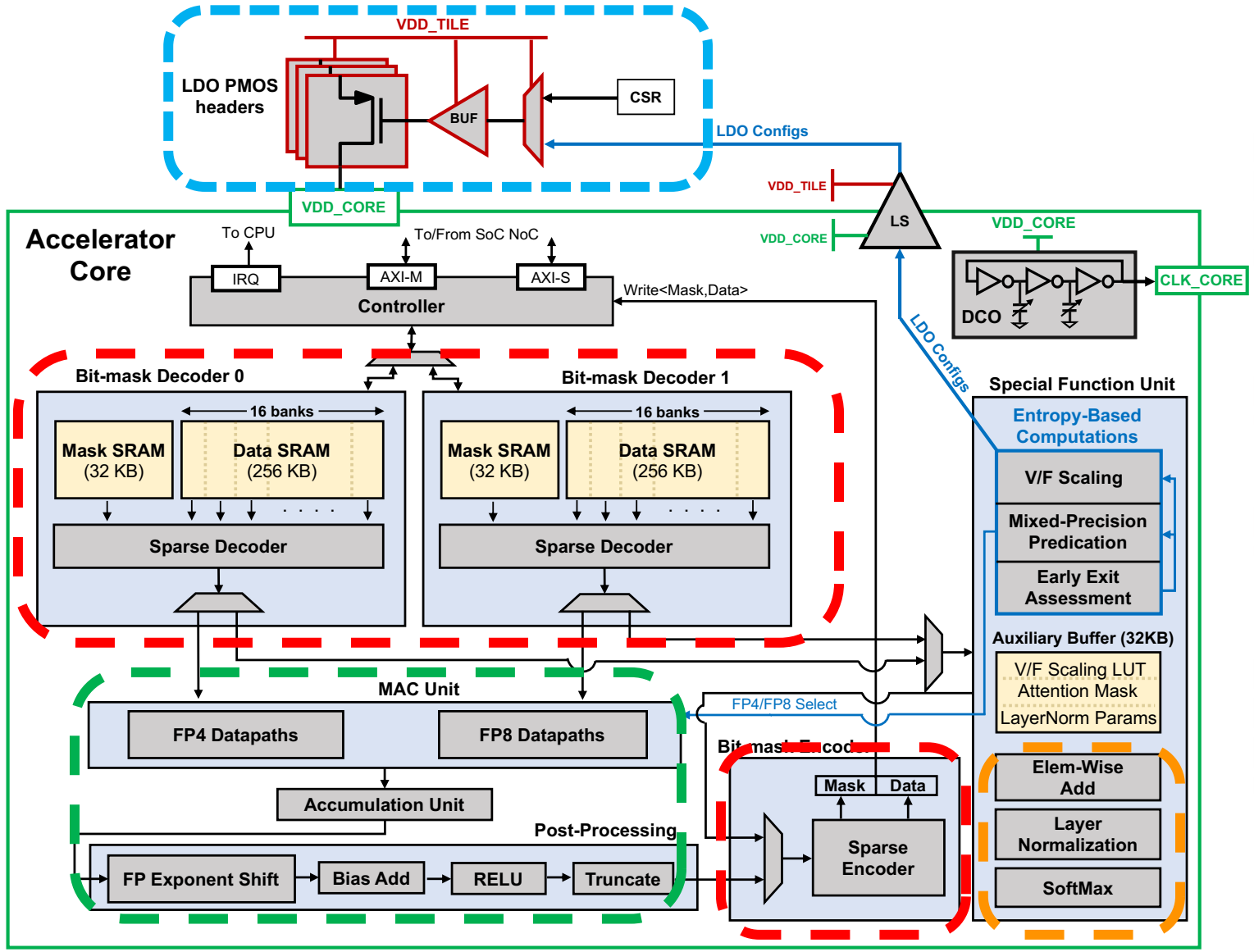
- **Bit-Mask Sparse Decoder**
- **Mixed-Precision FP4/FP8 Datapath**
- **Sparse Encoder**

Entropy-Controlled Voltage-Frequency Scaling



- **Entropy-Controlled Voltage-Frequency Scaling using:**
  - Open-loop free running LDO
  - Cell-based PMOS power headers
  - 16 pre-characterized LUT entropy values control the LDO drive strength
  - DCO powered from LDO output

Proposed Sparse Transformer Processor



- **Entropy-Controlled Voltage/Frequency Scaling**
- **Bit-Mask Encoding**
- **Special Function Unit**
- **Mixed-Precision FP4 (E3M0) / FP8 (E4M3) Datapath**

Outline

- Motivation
- Entropy-Driven Optimizations
  - Early Exit
  - Latency-Aware Voltage-Frequency Scaling
- 12nm Transformer Accelerator Architecture
  - **Mixed-Precision FP4/FP8 Datapath**
- Chip Measurement Results
- Summary

Efficient Number Systems

FP8 (E4M3)

$$-1^{sign} \times mant \times 2^{\alpha}$$

FP4/LOG4 (E3M0)

$$-1^{sign} \times 2^{\alpha/\gamma}$$

$\gamma \propto$ distance between values

Source: Zhao, Jiawei, et al. "LNS-Madam: Low-Precision Training in Logarithmic Number System Using Multiplicative Weight Update." *IEEE Transactions on Computers*, 2022.

Tensor Multiplication in Logarithmic Number System (LNS)

$$a = sign_a \times 2^{\tilde{a}/\gamma} \qquad b = sign_b \times 2^{\tilde{b}/\gamma}$$

$$a^T b = \sum XOR(sign_a, sign_b) \times 2^{(\tilde{a}+\tilde{b})/\gamma}$$

No Need for Multipliers!
(Only Adders + Shifters)

Source: Zhao, Jiawei, et al. "LNS-Madam: Low-Precision Training in Logarithmic Number System Using Multiplicative Weight Update." *IEEE Transactions on Computers*, 2022.

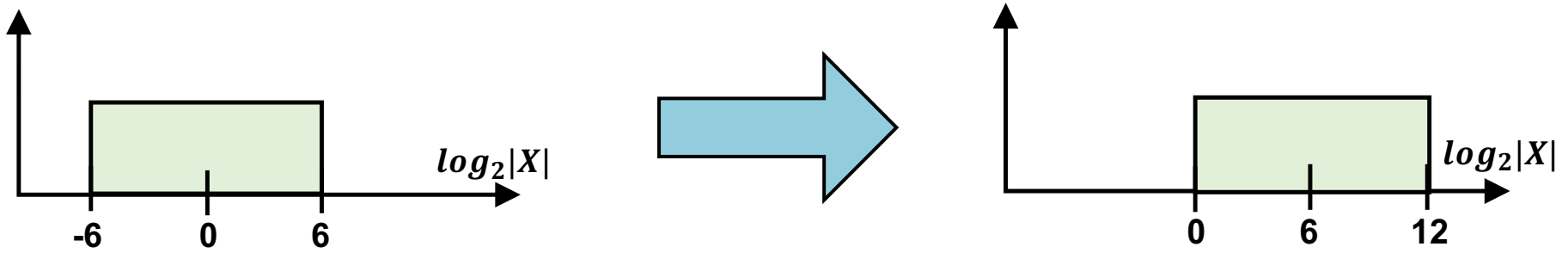
Tensor Scaling

FP8 (E4M3)

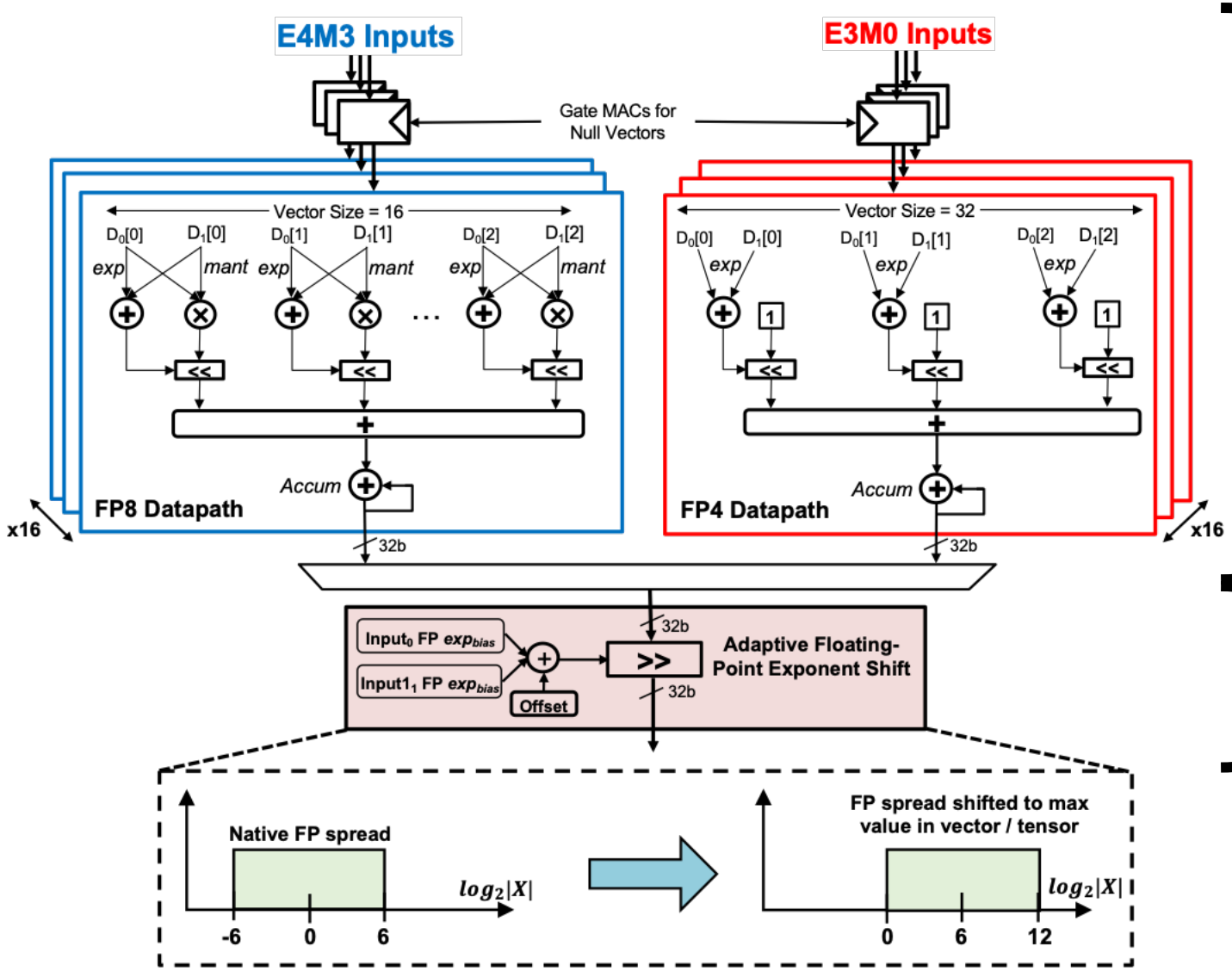
$$-1^{sign} \times mant \times 2^{\alpha + \boldsymbol{bias}}$$

FP4/LOG4 (E3M0)

$$-1^{sign} \times 2^{\alpha/\gamma + \boldsymbol{bias}}$$



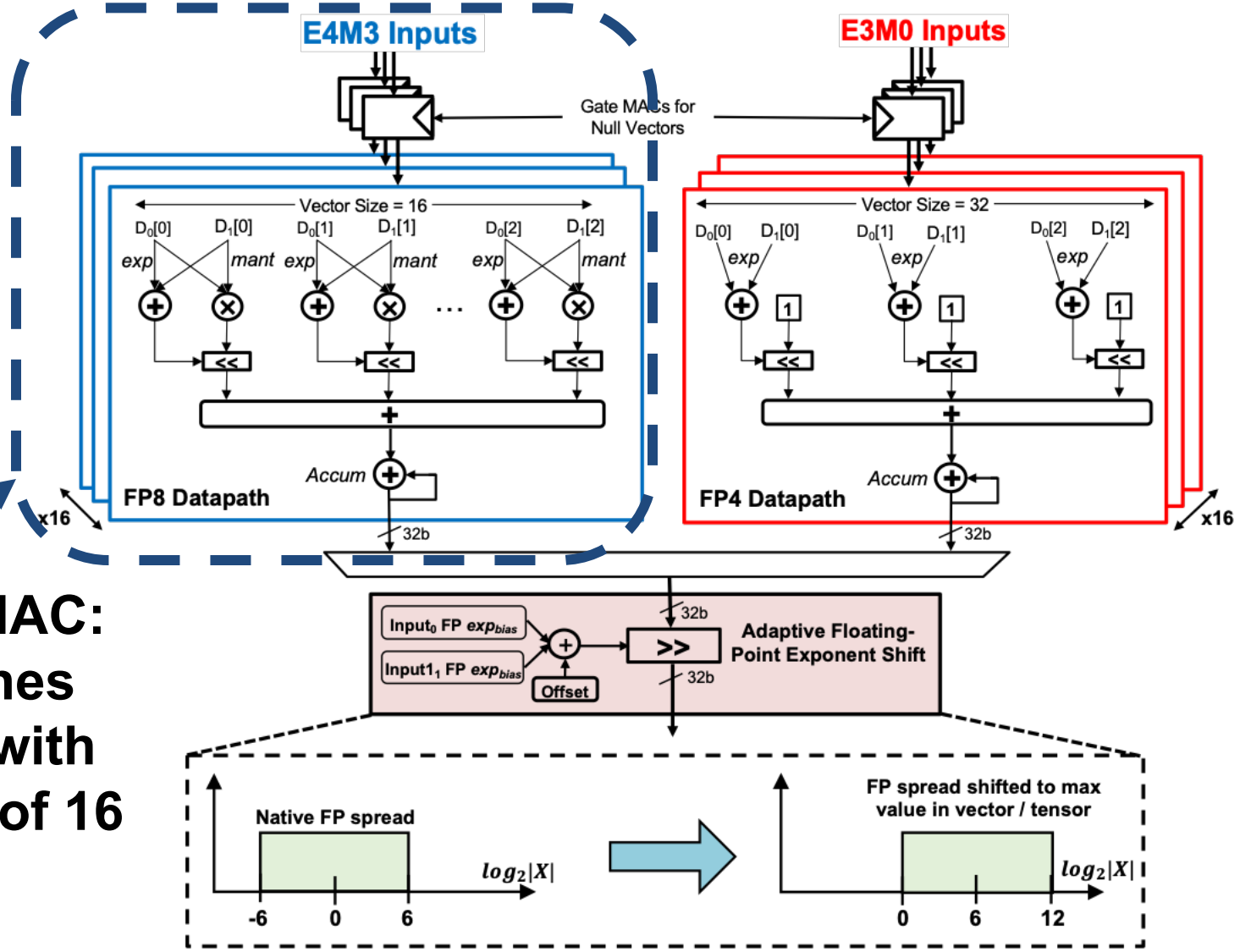
Mixed-Precision MAC Datapath



High Accuracy of Floating-Point Computations

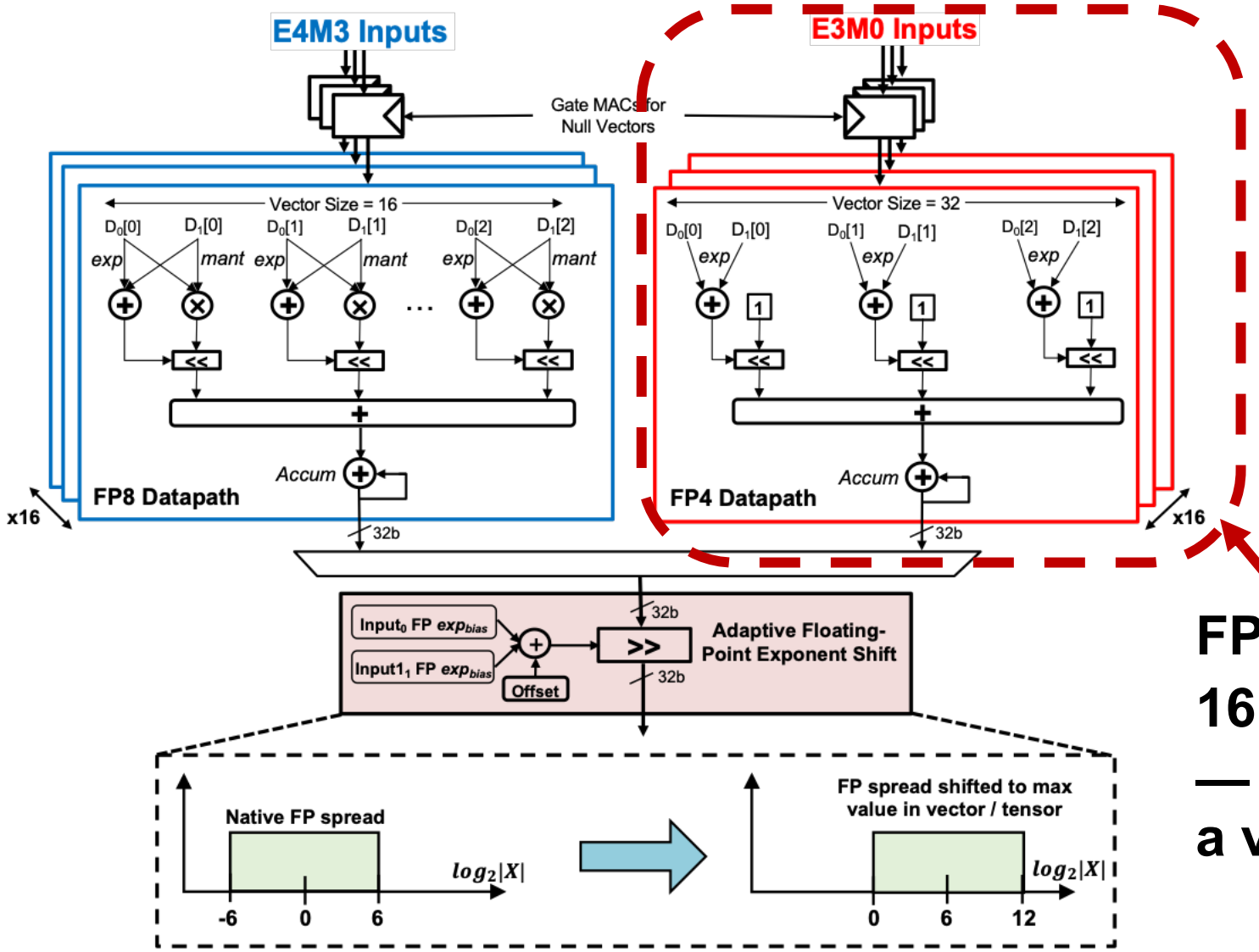
Greater Hardware Density of Fixed-Point Post-Processing

FP8 (E4M3) MAC



FP8 (E4M3) MAC: 16 parallel lanes — each lane with a vector size of 16

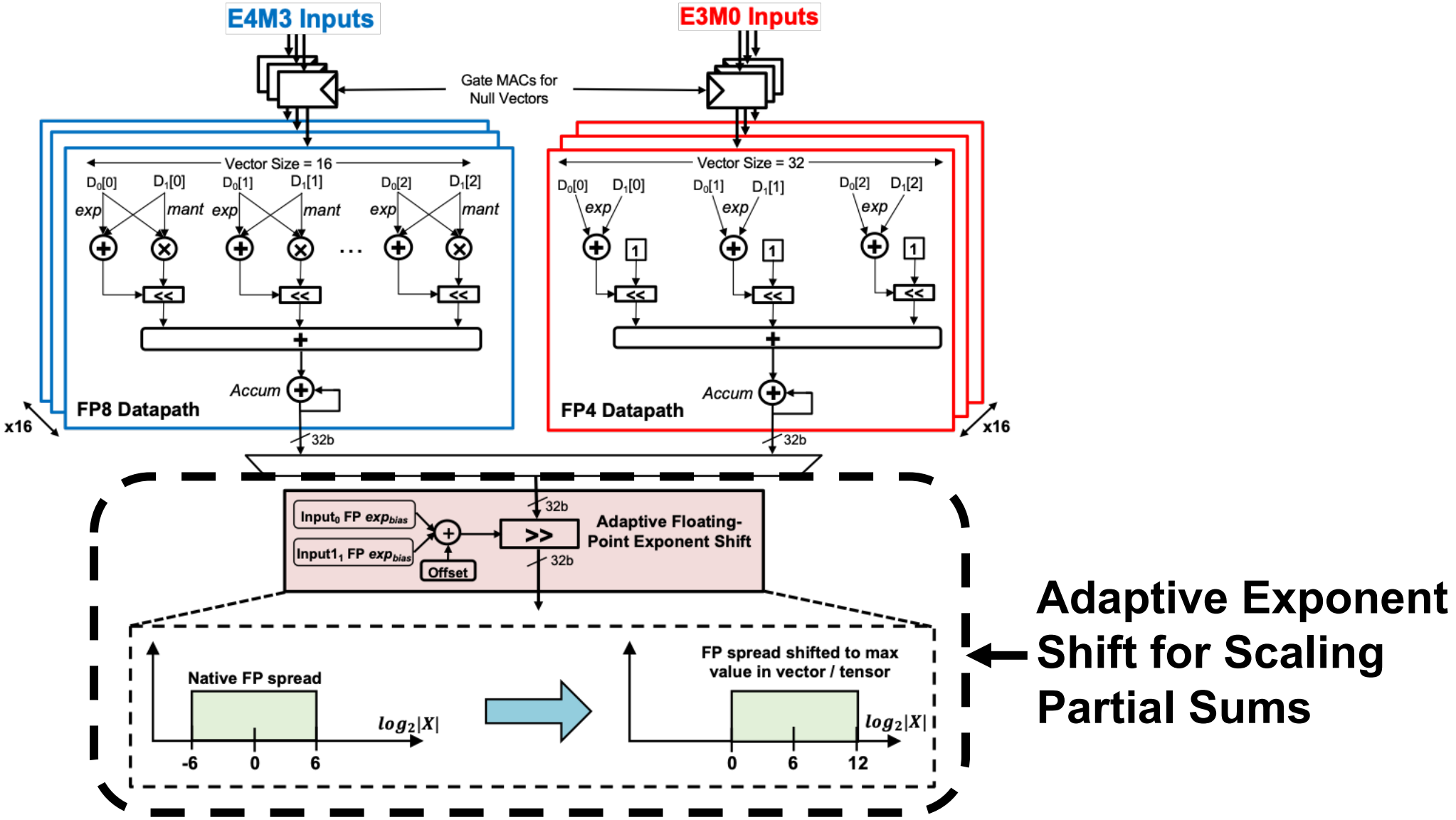
FP4/LOG4 (E3M0) MAC



FP4 (E3M0) MAC: 16 parallel lanes — each lane with a vector size of 32

- 2× throughput

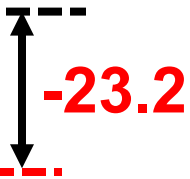
Tensor Scaling Unit



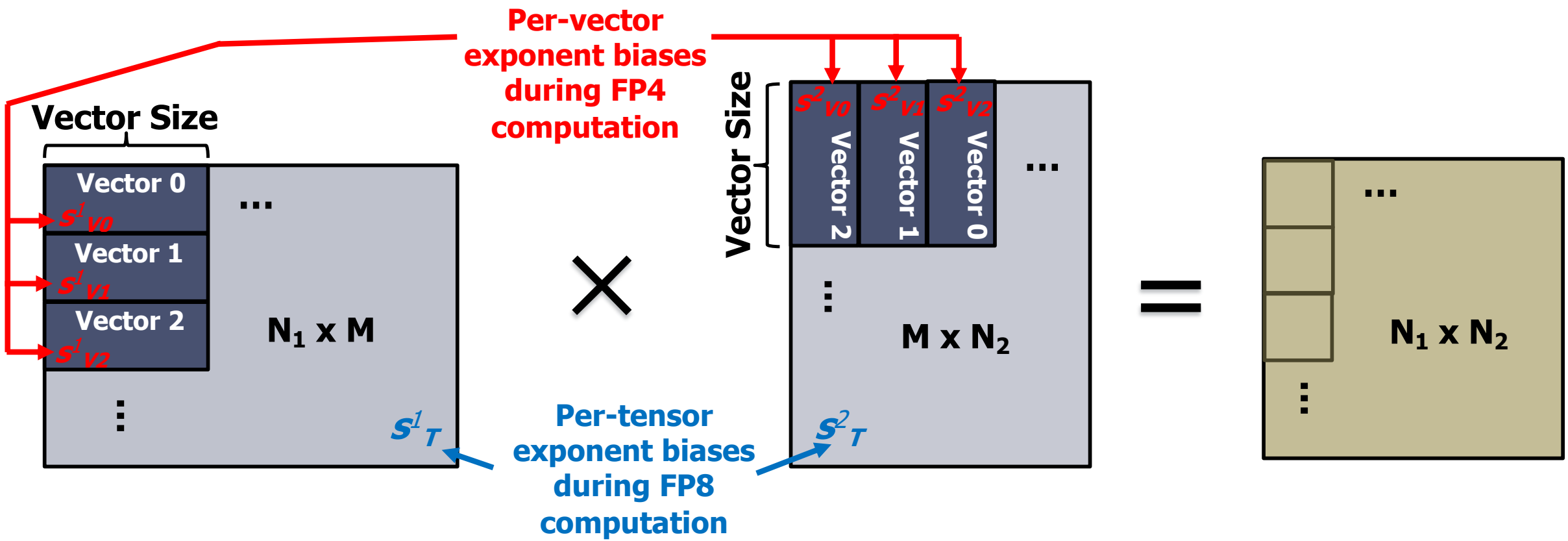
Adaptive Exponent Shift for Scaling Partial Sums

Steep Accuracy Loss w/ Per-Tensor Bias in FP4

Baseline SST-2 Acc.	92.2
w/ FP8 per-tensor exponent bias	92.2
w/ FP4 per-tensor exponent bias	69.0



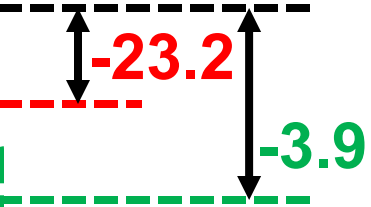
Per-Vector Exponent Scaling when using FP4



- To avoid steep accuracy loss, we adopt per-vector exponent bias scaling in the FP4 regime

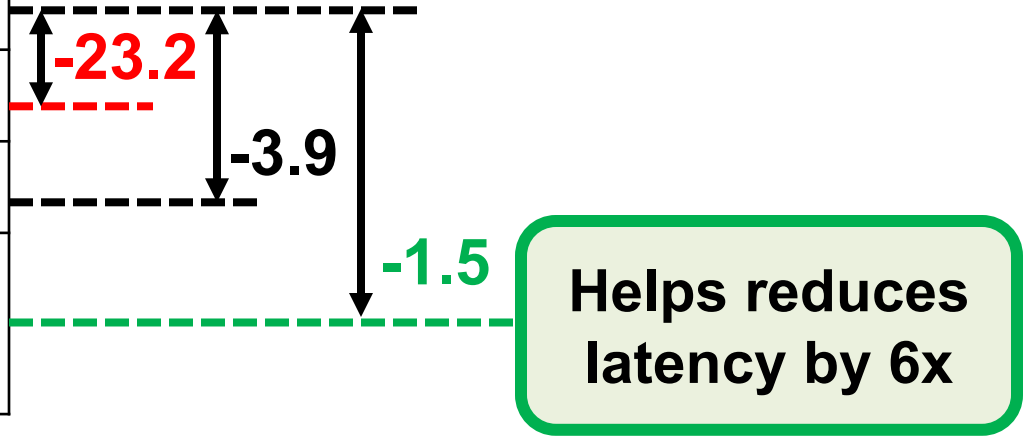
Per-Vector Scaling in FP4 Averts Steep Accuracy Loss

Baseline SST-2 Acc.	92.2
w/ FP8 per-tensor scaling	92.2
w/ FP4 per-tensor scaling	69.0
w/ FP4 per-vector scaling	88.3



Entropy-Controlled Precision Selection

Baseline SST-2 Acc.	92.2
w/ FP8 per-tensor scaling	92.2
w/ FP4 per-tensor scaling	69.0
w/ FP4 per-vector scaling	88.3
w/ entropy-controlled precision selection	90.7

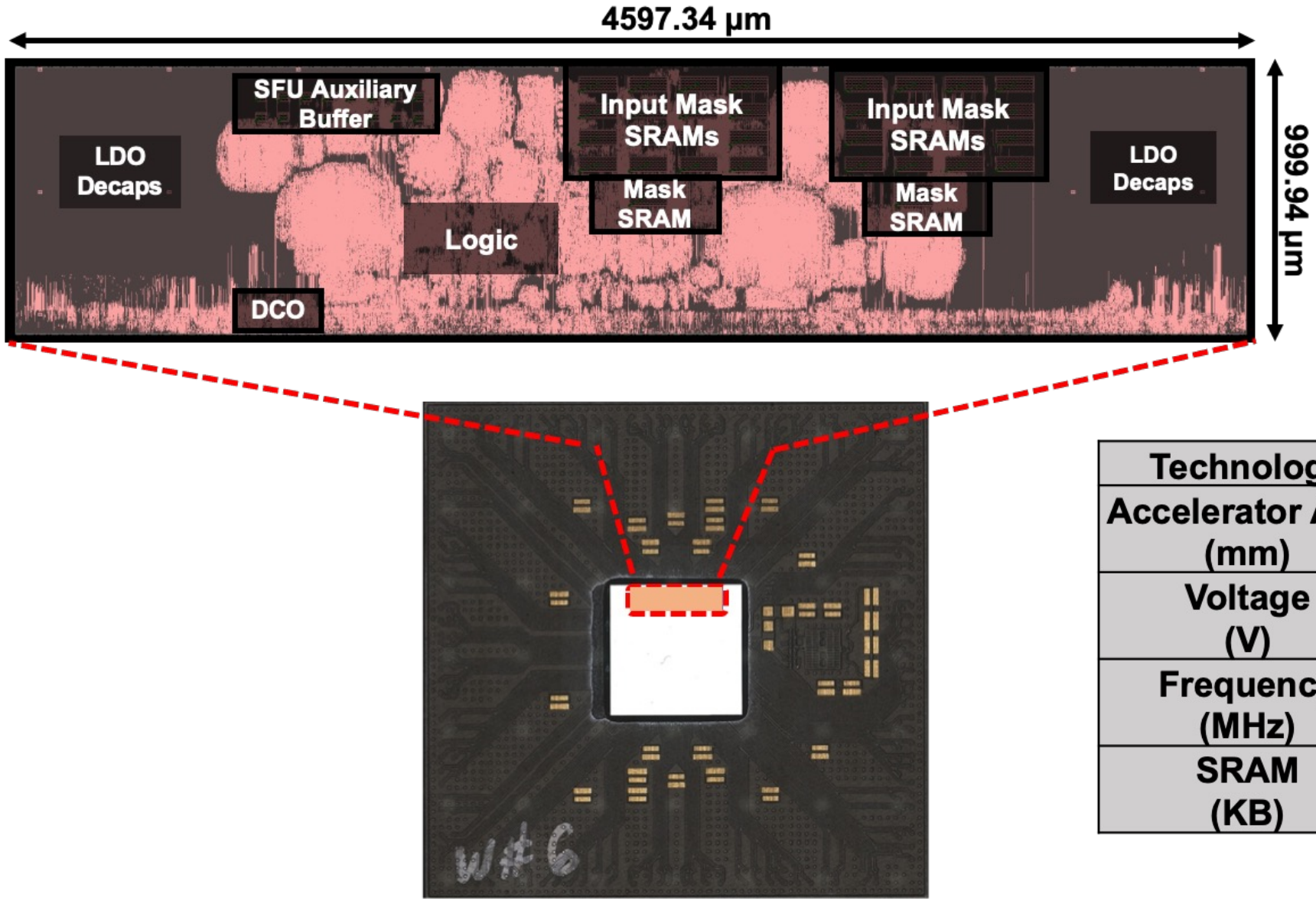


- Pre-calibrated entropy predication selects between FP4 and FP8 MAC during mixed-precision operation

Outline

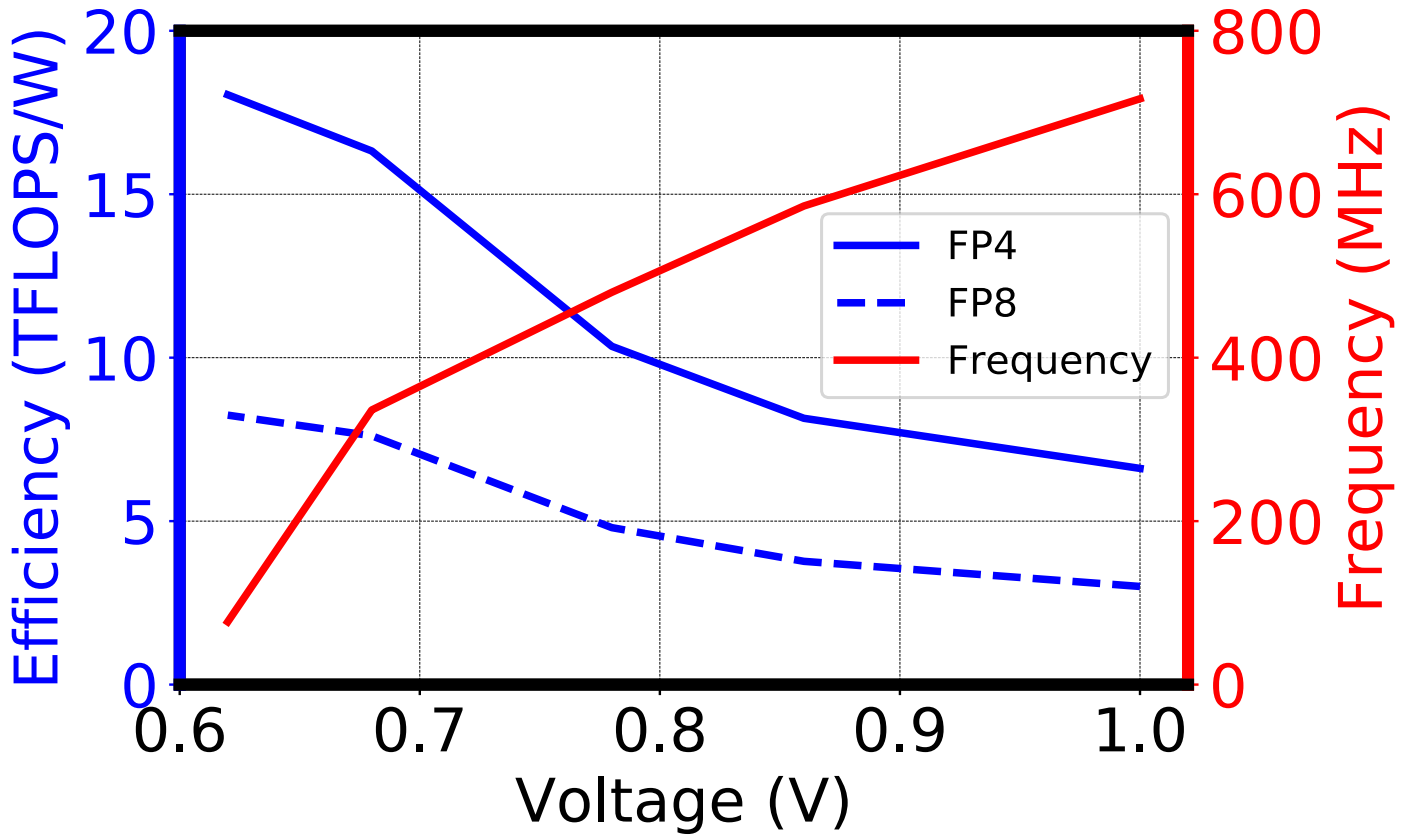
- Motivation
- Entropy-Driven Optimizations
  - Early Exit
  - Latency-Aware Voltage-Frequency Scaling
- 12nm Transformer Accelerator Architecture
  - Mixed-Precision FP4/FP8 Datapath
- **Chip Measurement Results**
- Summary

12nm Chip Tapeout



Technology	GF 12 nm FinFET
Accelerator Area (mm)	4.597 x 0.999
Voltage (V)	0.62 – 1.0
Frequency (MHz)	77 – 717
SRAM (KB)	647

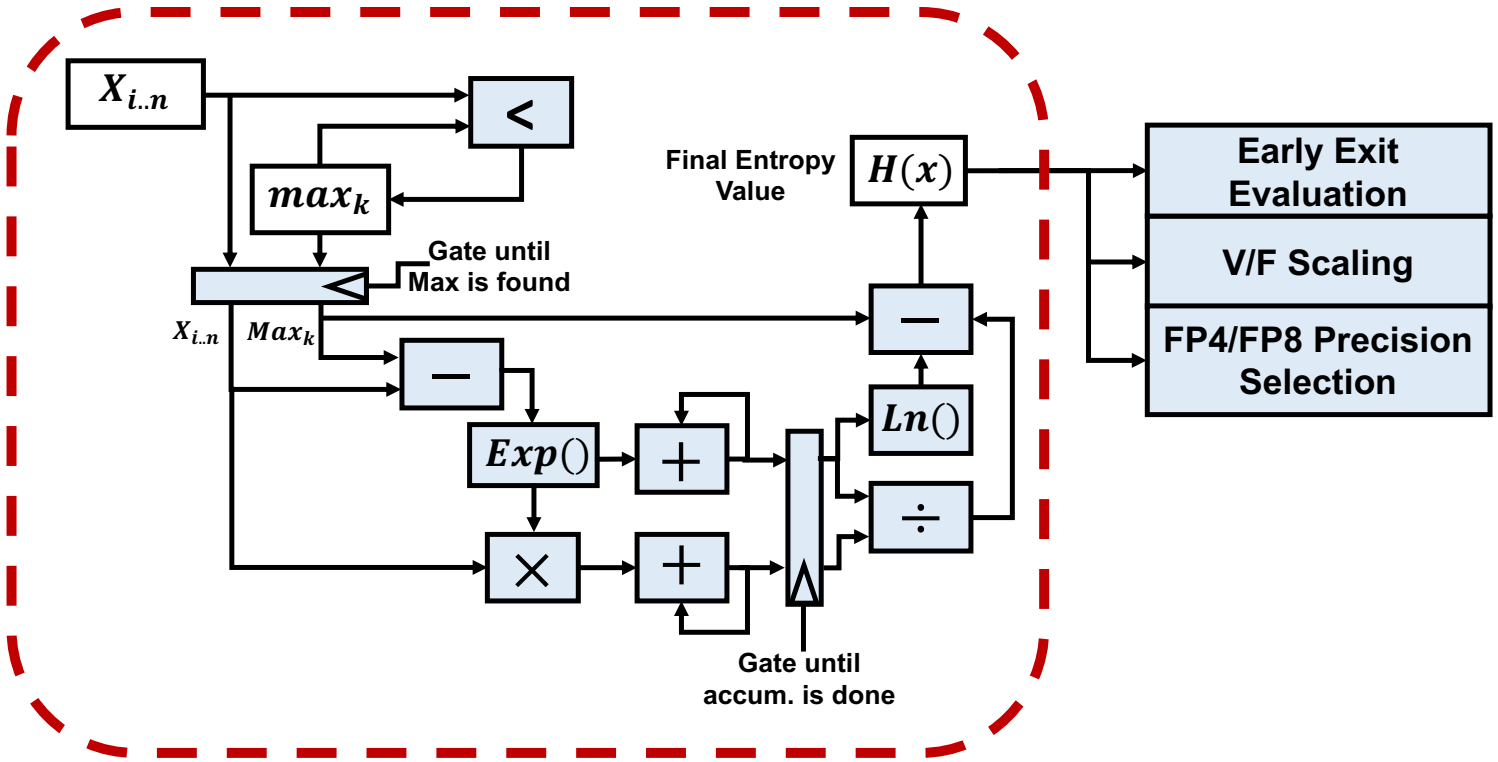
Accelerator Efficiency



Peak energy efficiency:

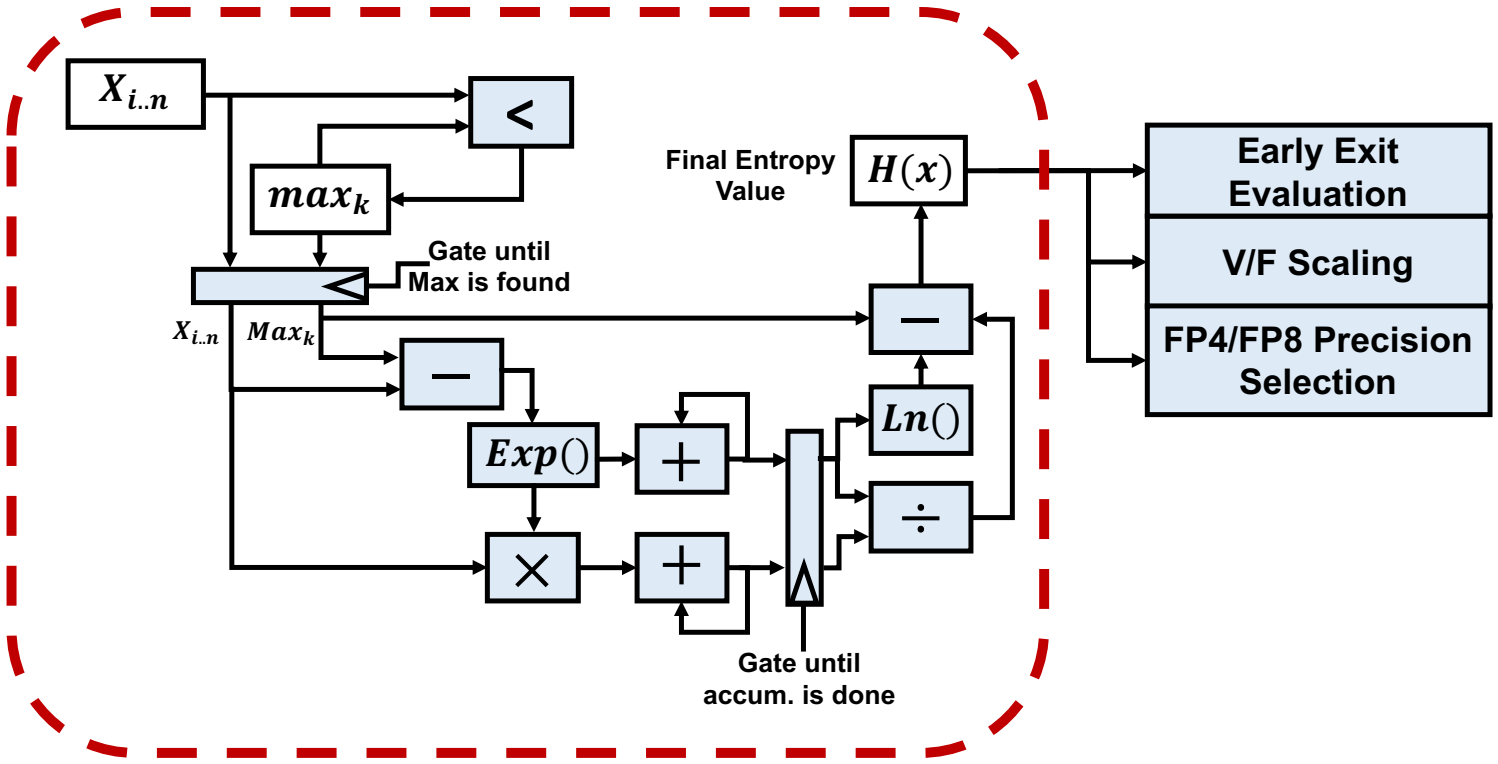
- **FP4: 18.1 TFLOPS/W**
- **FP8: 8.24 TFLOPS/W**

Entropy Hardware Unit

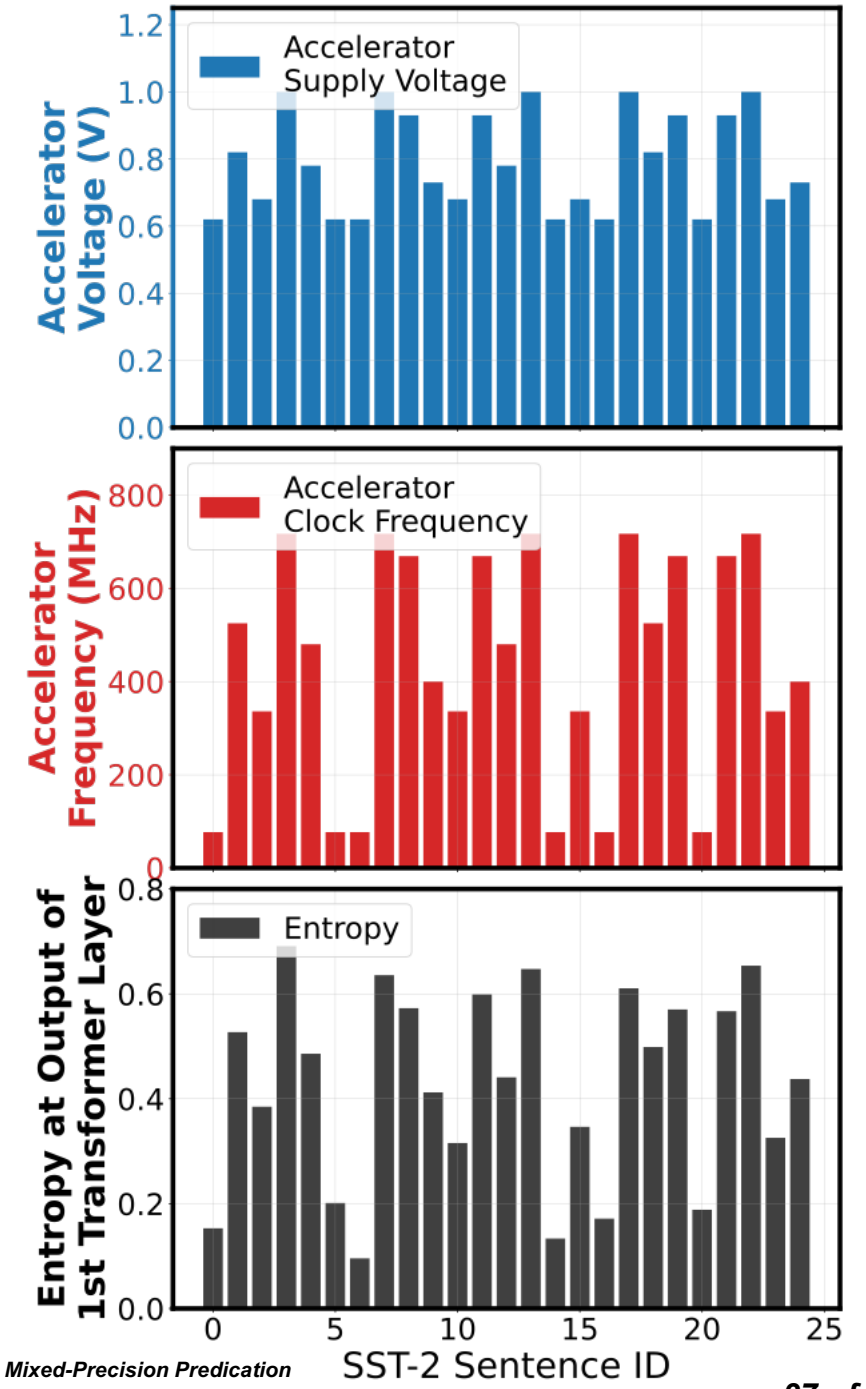


(Reformulated for numerical stability)

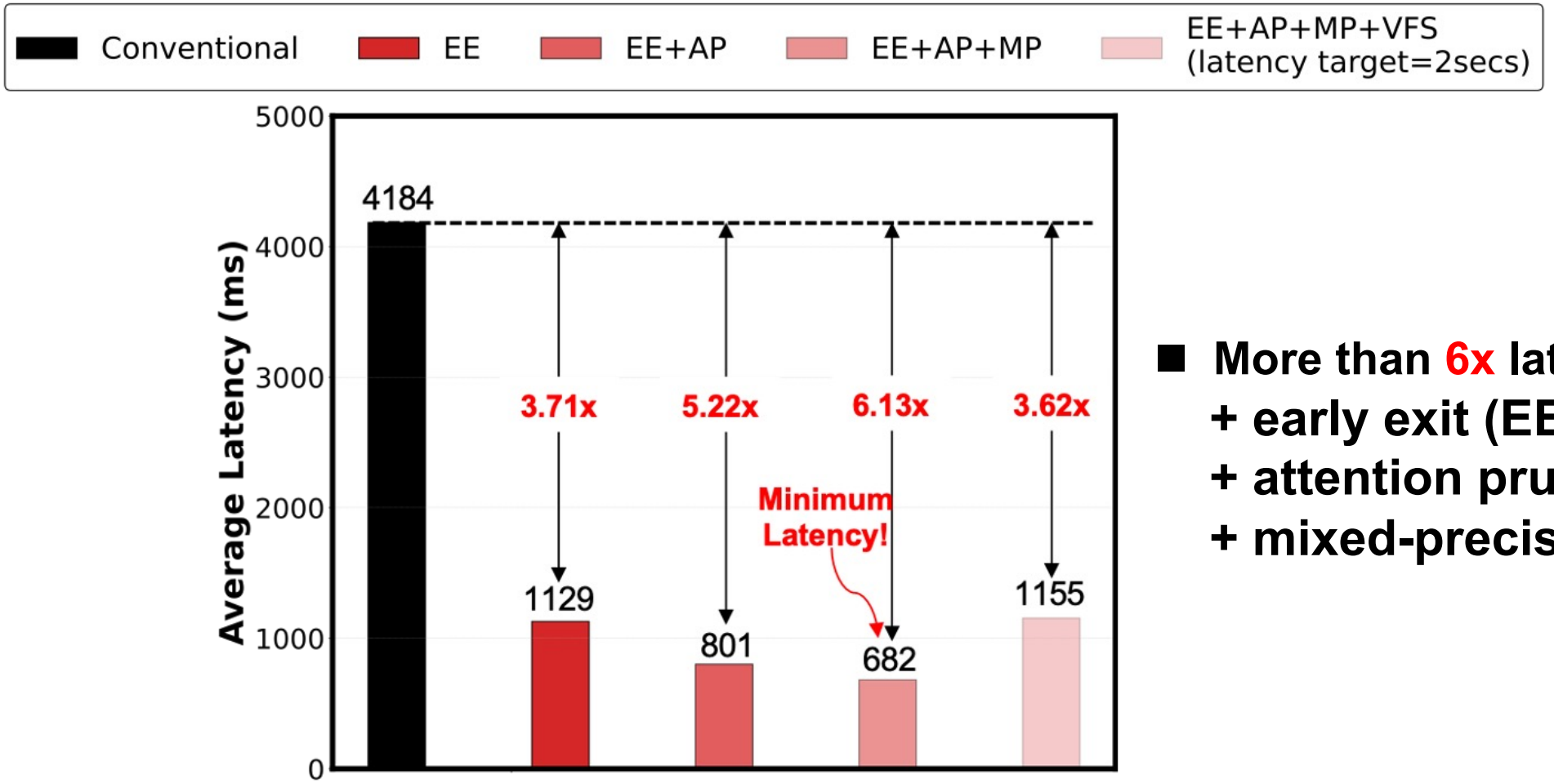
Entropy Hardware Unit



(Reformulated for numerical stability)

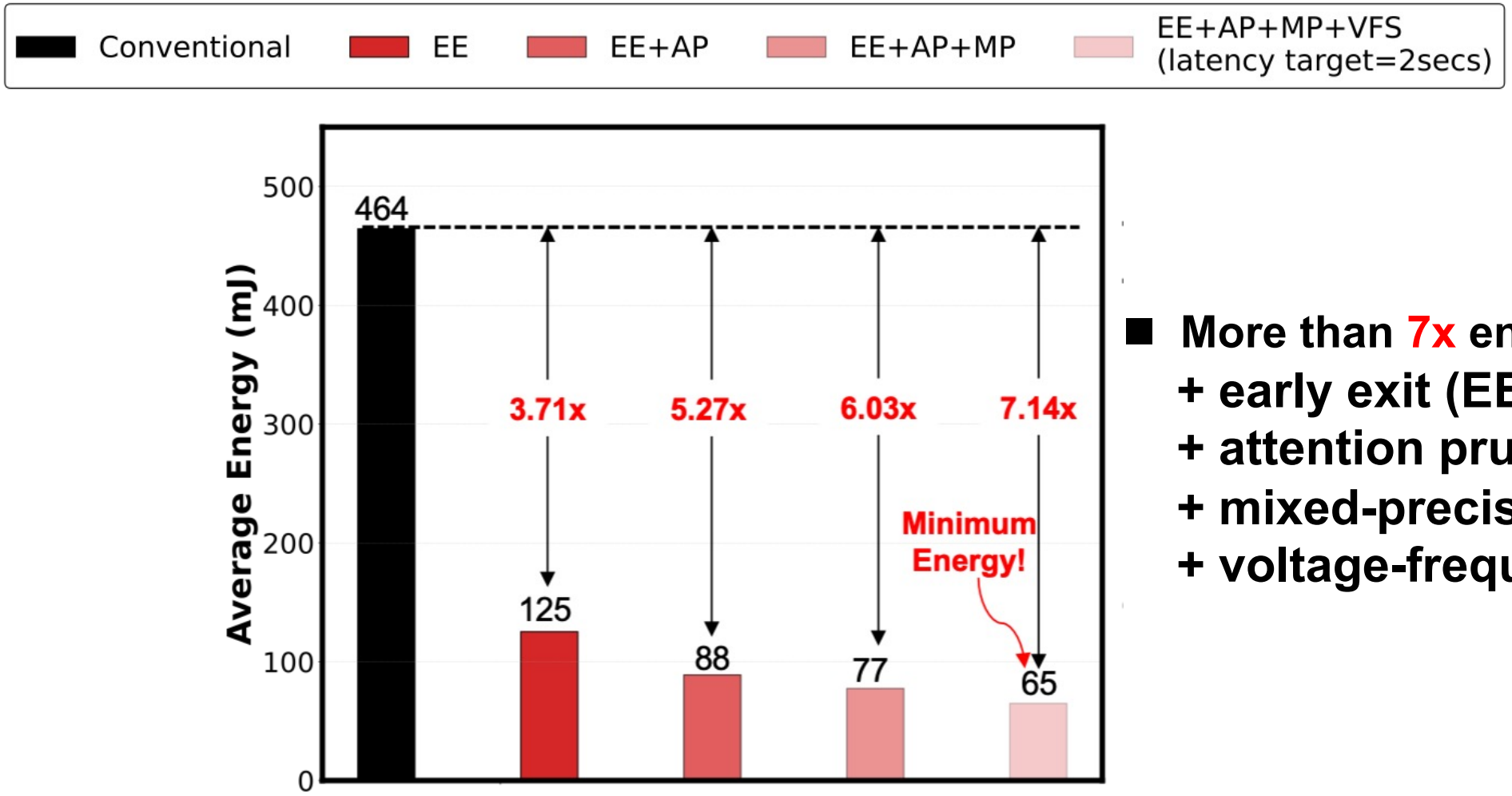


Measured Latency Results



- More than 6x latency reduction with
  + early exit (EE)
  + attention pruning (AP)
  + mixed-precision (MP)

Measured Energy Results



- More than 7x energy reduction with
  + early exit (EE)
  + attention pruning (AP)
  + mixed-precision (MP)
  + voltage-frequency scaling (VFS)

Summary

- **Large language models levy a hefty cost on low capacity edge devices**
- **This work enables fine-grained sentence-level latency and energy optimizations for BERT inference aided by:**
  - Entropy-based early exit
  - Entropy-based voltage/frequency scaling
  - FP4/FP8 mixed-precision MAC
- **Measurements on test chip show:**
  - Up to 6x latency reduction and 7x energy reduction over conventional inference
  - Peak throughput of 18.1TFLOPs/W

This Work is Dedicated to our Friend and Collaborator: Davide Giri

1990 – 2021

Thank You!

Acknowledgements

- This work is supported in part by DARPA, JUMP ADA, NSF Awards 1704834 and 1718160, Intel Corp., and Arm Inc.
- We thank our DARPA collaborators from IBM, Pradip Bose, Martin Cochet, and Karthik Swaminathan for helping support this work.