



# Hardware for Deep Learning

Hot Chips  
August 29, 2023

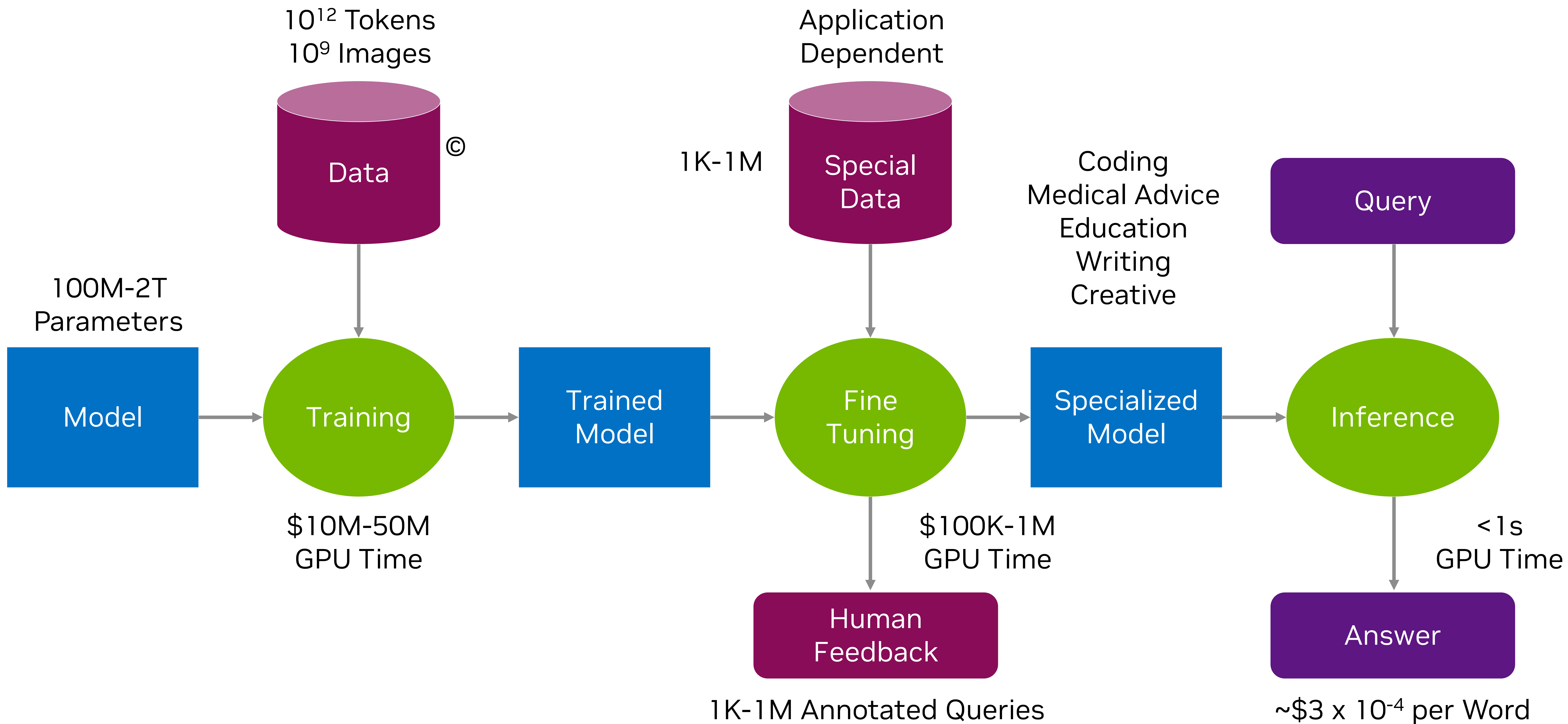
Bill Dally  
Chief Scientist and SVP of Research, NVIDIA Corporation  
Adjunct Professor of CS and EE, Stanford

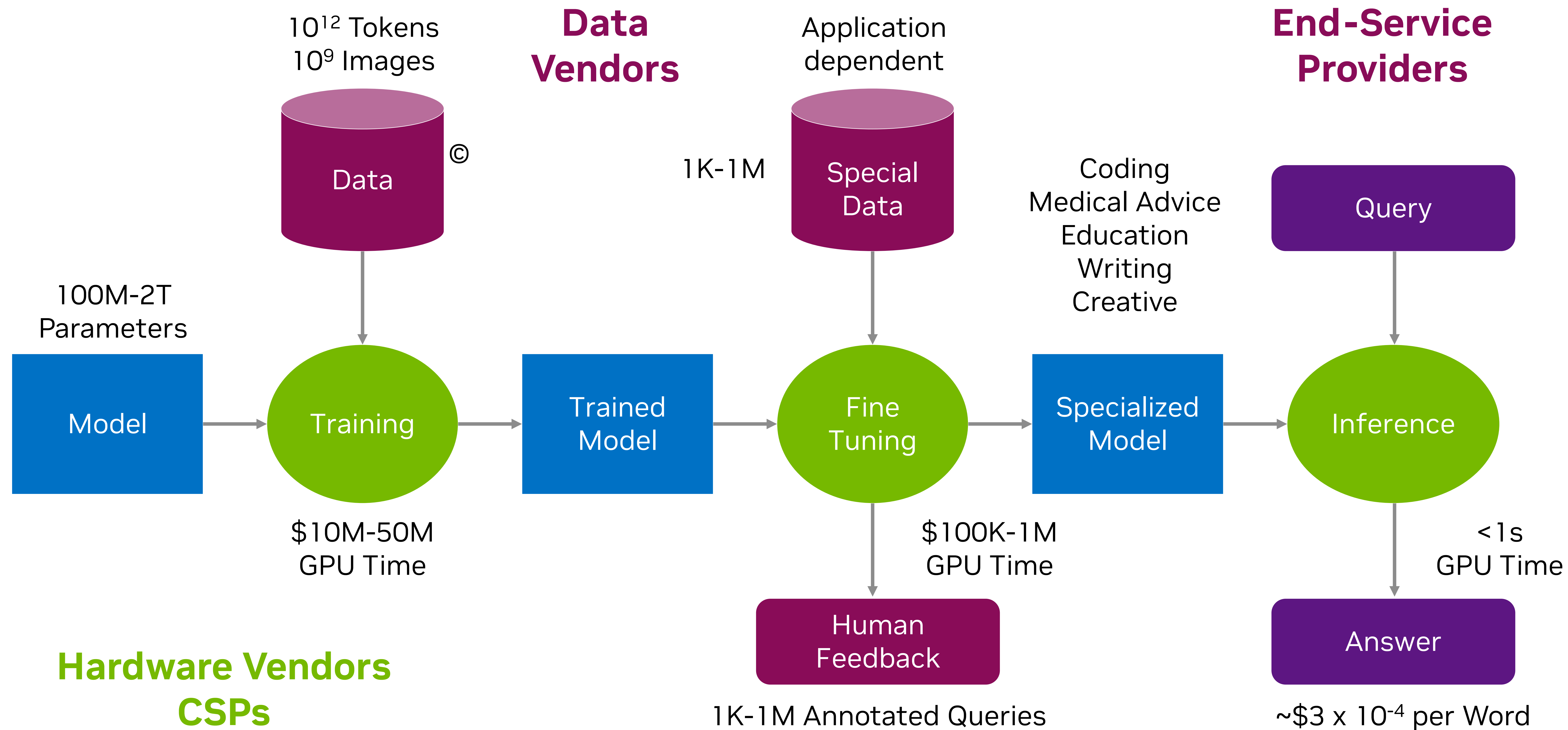
Q: What would Bill Dally say about directions in deep learning hardware

Chat GPT: Bill Dally has discussed several directions in deep learning hardware that he believes are important for the future of the field. Some of these directions include:

1. Specialized hardware for specific deep learning tasks: Dally has argued that as deep learning becomes more complex and diverse, specialized hardware tailored to specific tasks, such as natural language processing or computer vision, will become increasingly important.
2. Energy efficiency: As deep learning systems grow in size and complexity, their energy requirements can become prohibitively high. Dally has emphasized the need for energy-efficient hardware designs that can handle the demands of deep learning workloads.
3. Hardware-software co-design: Dally believes that hardware and software must be designed together to achieve optimal performance. This includes optimizing the hardware architecture for specific deep learning algorithms and developing software tools that can take advantage of the hardware's unique features.
4. Low-precision arithmetic: Dally has proposed using low-precision arithmetic, such as 8-bit or 4-bit arithmetic, in deep learning hardware to reduce energy consumption while still maintaining high accuracy.
5. Memory hierarchy optimization: Dally has also stressed the importance of optimizing the memory hierarchy in deep learning hardware to reduce energy consumption and improve performance.

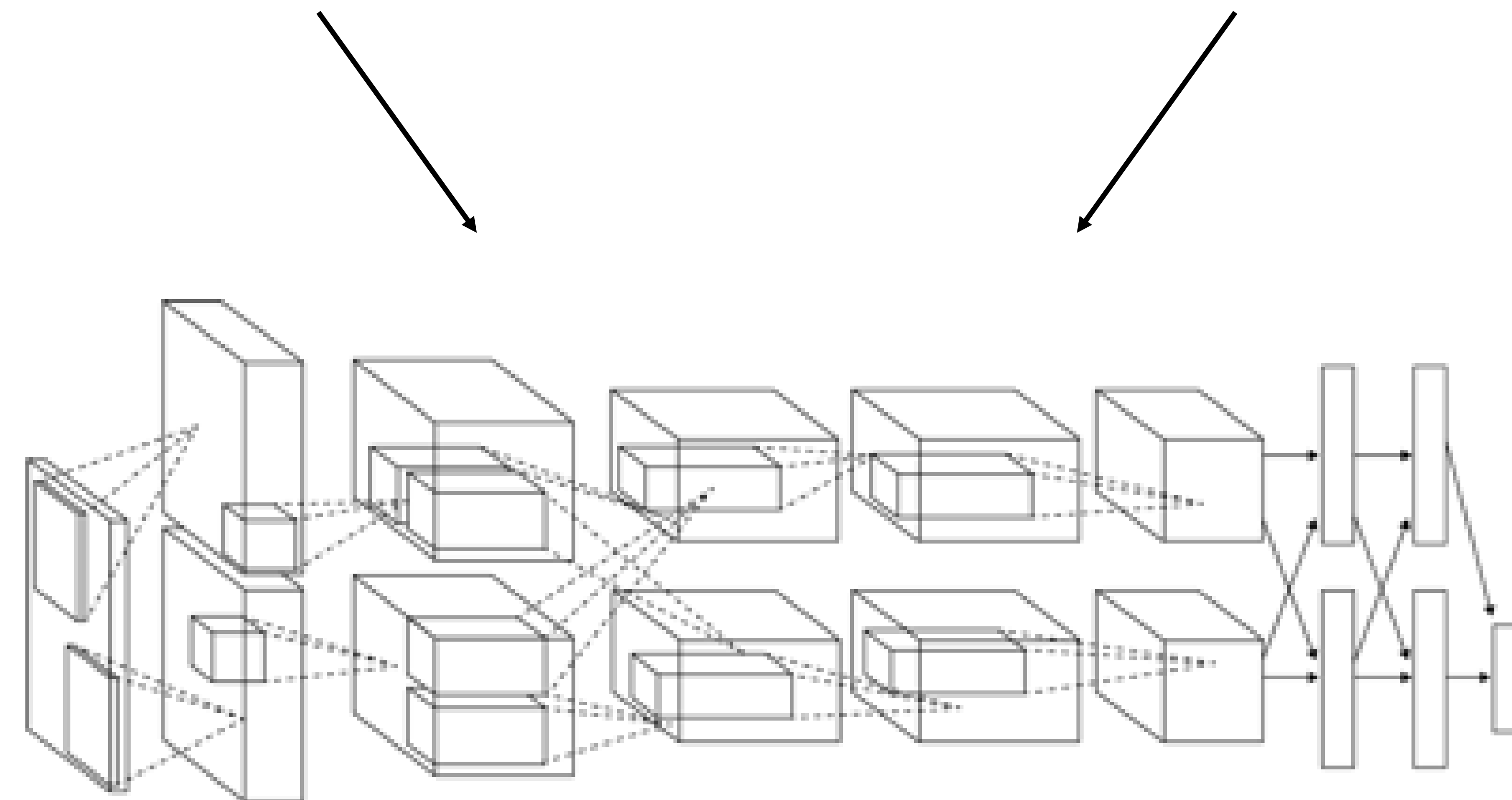
In summary, Bill Dally believes that deep learning hardware must be tailored to the specific needs of different tasks, be energy-efficient, use low-precision arithmetic, optimize the memory hierarchy, and be co-designed with software to achieve optimal performance.





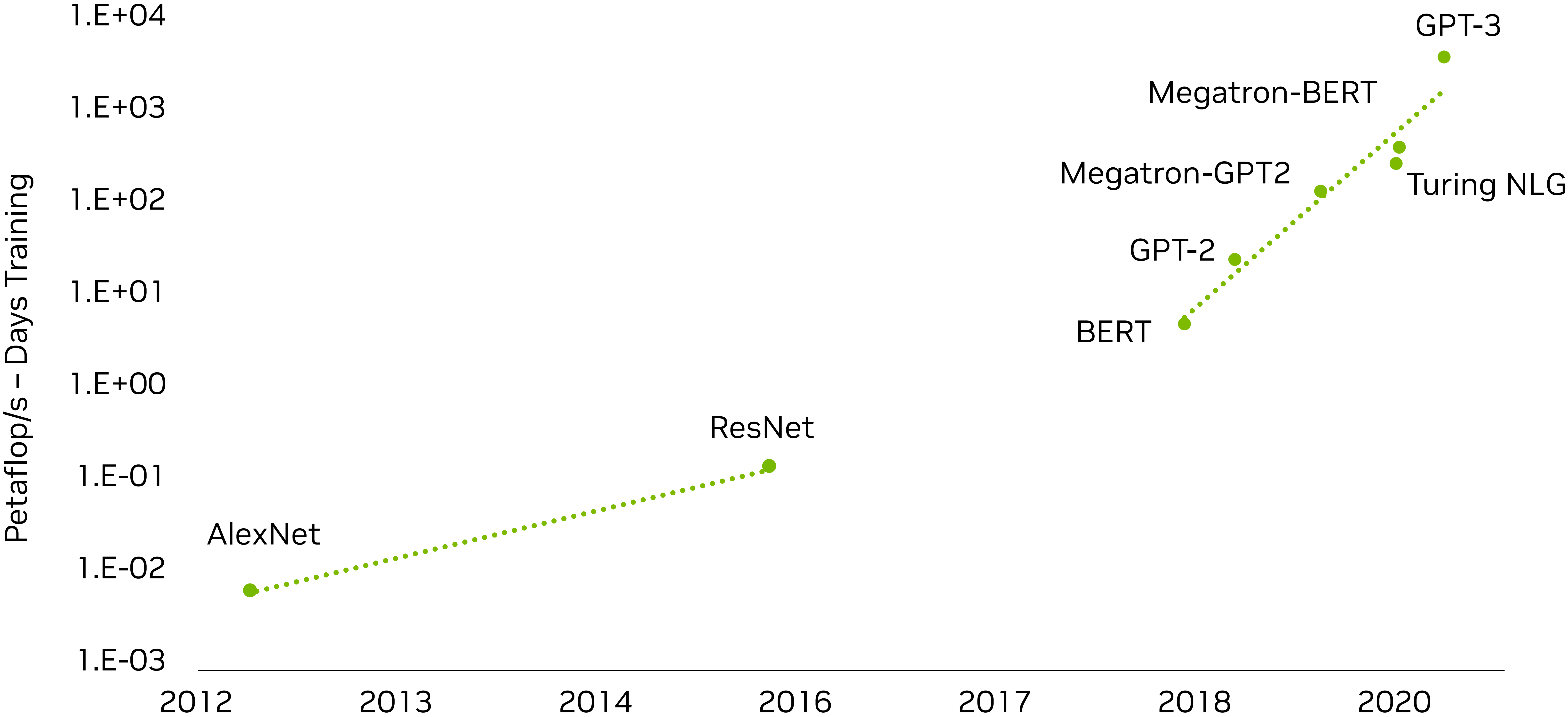
# Motivation

# Deep Learning was Enabled by Hardware



# Deep Learning is Gated by Hardware

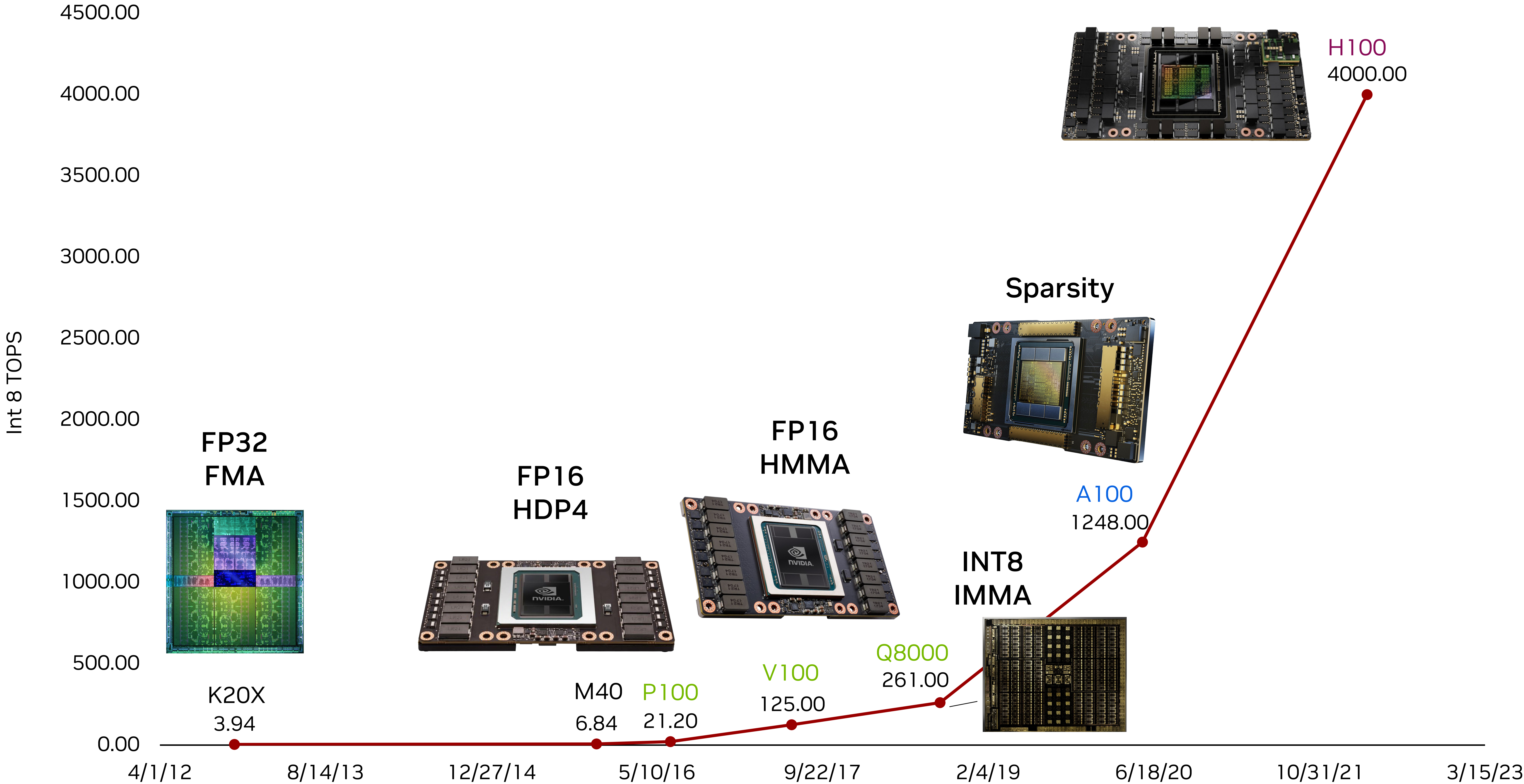
**GPT-4  
est**





# Some History

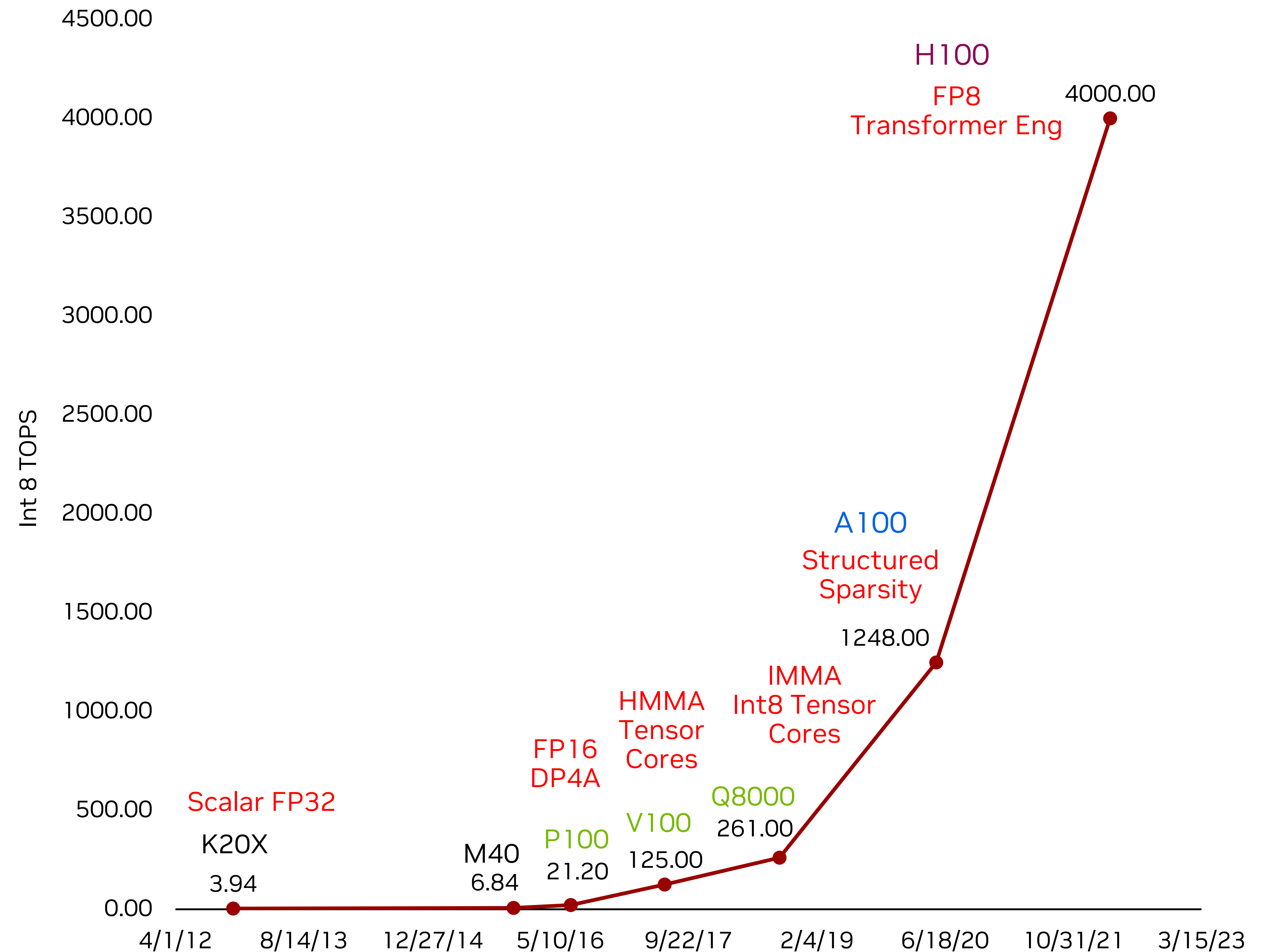
# Single-Chip Inference Performance - 1000X in 10 years



## Gains from

- Number Representation
  - FP32, FP16, Int8
  - (TF32, BF16)
  - ~16x
- Complex Instructions
  - DP4, HMMA, IMMA
  - ~12.5x
- Process
  - 28nm, 16nm, 7nm, 5nm
  - ~2.5x
- Sparsity
  - ~2x
- Model efficiency has also improved – overall gain > 1000x

### Single-Chip Inference Performance - 1000X in 10 years



# Specialized Instructions Amortize Overhead

Operation	Energy**	Overhead*
HFMA	1.5pJ	2000%
HDP4A	6.0pJ	500%
HMMA	110pJ	22%
IMMA	160pJ	16%

\*Overhead is instruction fetch, decode, and operand fetch – 30pJ

\*\*Energy numbers from 45nm process

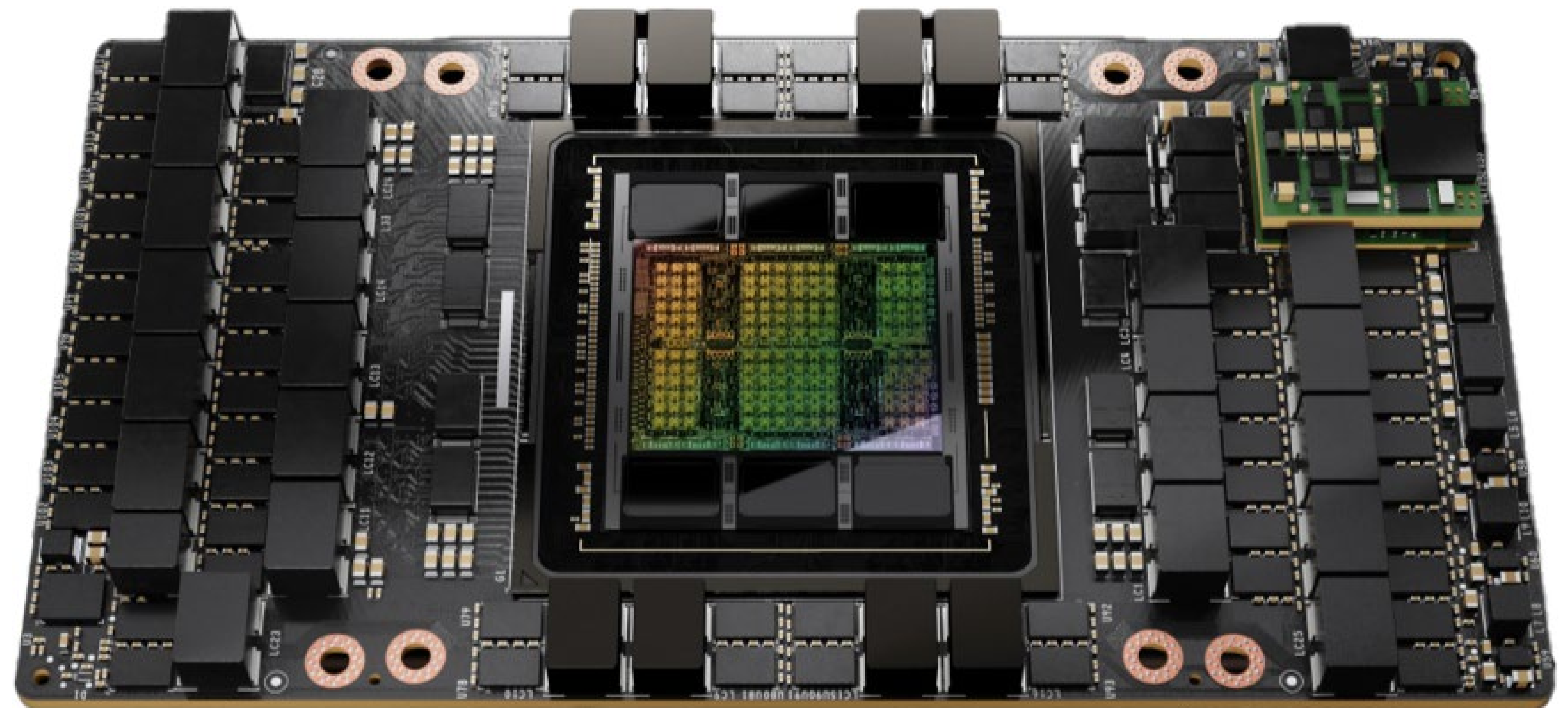
# Hopper H100

1 PFLOPS (TF32)  
1 / 2 PLFLOPS (FP16 or BF16) (dense/sparse)  
2 / 4 PLFLOPS (FP8 or Int8) (dense/sparse)

3.4TB/s (HBM3) 94GB  
18 NVLINK ports  
400Gb/s each 900GB/s total  
700W

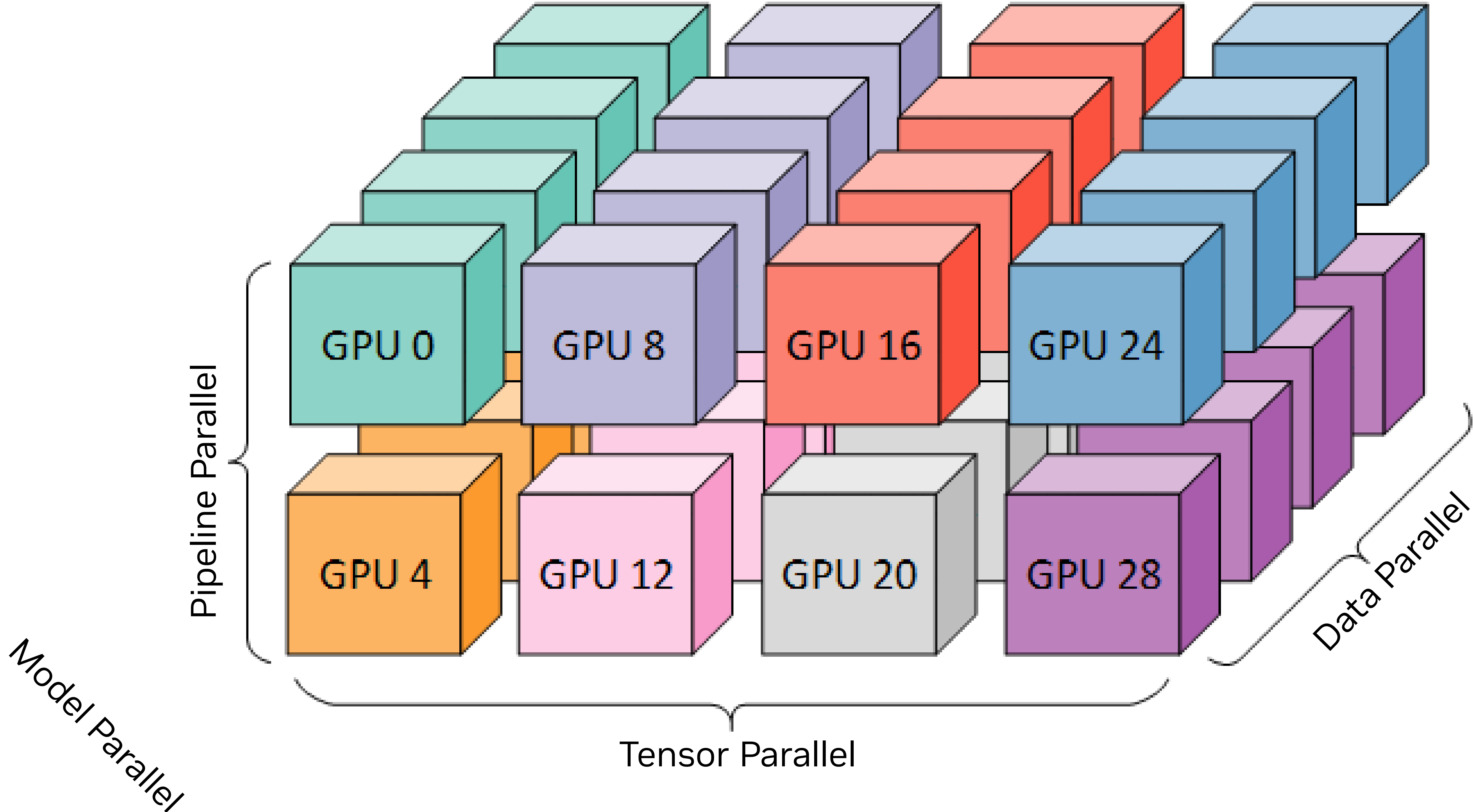
Transformer Engine  
Dynamic Programming Instructions

9 TOPS/W (Int8/FP8)



4PF Sparse FP8, 900GB/s, 700W

# 3D Parallelism



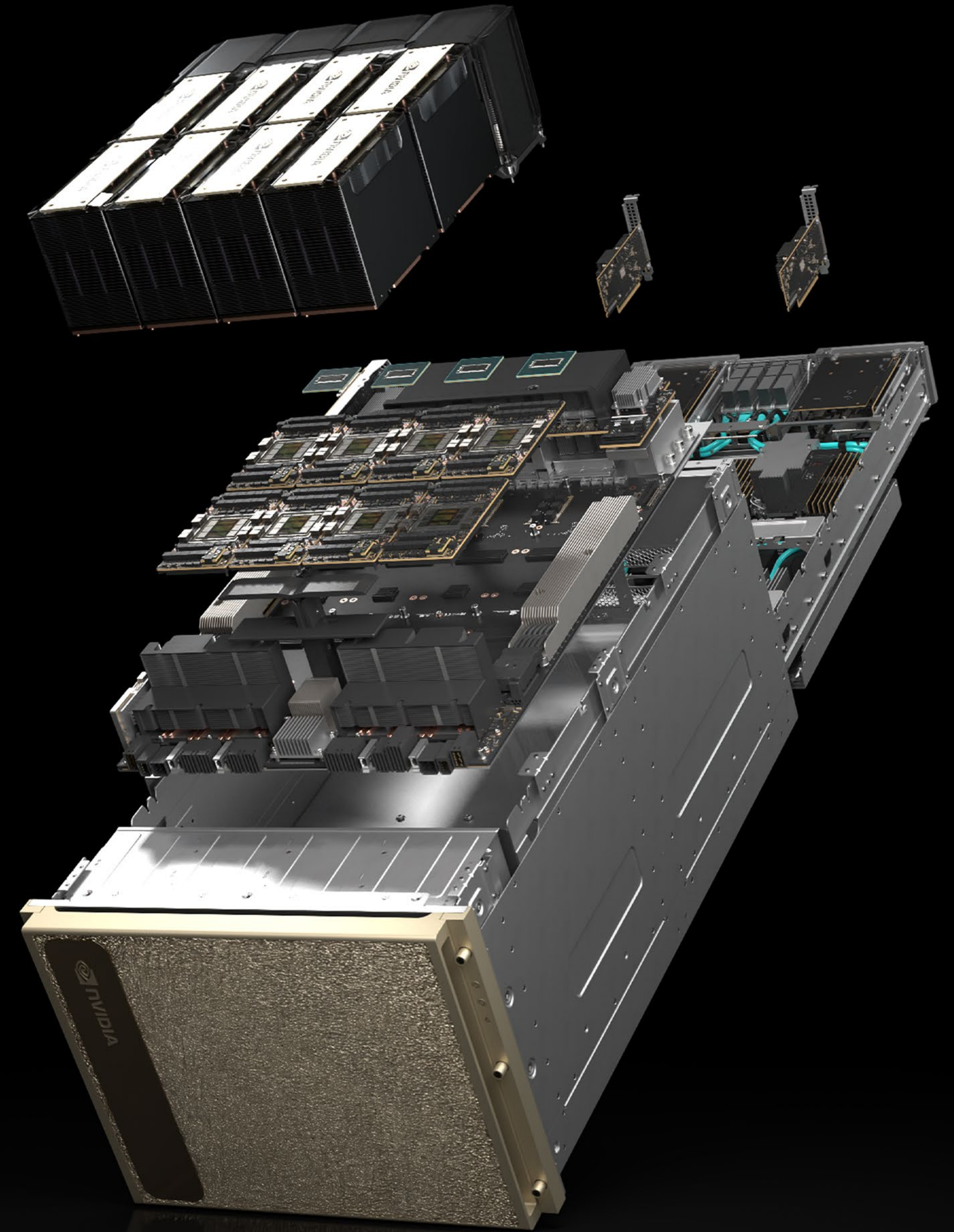
It takes 20 GPUs to hold one copy of GPT4 model parameters

# DGX H100 Server

## 8-H100 4-NVSwitch Server

- 32 PFLOPS of AI Performance
- 640 GB aggregate GPU memory
- 18 NVLink Network OSFPs
- 3.6 TBps of full-duplex NVLink Network bandwidth (72 NVLinks)
- 8x 400 Gb/s ConnectX-7 InfiniBand/Ethernet ports
- 2 dual-port Bluefield-3 DPUs
- Dual Sapphire Rapids CPUs
- PCIe Gen5

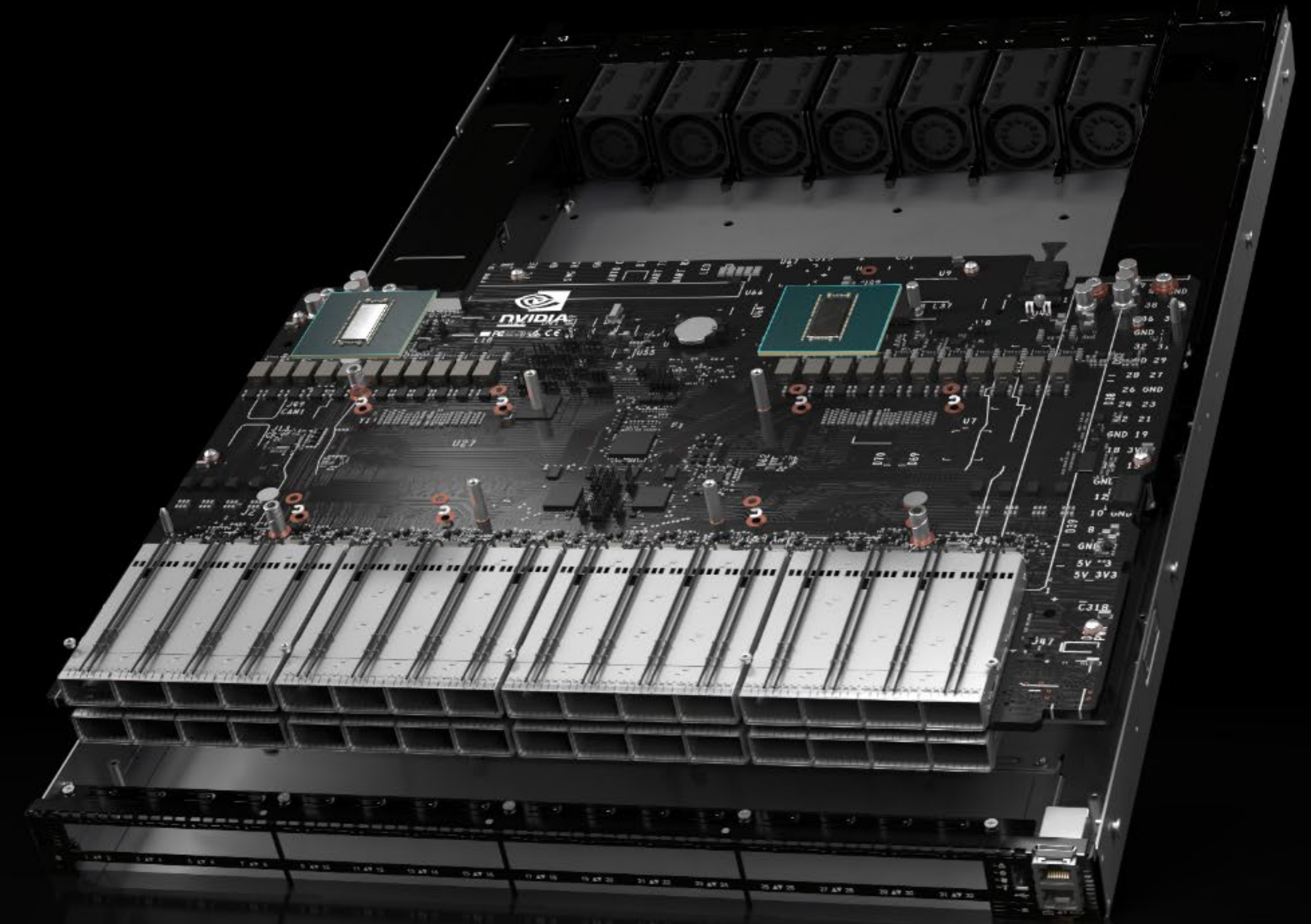
32PF Sparse FP8, 11.3kW, 900GB/s  
One Big GPU



# DGX H100 Superpod: NVLink Switch

## NVLink Switch

- Standard 1RU 19-inch formfactor highly leveraged from InfiniBand switch design
- Dual NVLink4 NVSwitch chips
- 128 NVLink4 ports
- 32 OSFP cages
- 6.4 TB/s full-duplex BW
- Managed switch with out-of-band management communication
- Support for passive-copper, active-copper and optical OSFP cables (custom FW)







Scale-up – NVLink and NVSwitch – to 256 GPUs

Scale-out – IB to 10,000s of GPUs

Collectives Double Effective Network Bandwidth (AllReduce)

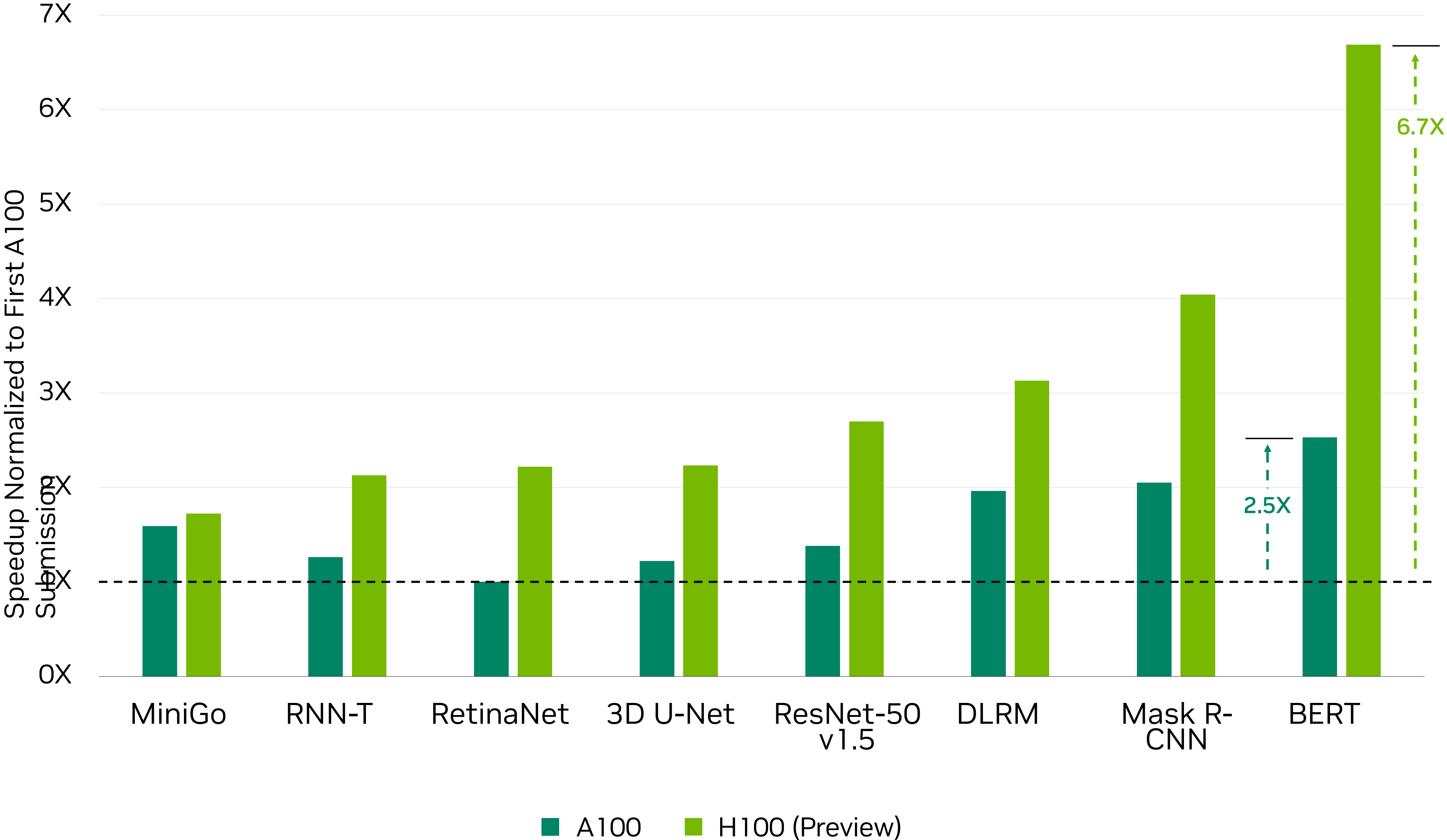
**Software**



# NVIDIA AI and H100 Deliver 6.7X in 2.5 Years

Full-stack innovation fuels continuous performance gains

MLPerf™ Training v2.1 Performance



Up to

# 6.7X

Higher performance with new H100 GPUs

Up to

# 2.5X

Speedup on existing A100 GPUs with software

ResNet-50 v1.5: 8x NVIDIA 0.7-18, 8x NVIDIA 2.1-2060, 8x NVIDIA 2.1-2091 | BERT: 8x NVIDIA 0.7-19, 8x NVIDIA 2.1-2062, 8x NVIDIA 2.1-2091 | DLRM: 8x NVIDIA 0.7-17, 8x NVIDIA 2.1-2059, 8x NVIDIA 2.1-2091 | Mask R-CNN: 8x NVIDIA 0.7-19, 8x NVIDIA 2.1-2062, 8x NVIDIA 2.1-2091 | RetinaNet: 8x NVIDIA 2.0-2091, 8x NVIDIA 2.1-2061, 8x NVIDIA 2.1-2091 | RNN-T: 8x NVIDIA 1.0-1060, 8x NVIDIA 2.1-2061, 8x NVIDIA 2.1-2091 | Mini Go: 8x NVIDIA 0.7-20, 8x NVIDIA 2.1-2063, 8x NVIDIA 2.1-2091 | 3D U-Net: 8x NVIDIA 1.0-1059, 8x NVIDIA 2.1-2060, 8x NVIDIA 2.1-2091  
 First NVIDIA A100 Tensor Core GPU results normalized for throughput due to higher accuracy requirements introduced in MLPerf™ Training 2.0 where applicable. MLPerf™ name and logo are trademarks. See [www.mlperf.org](http://www.mlperf.org) for more information.



- Home
- Topics
- Sectors
- Exascale
- Specials
- Resource Library

## NVIDIA H100 GPUs Set Standard for Generative AI in Debut MLPerf Benchmark

June 28, 2023

June 28, 2023 — Leading users and industry-standard benchmarks agree: **NVIDIA H100 Tensor Core GPUs deliver the best AI performance, especially on the large language models (LLMs) powering generative AI.**

H100 GPUs set new records on all eight tests in the latest MLPerf training benchmarks released this week, excelling on a new MLPerf test for generative AI. **That excellence is delivered both per-accelerator and at-scale in massive servers.**



- Home
- Topics
- Sectors

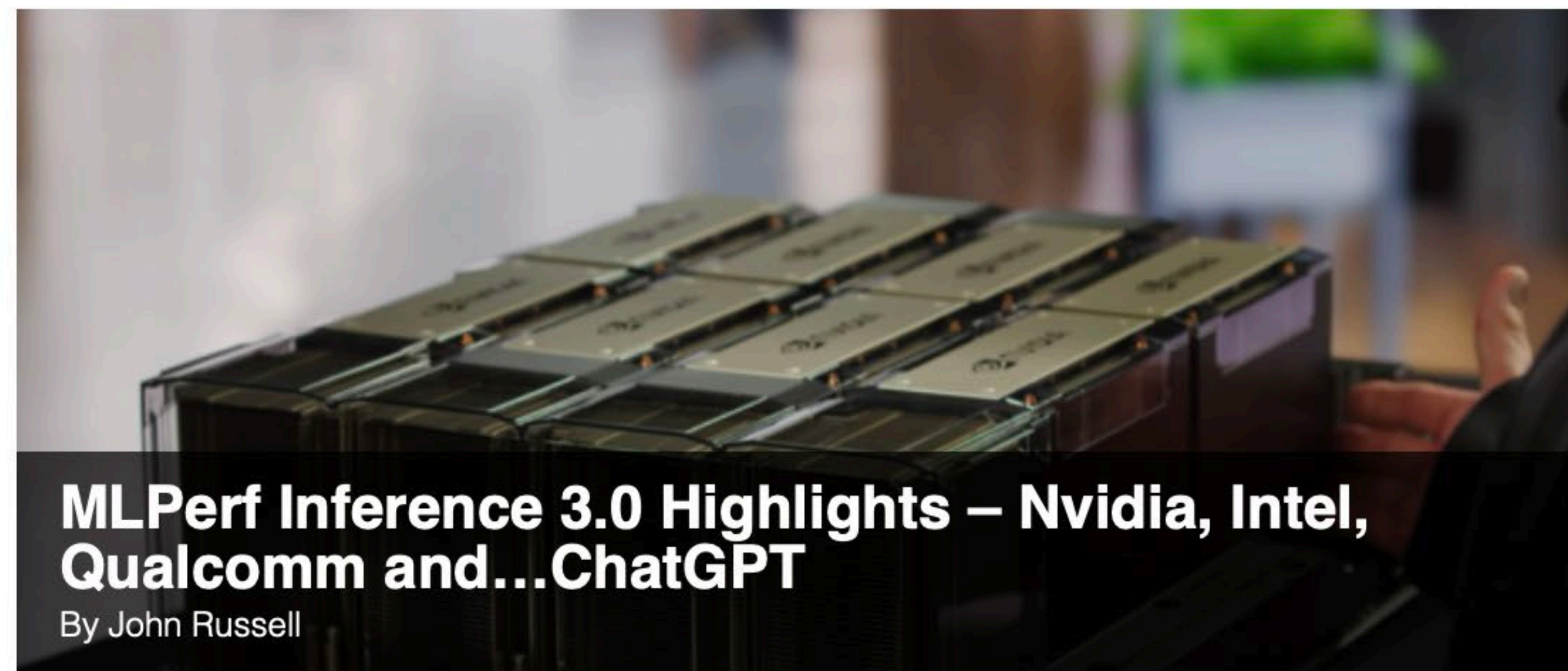
## Nvidia Hopper, Ampere GPUs Sweep MLPerf Benchmarks in AI Training

November 9, 2022

Nov. 9, 2022 — Two months after [their debut](#) sweeping MLPerf inference benchmarks, **NVIDIA H100 Tensor Core GPUs** set world records across enterprise AI workloads in the industry group's latest tests of AI training.

Together, the results show H100 is the best choice for users who demand utmost performance when creating and deploying advanced AI models.

- Home
- Topics
- Sectors
- Exascale
- Specials
- Resource Library
- Podcast**



April 5, 2023

MLCommons today released the latest MLPerf Inferencing (v3.0) results for the datacenter and edge. While **Nvidia continues to dominate the results** – topping all performance categories – other companies are joining the MLPerf

- Home
- Technologies
- Sectors
- AI/ML/DL
- Exascale



April 6, 2022

# Future Directions

# Future Directions

## Number representation

- Log numbers
- Vector scaling (VS-Quant)
- Optimal Clipping
- Much cheaper math
- Smaller numbers

## Sparsity

- Activations
- Lower density (vs 2:4 in A100/H100)

## Better tiling

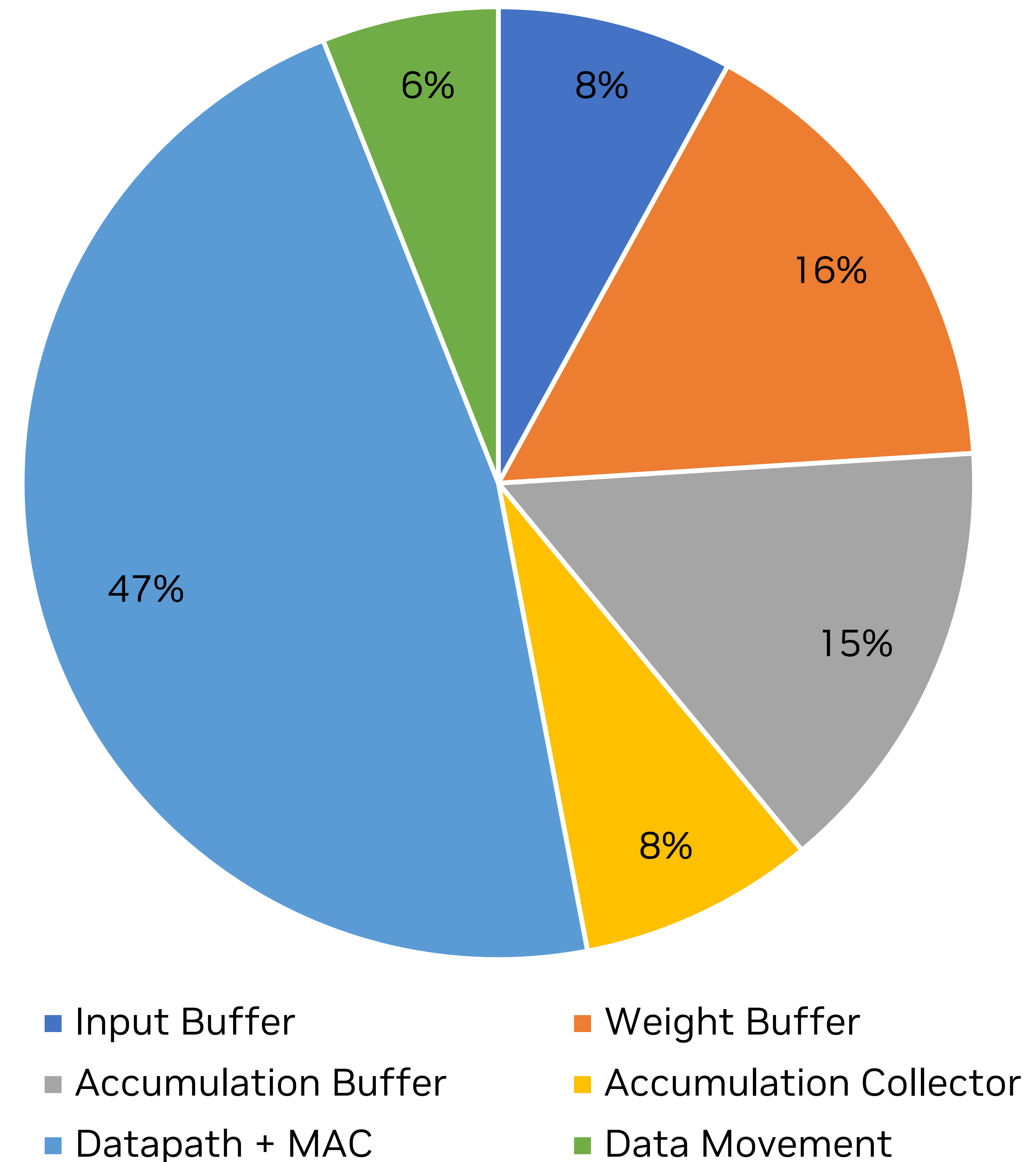
- Lower memory energy

## Circuits

- Memory
- Communication
- 3D memory

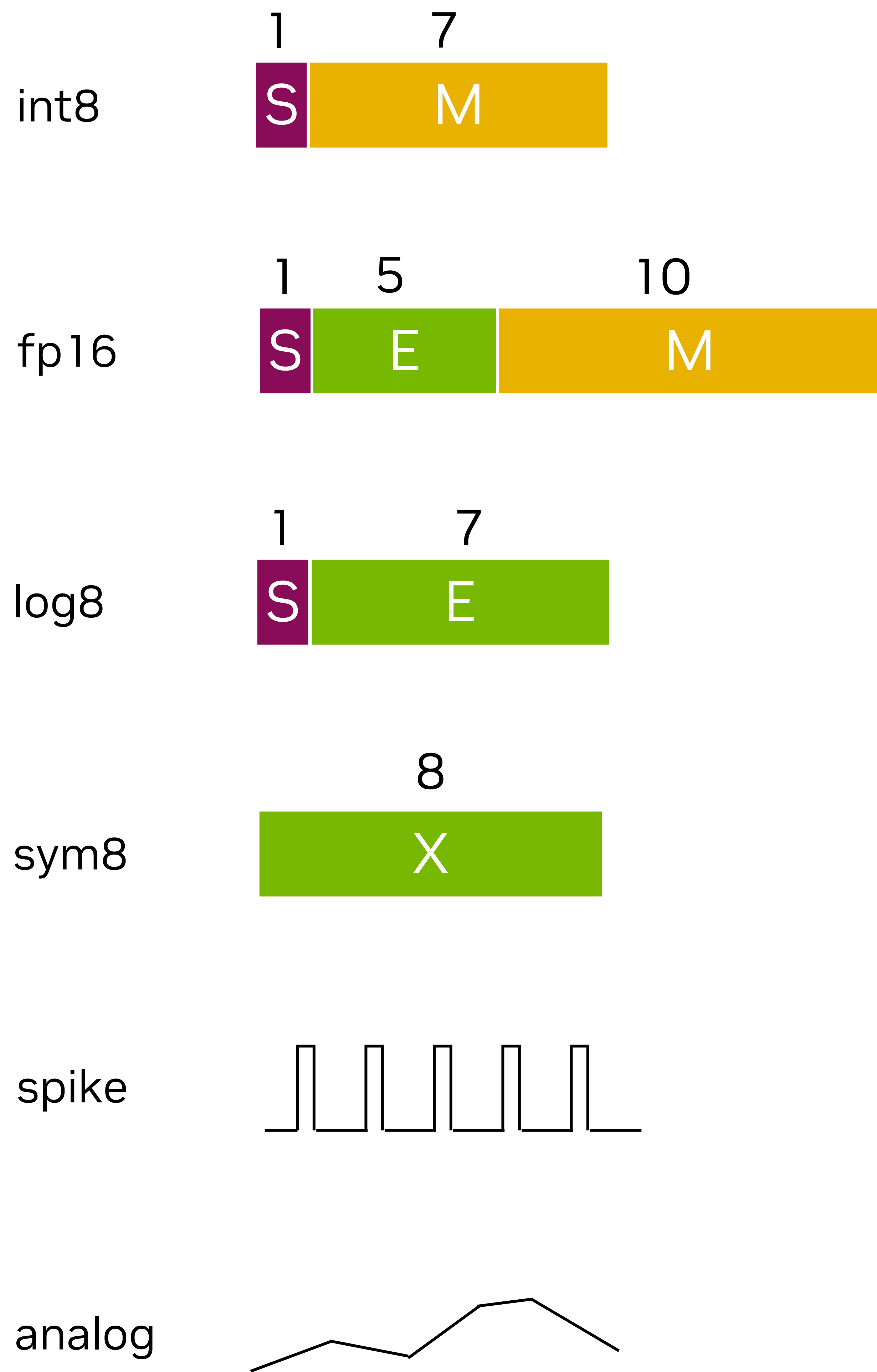
## Process

- Capacitance scaling

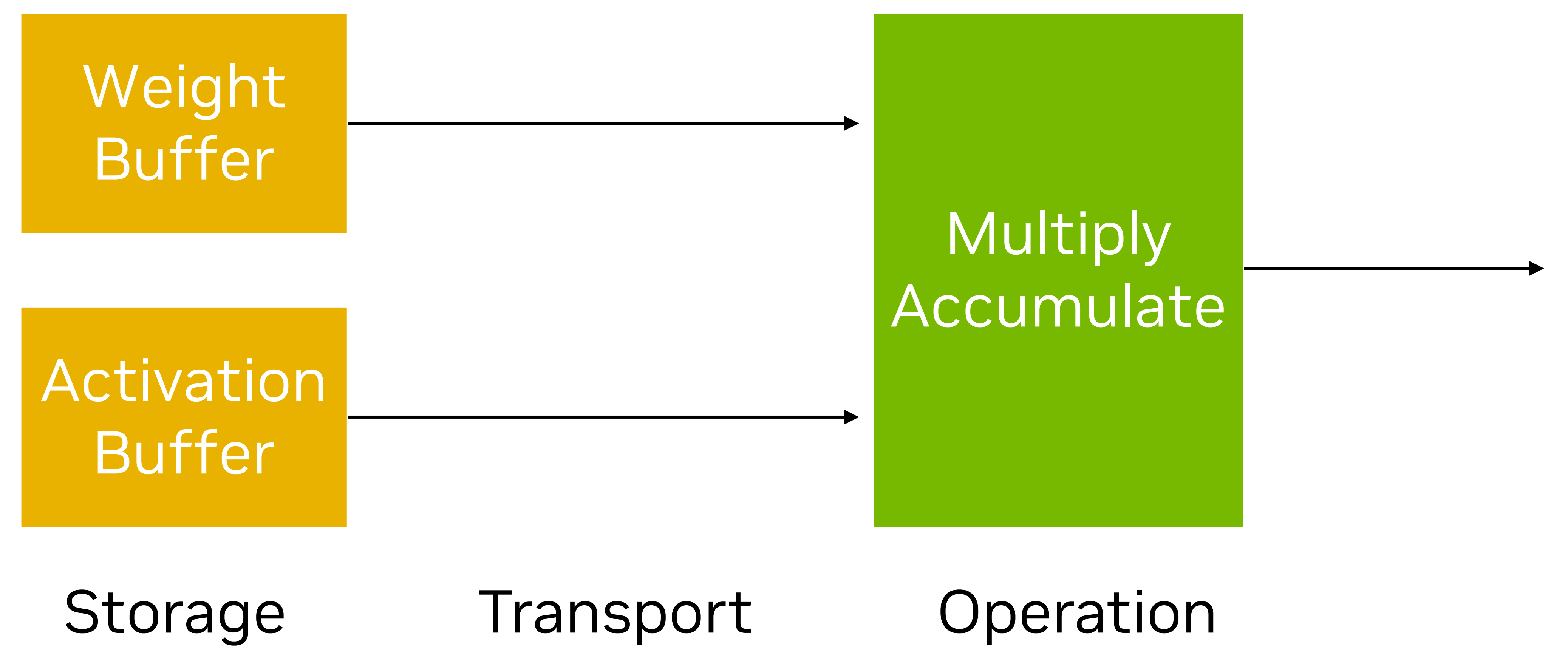


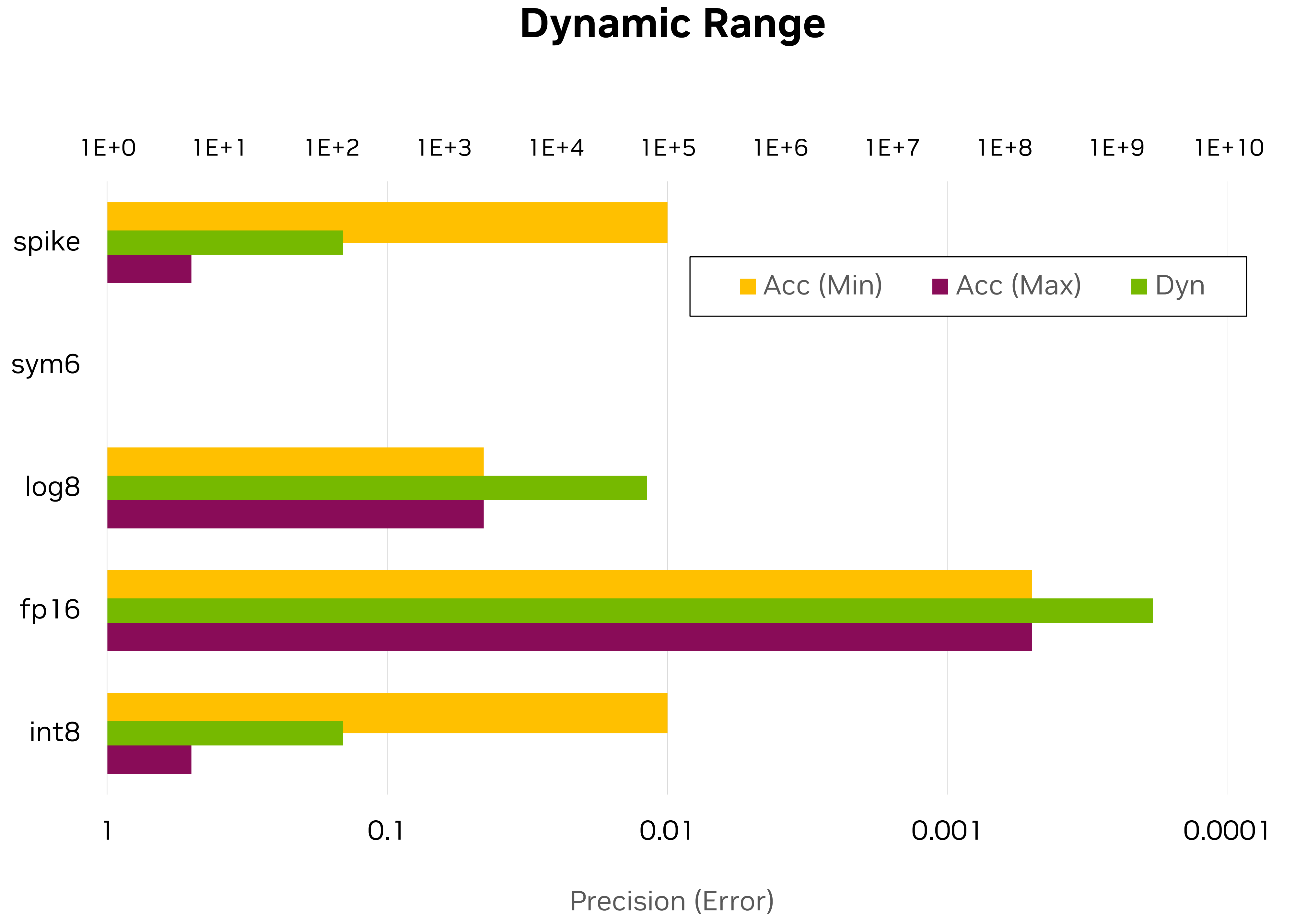
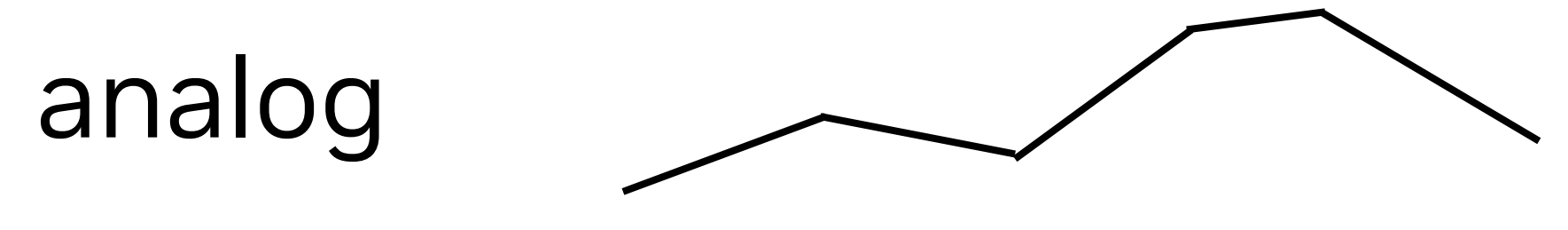
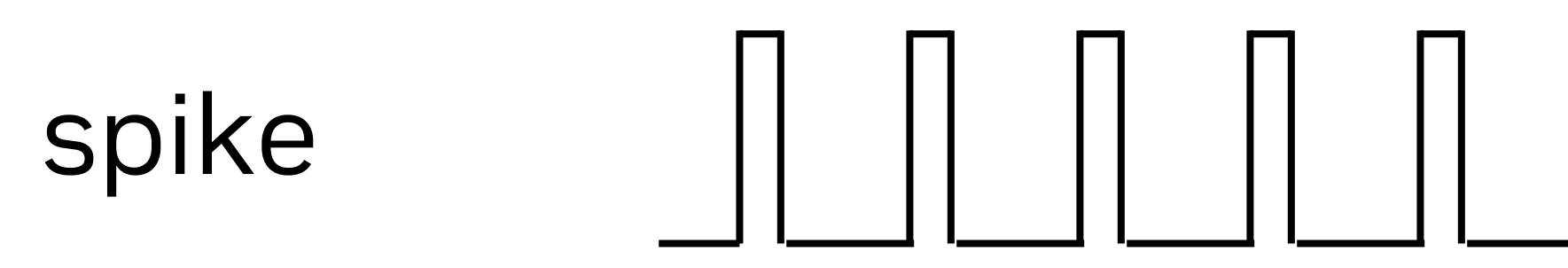
# Number Representation



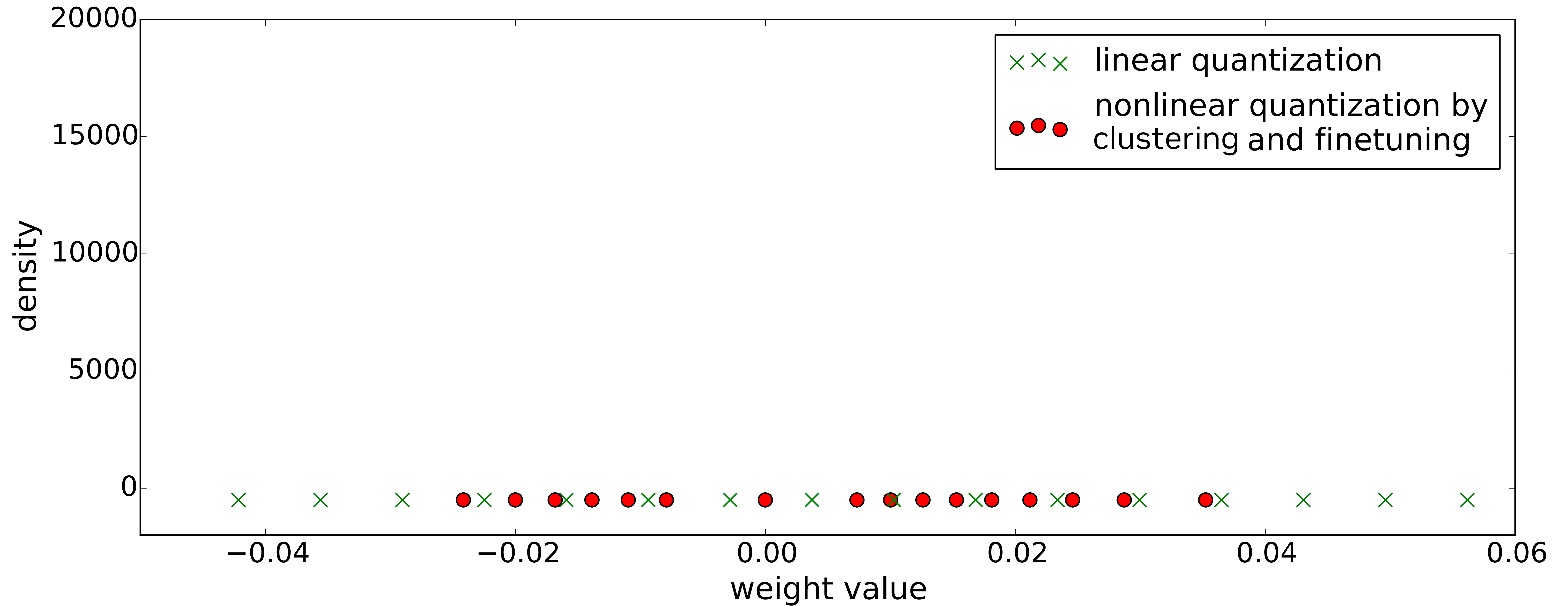


- Attributes:
  - Cost
    - Operation energy
    - Movement energy
  - Accuracy
    - Dynamic range
    - Precision (error)

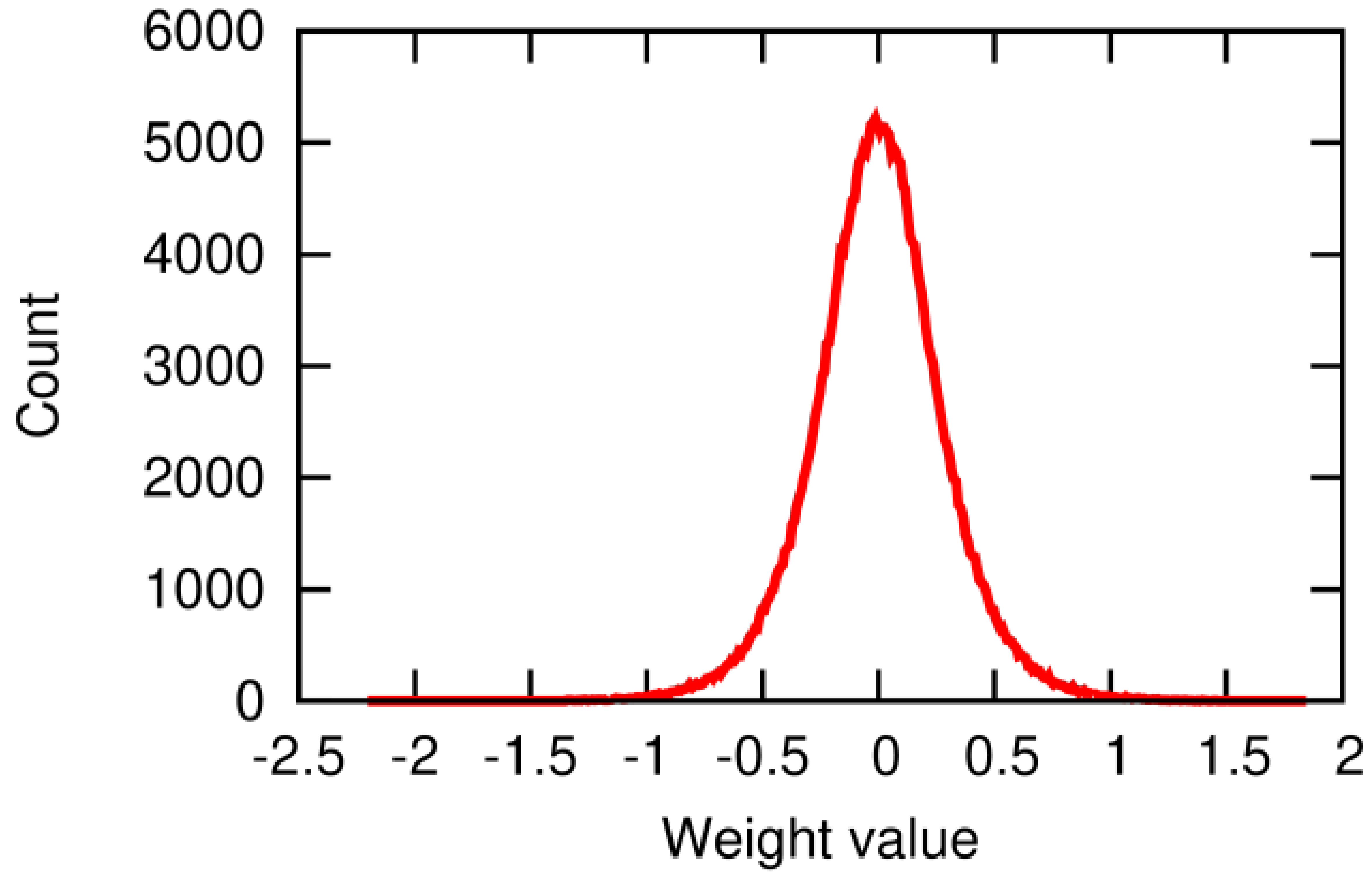




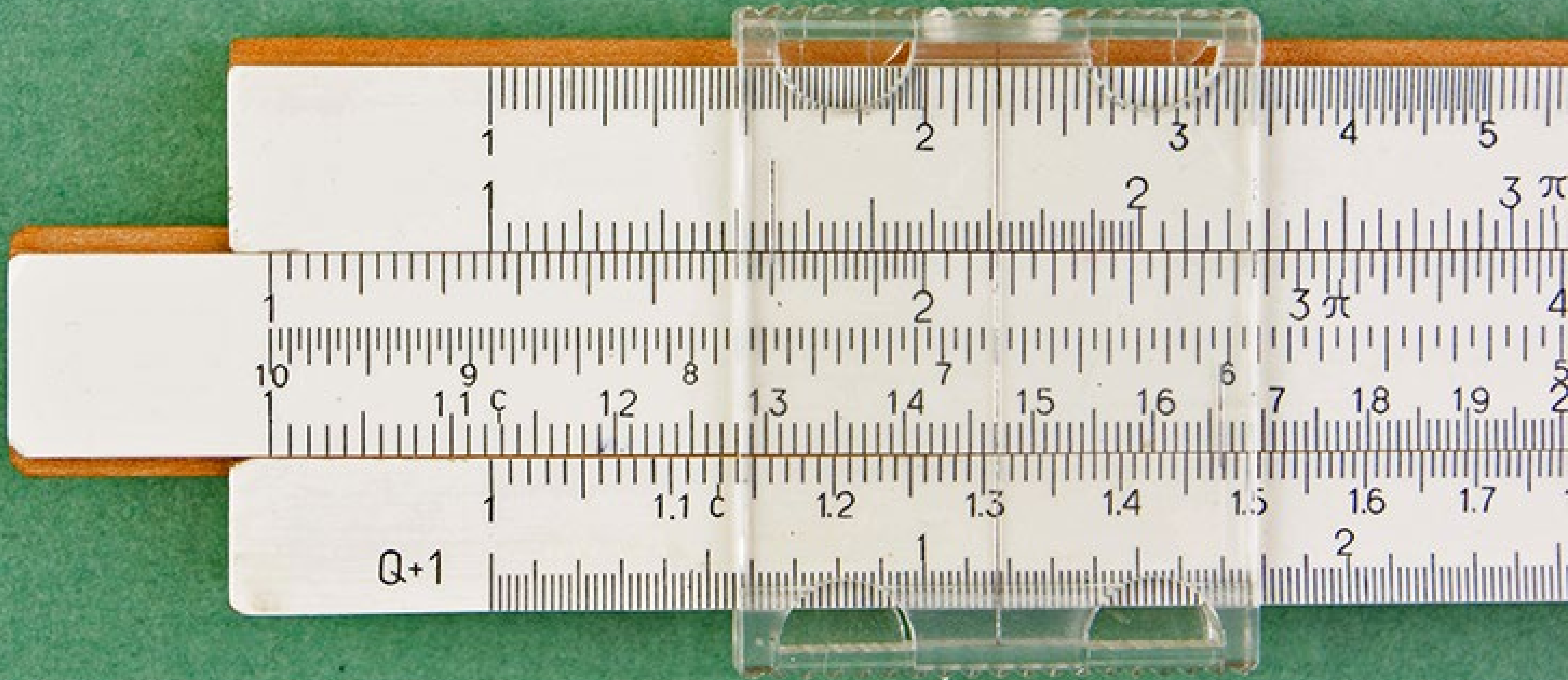
# Symbol Representation (Codebook)



Weight distribution of layer 1 (PTB small)



# Log Representation



Q+1

1

1.1 c

1.2

1.3

1.4

1.5

1.6

1.7

1

2

1

2

3

4

5

1

2

3  $\pi$

1

2

3  $\pi$

4

10

9

8

7

6

5

1.1 c

1.2

1.3

1.4

1.5

1.6

1.7

1.8

1.9

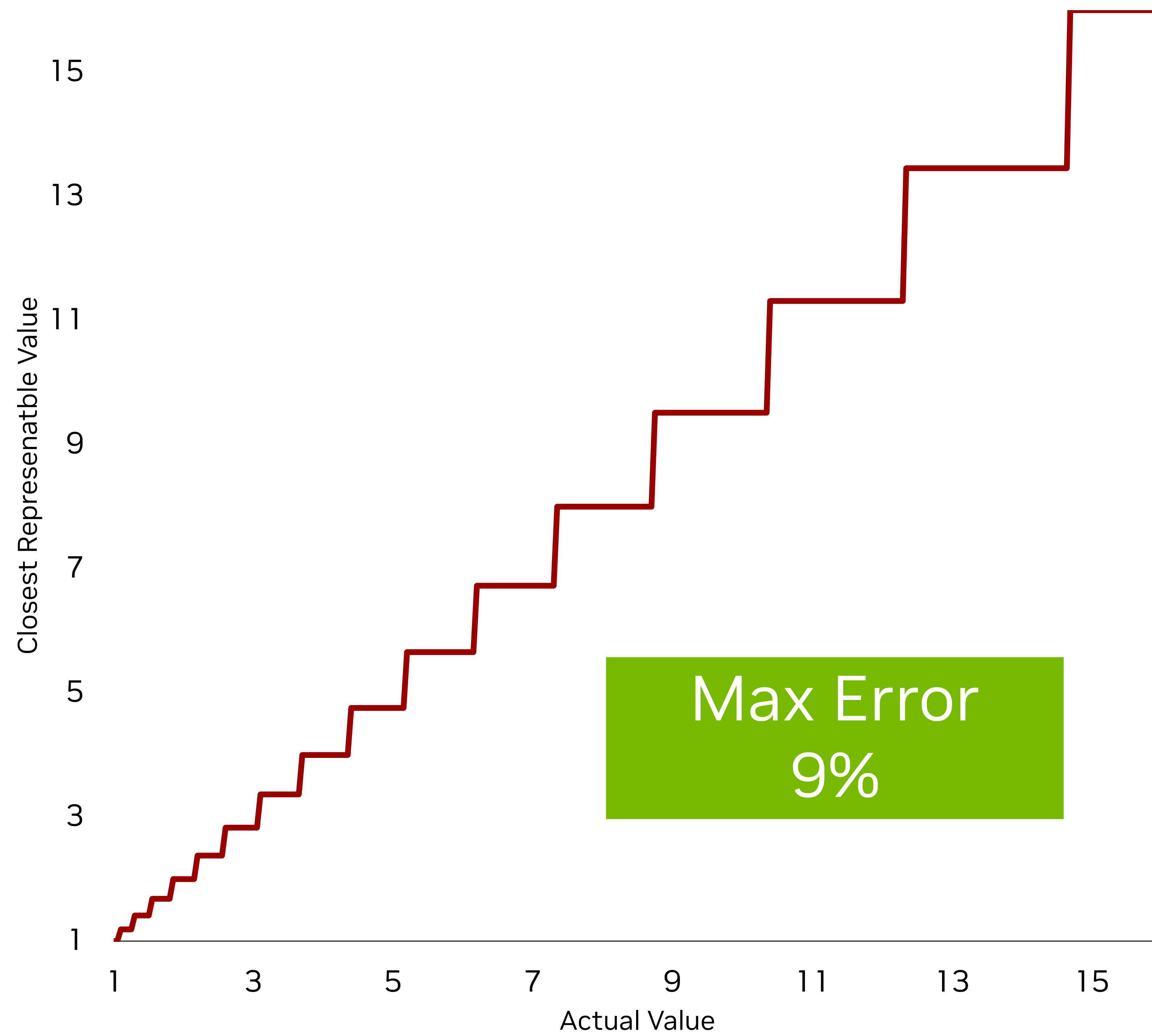
2

- Dynamic Range  $10^5$
- WC Accuracy 4%
  
- Vs Int8 – DR  $10^2$
- WC Accuracy 33%

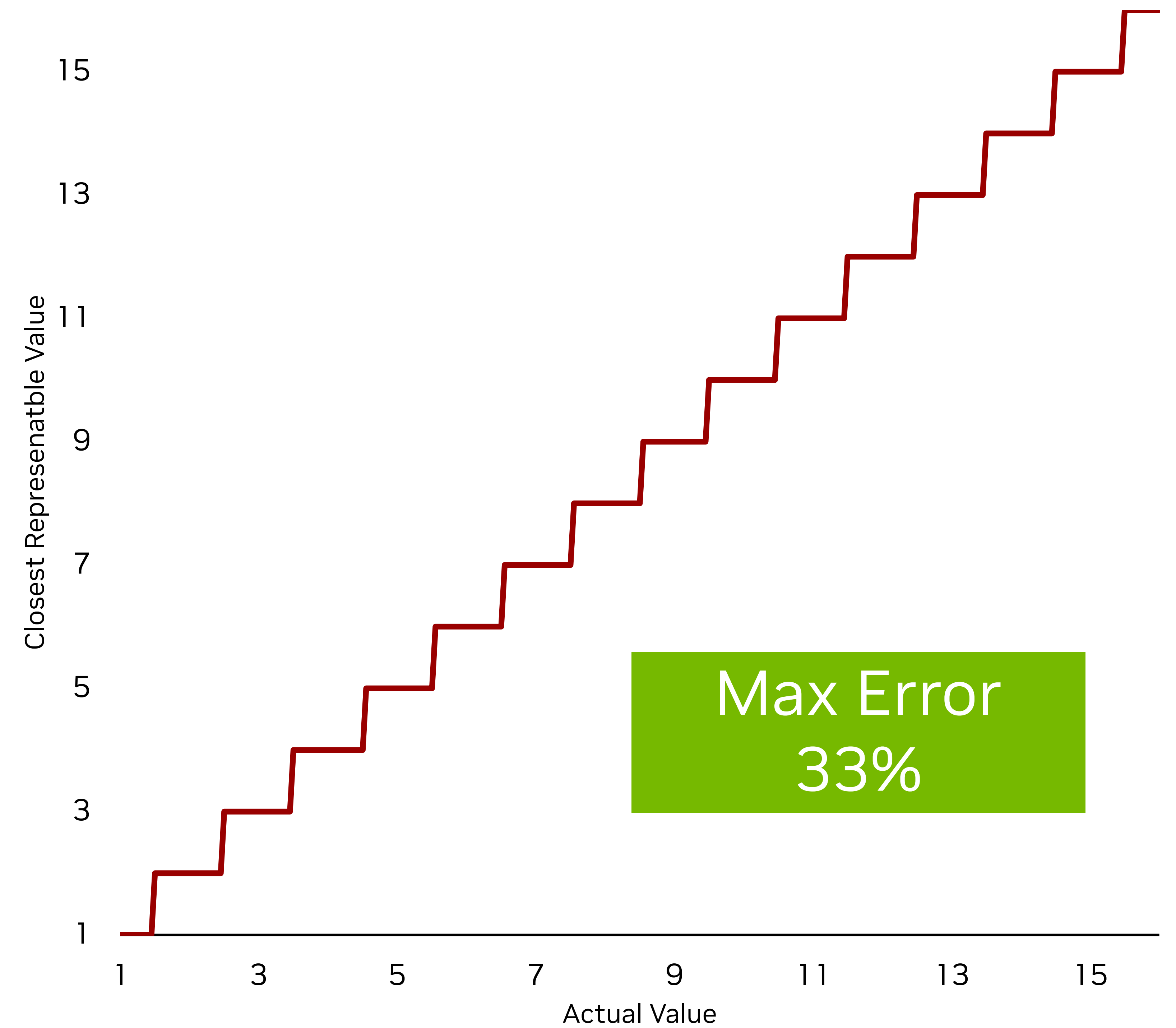
$$v = -1^s 2^{ei.ef}$$



### 4-bit Log Representation (L2.2)

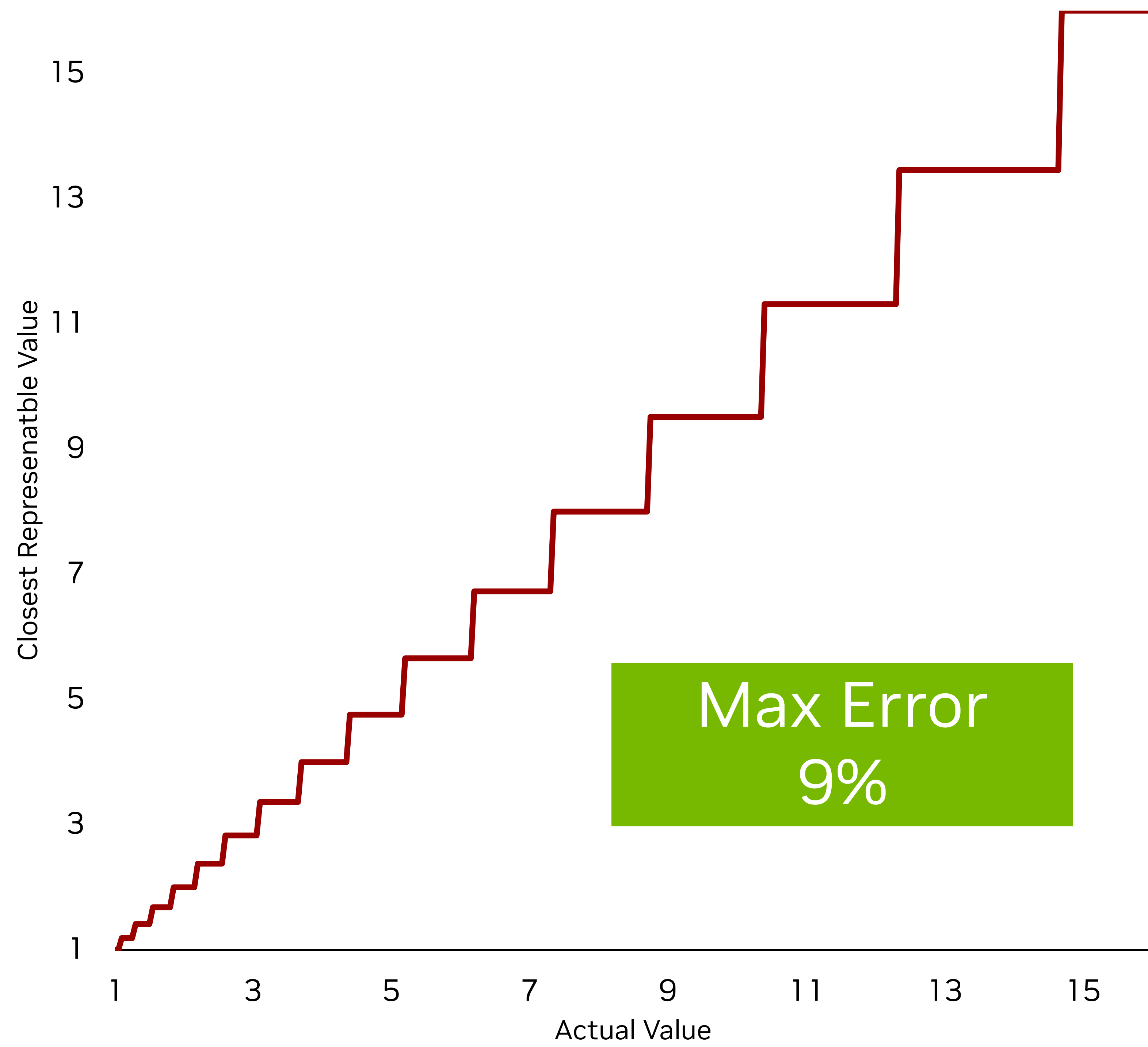


### 4-bit Integer Representation (Int4)

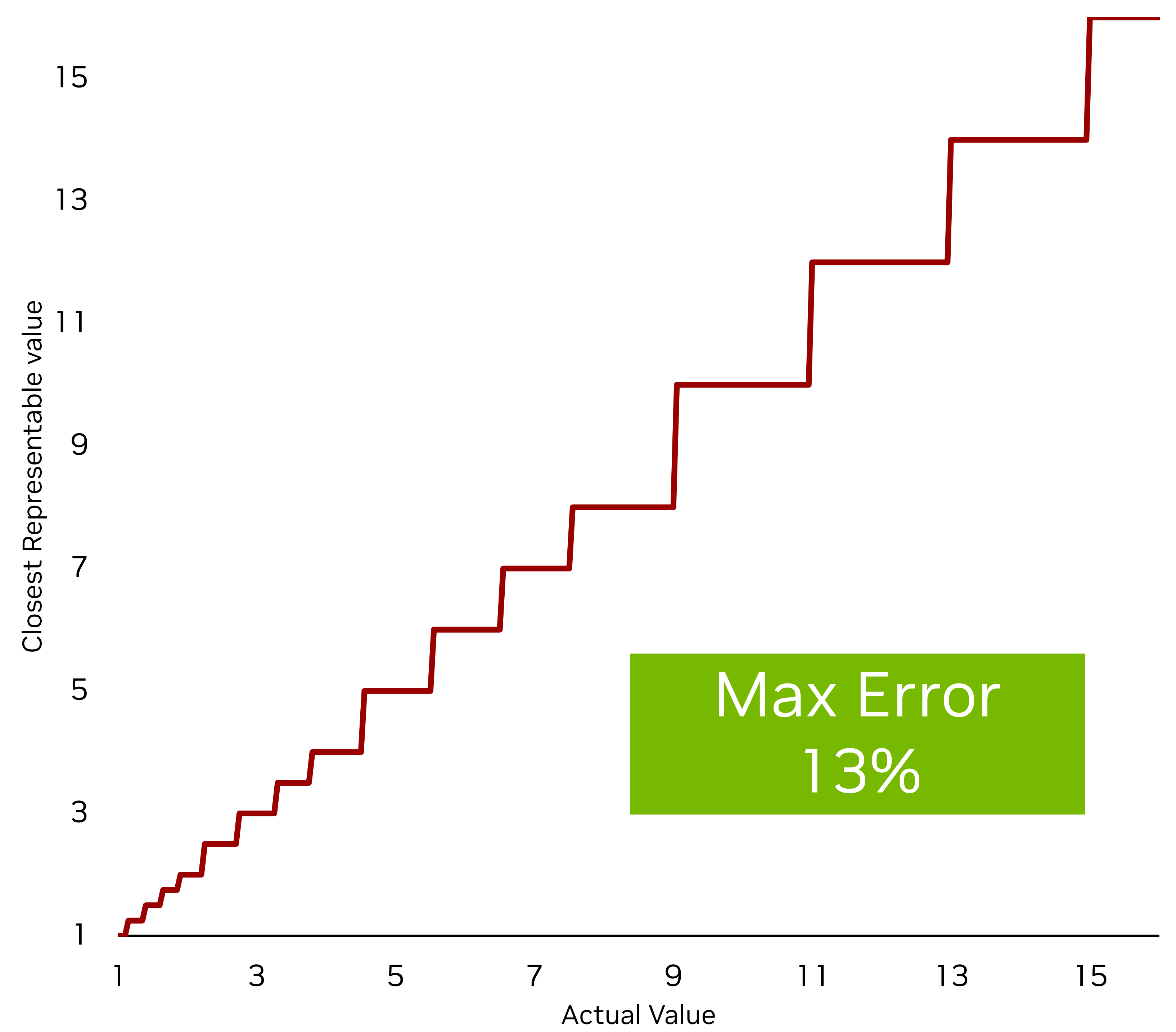




### 4-bit Log Representation (L2.2)



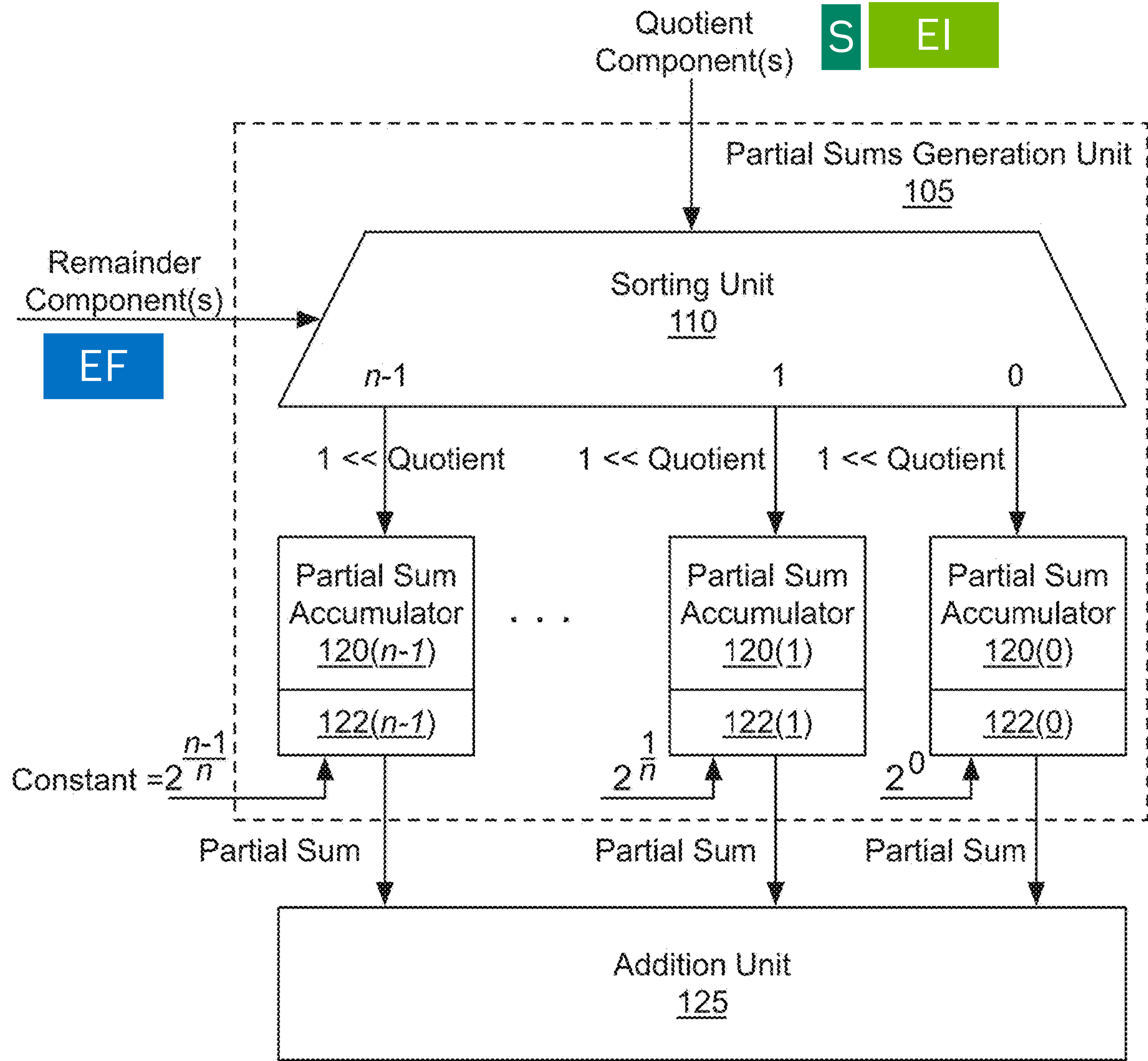
### FP2.2



- Log Numbers
- Multiplies are cheap – just an add
- Adds are hard – convert to integer, add, convert back
  - Fractional part of log is a lookup
  - Integer part of log is a shift
- Can factor the lookup outside the summation
  - Only convert back after summation (and NLF)



S EI

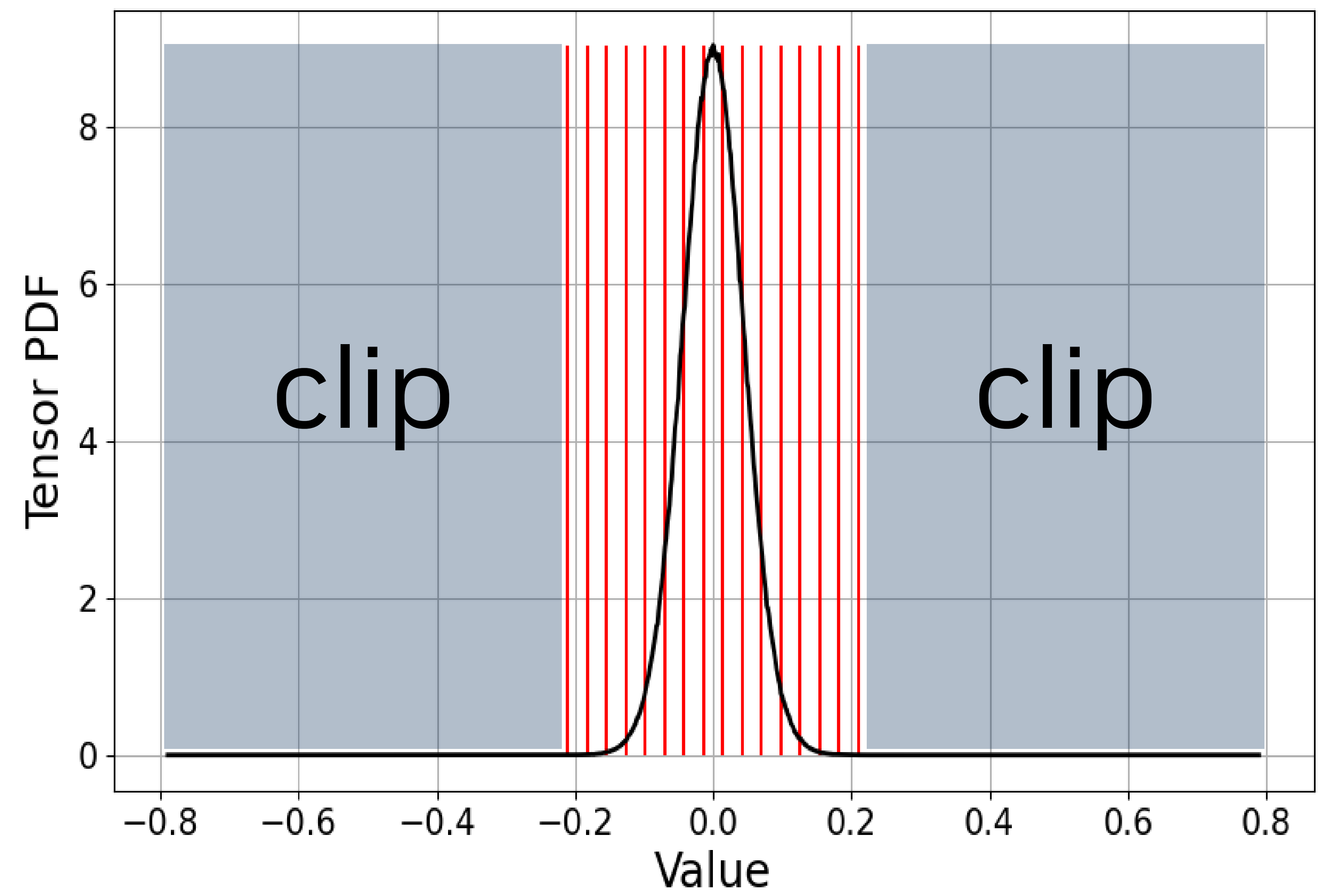
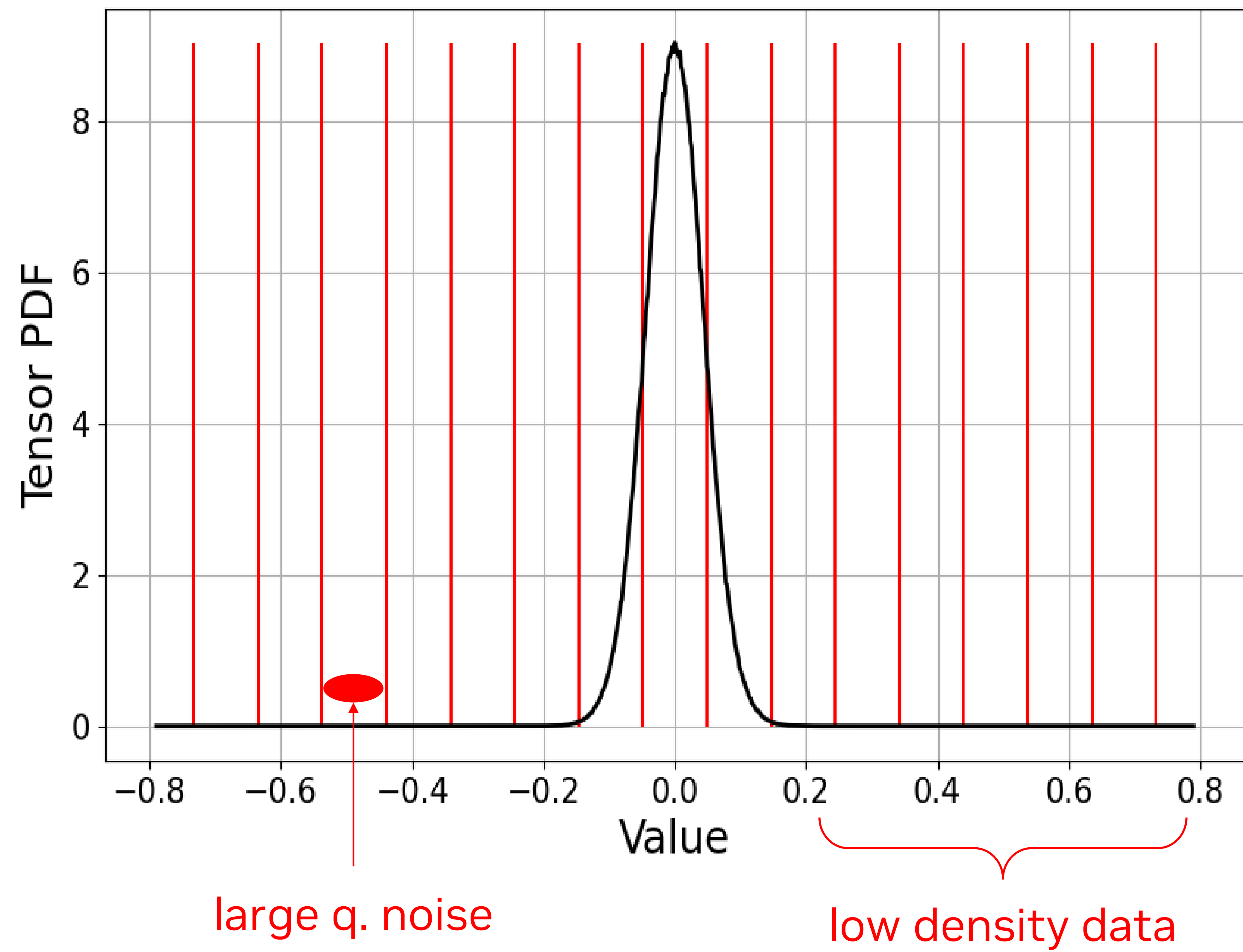


1 4 3  
S EI EF

# Optimum Clipping

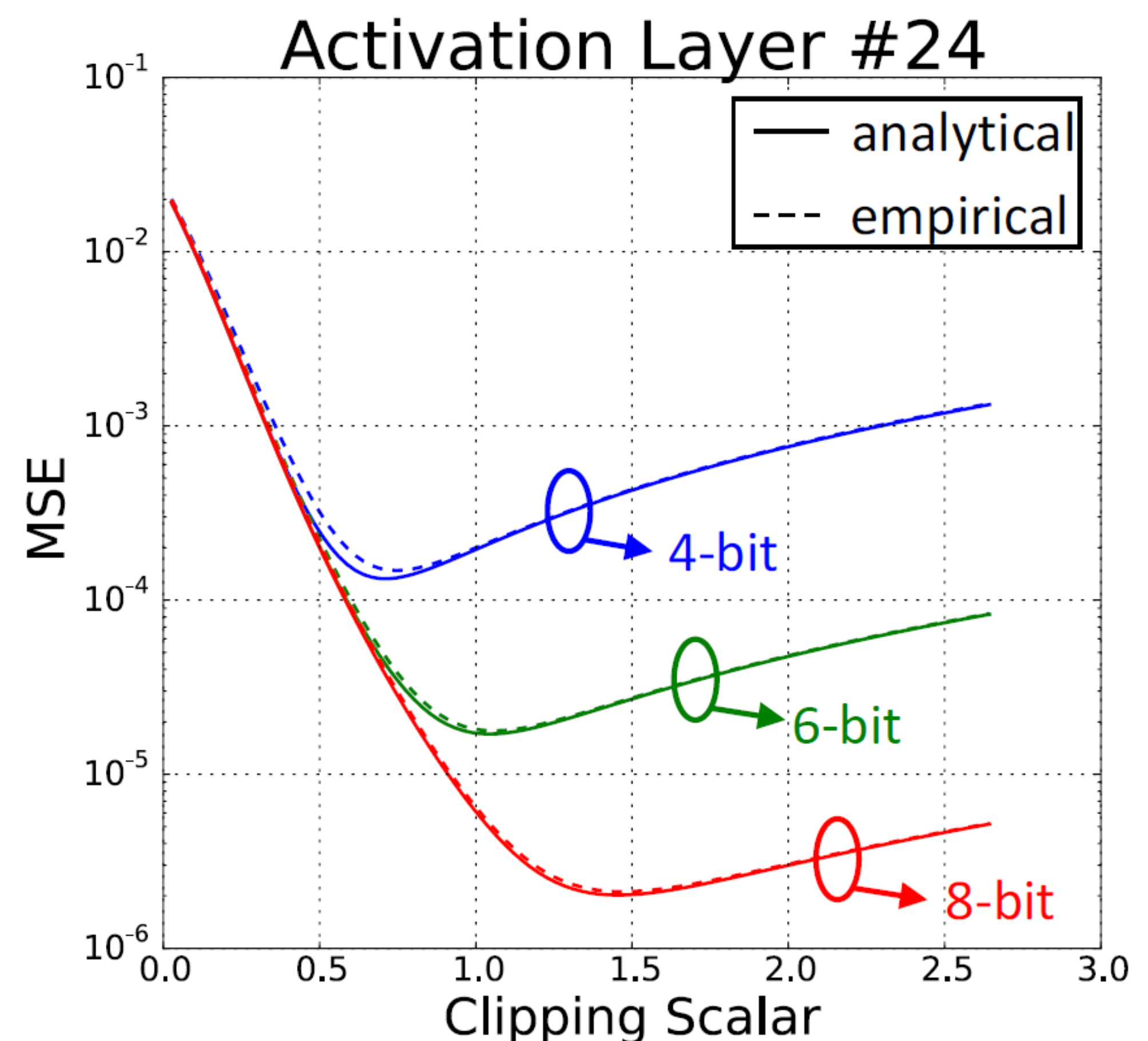
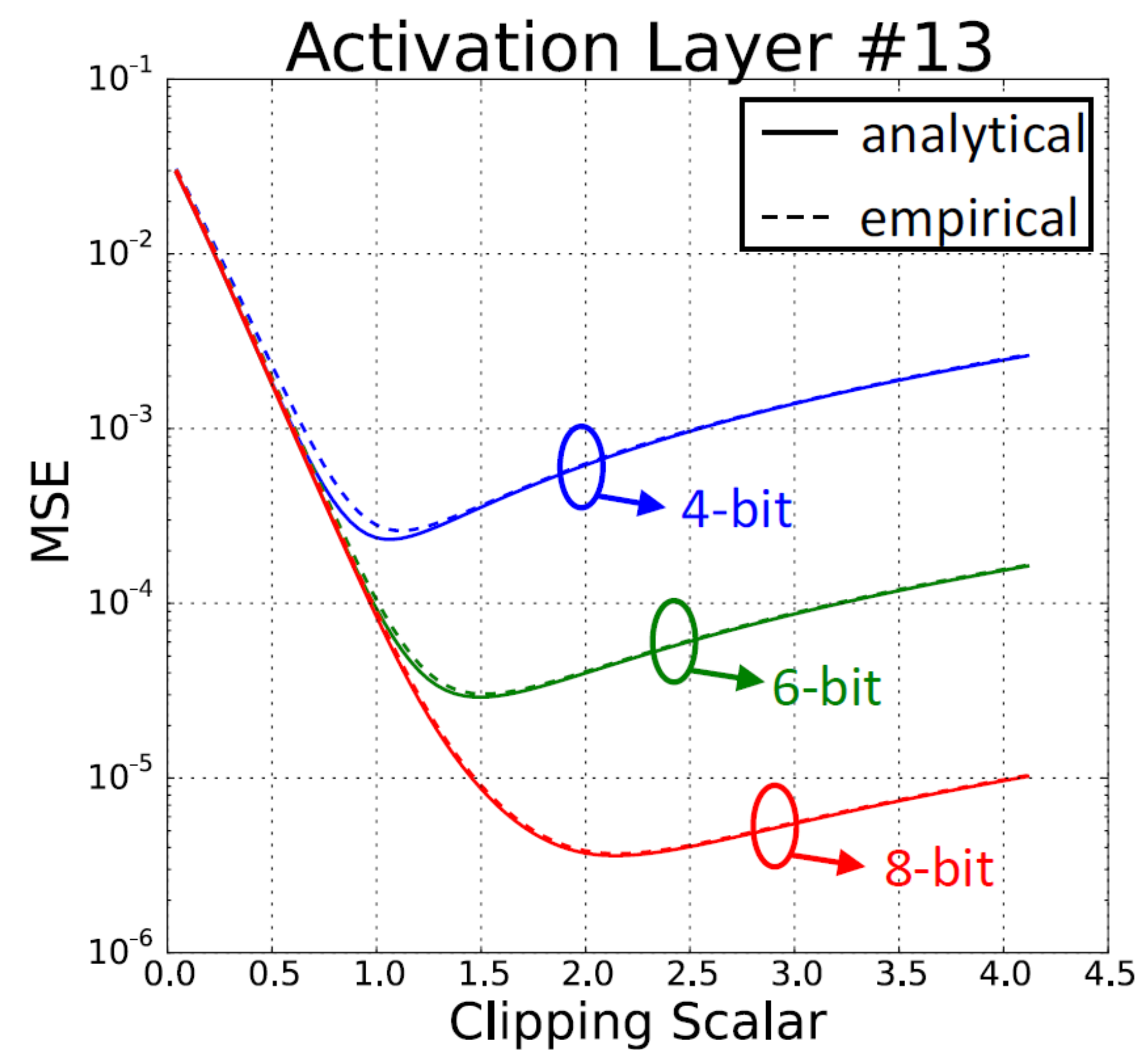
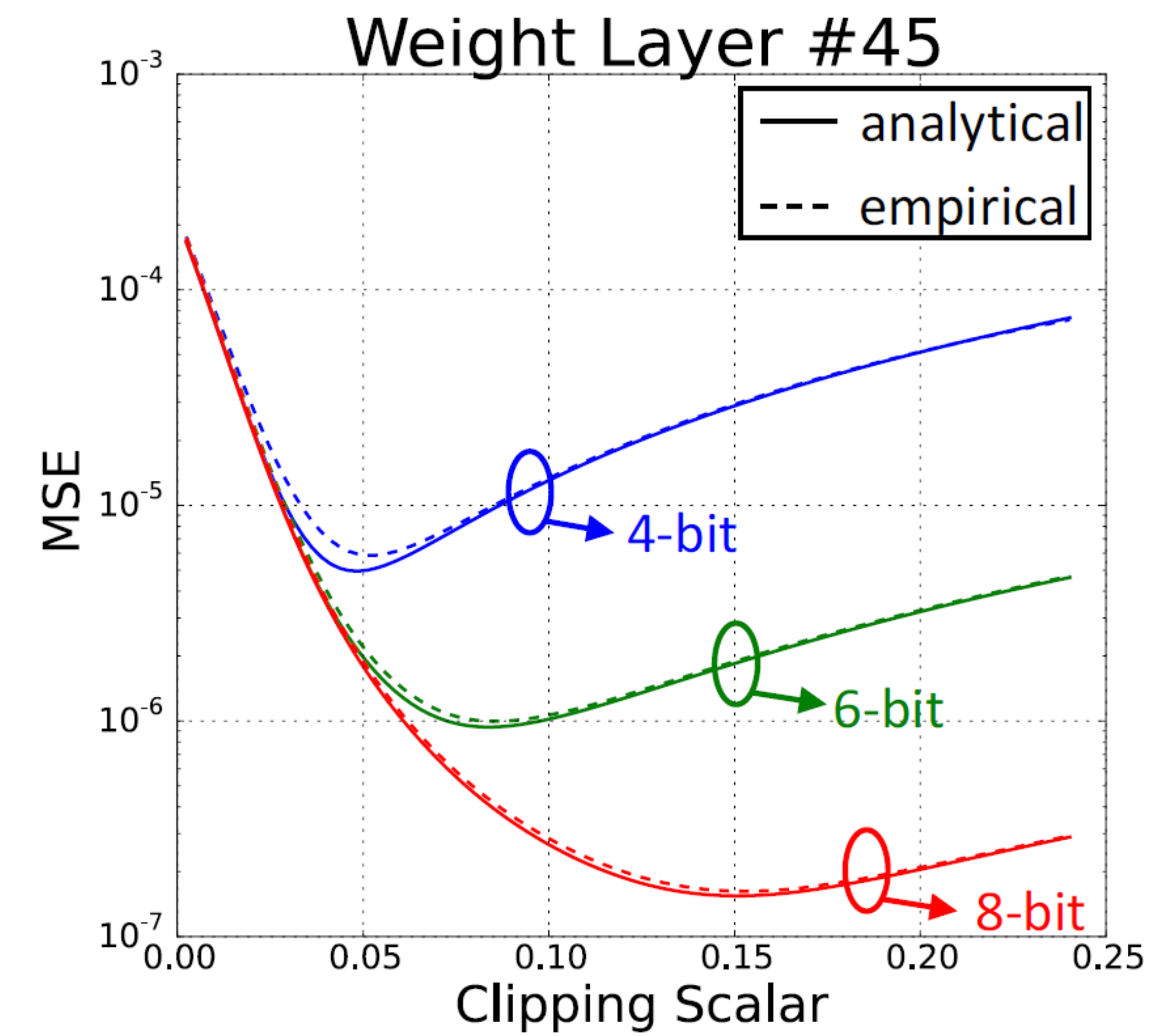
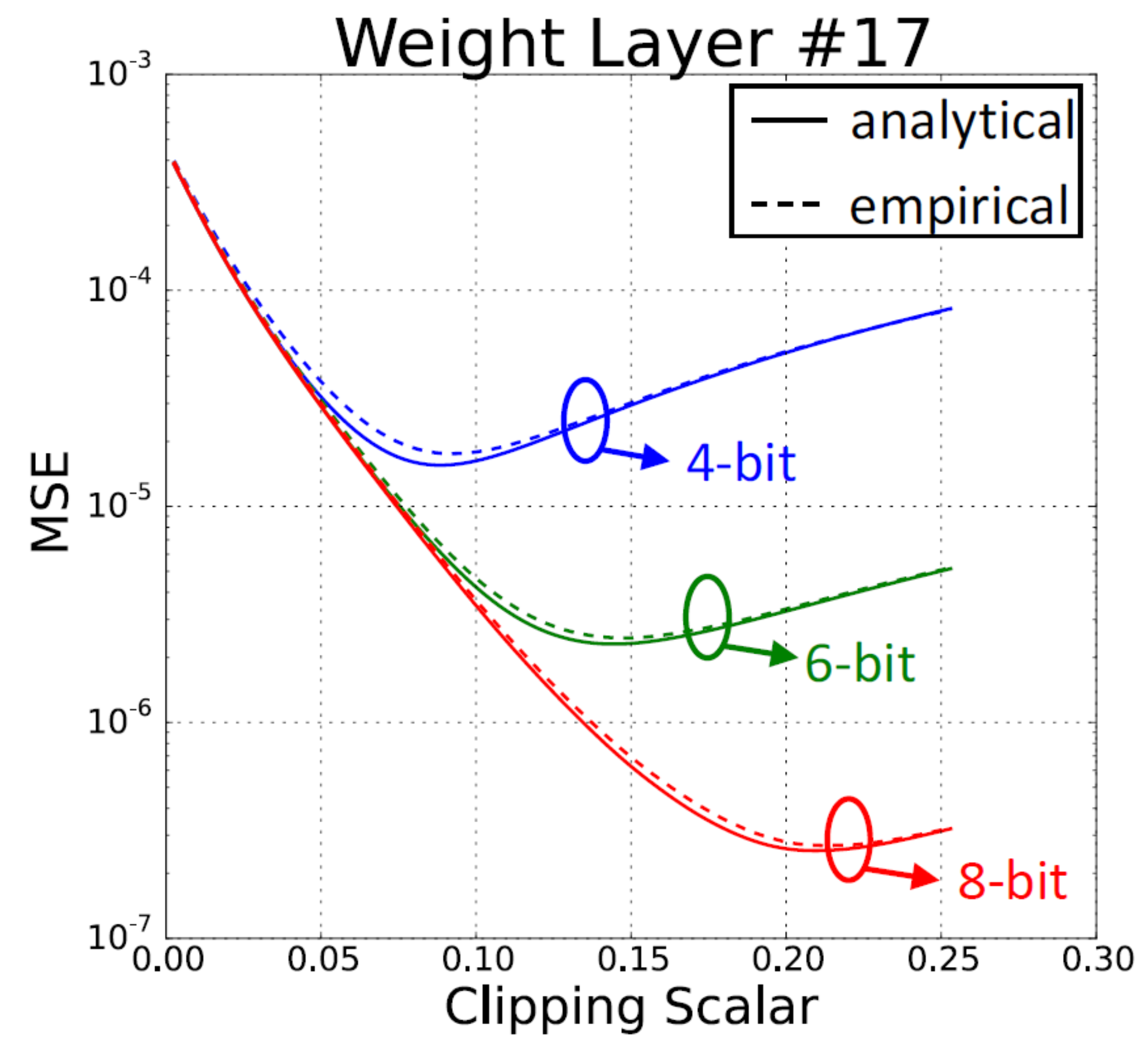
**Whatever number representation you use**

**Pick the range optimally**



$$J = \frac{4^{-B}}{3} s^2 \int_0^s f_{|X|}(x) dx + \int_s^\infty (s - x)^2 f_{|X|}(x) dx$$

$$s_{n+1} = \frac{\mathbf{E}[|X| \cdot \mathbf{1}_{\{|X| > s_n\}}]}{\frac{4^{-B}}{3} \mathbf{E}[\mathbf{1}_{\{|X| < s_n\}}] + \mathbf{E}[\mathbf{1}_{\{|X| > s_n\}}]}$$

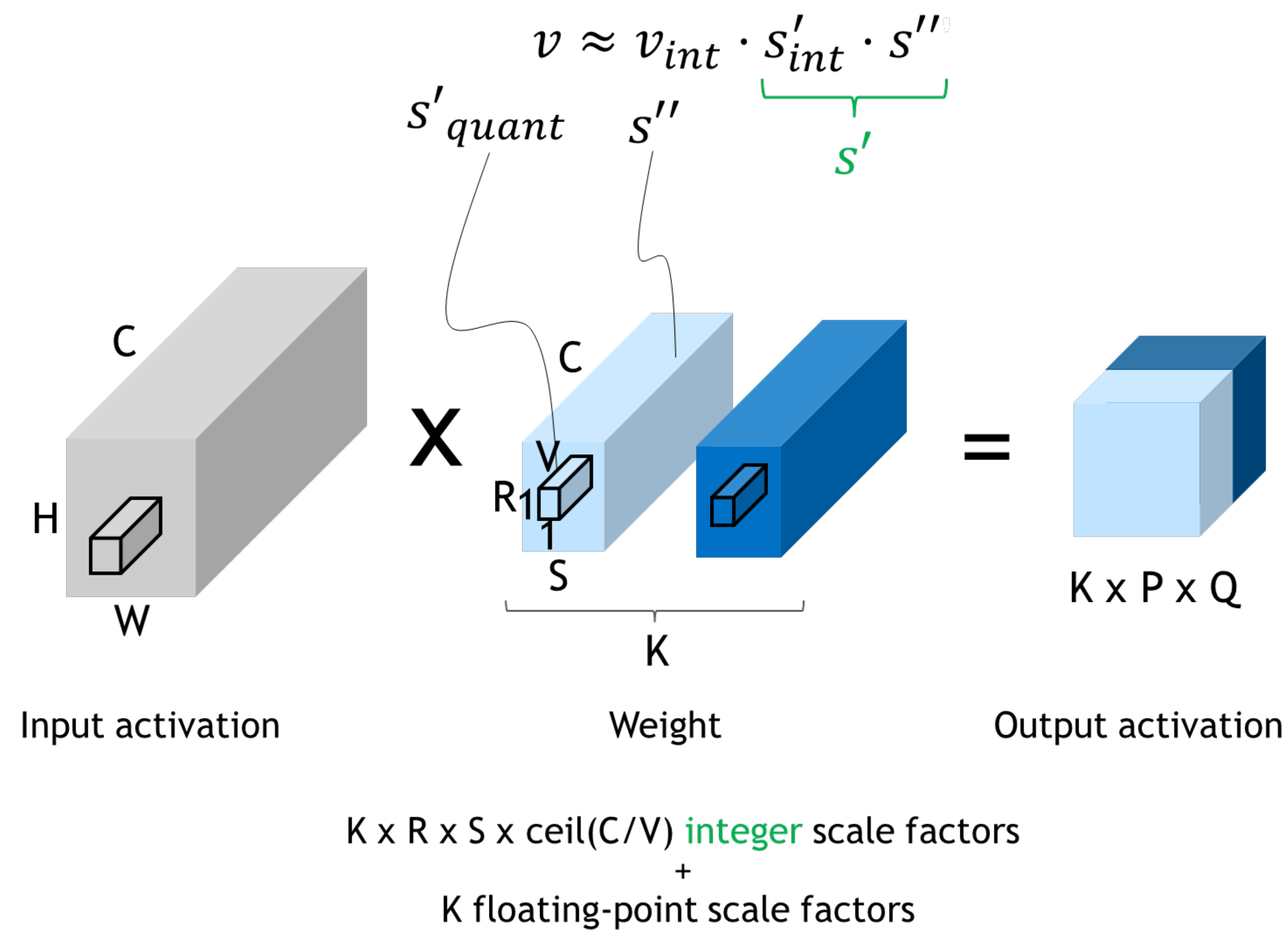


# Vector Scaling



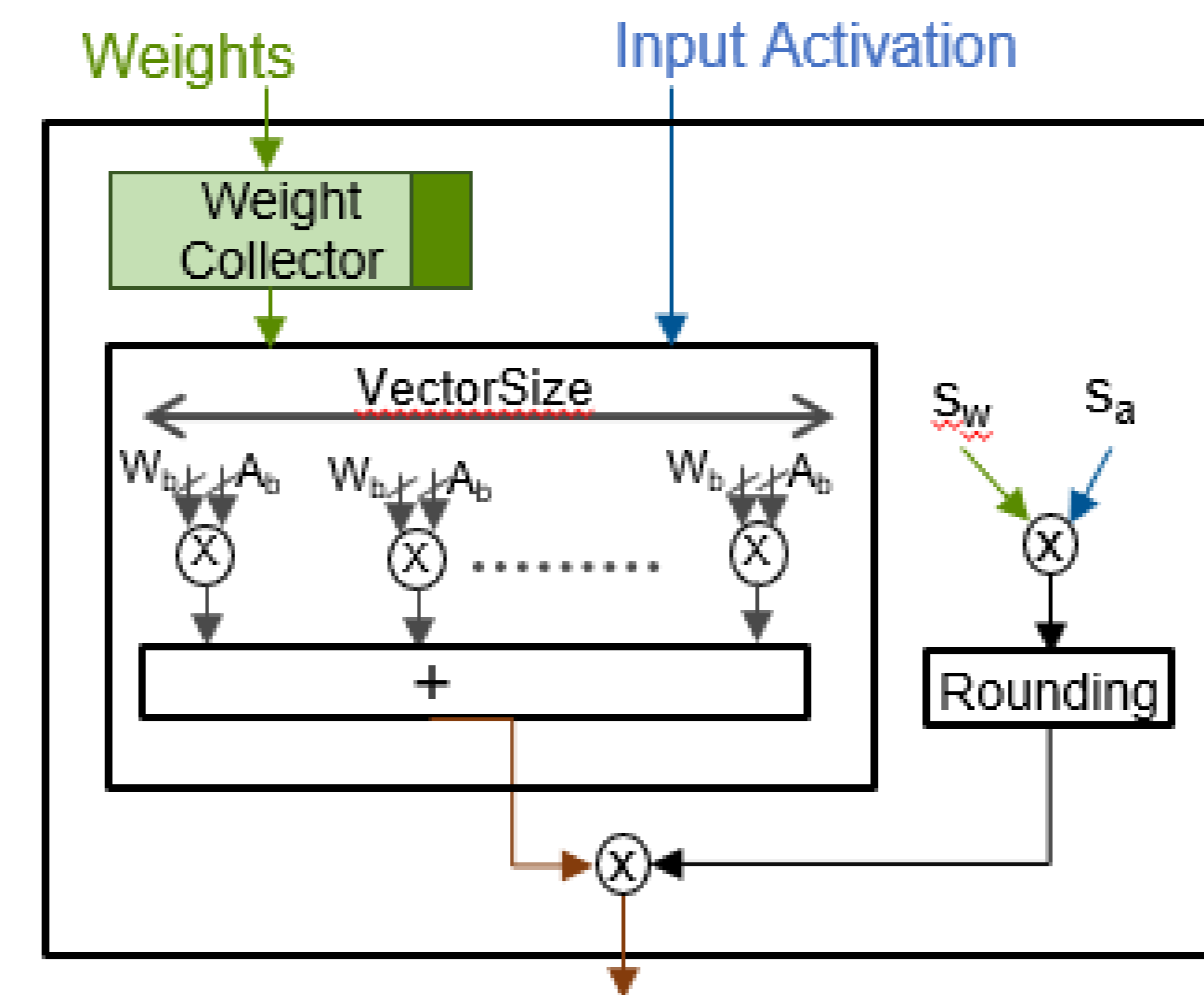
# VS-Quant

Per-vector scaled quantization for low-precision inference



*Fine-grained scale factors per vector*

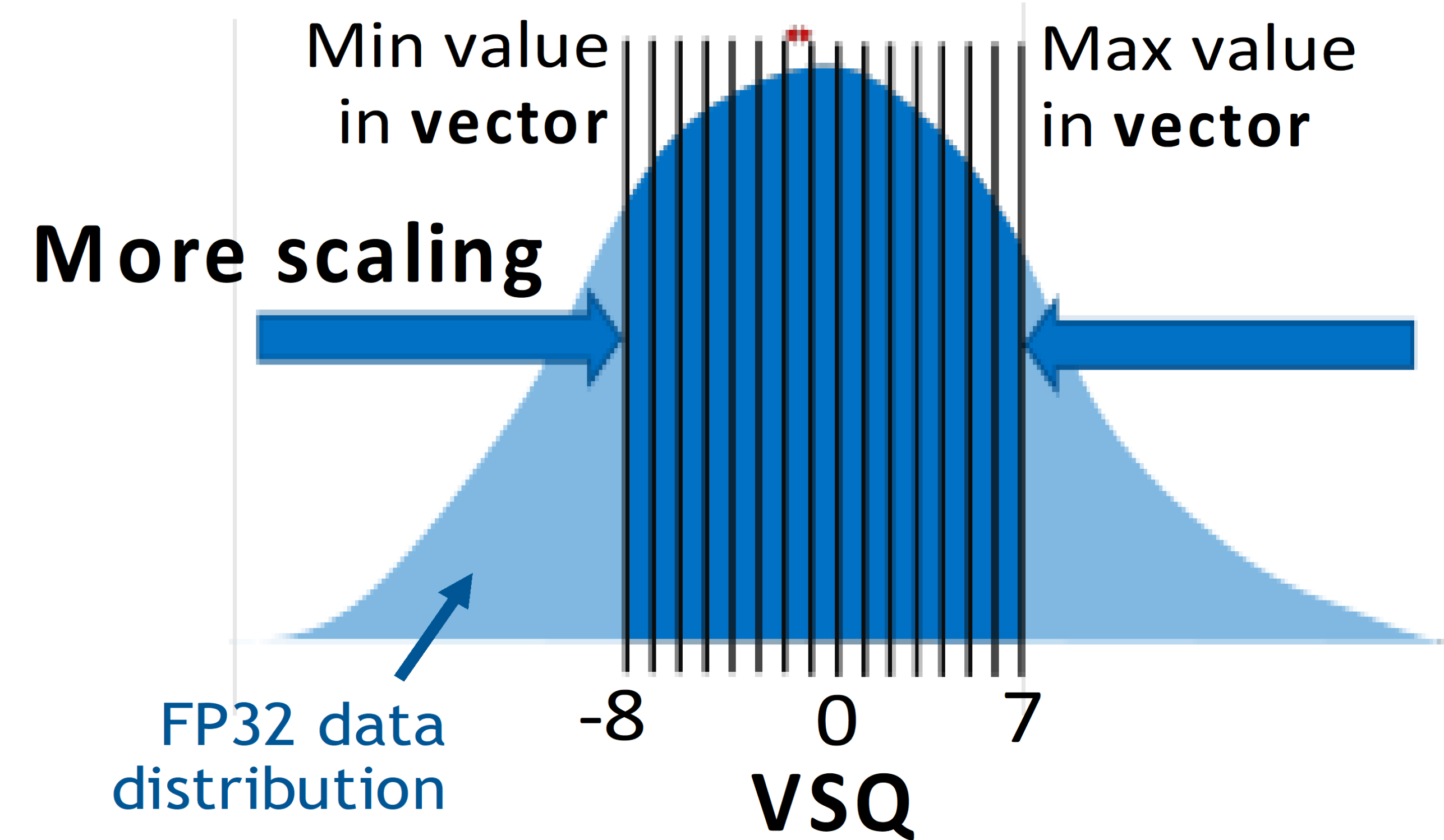
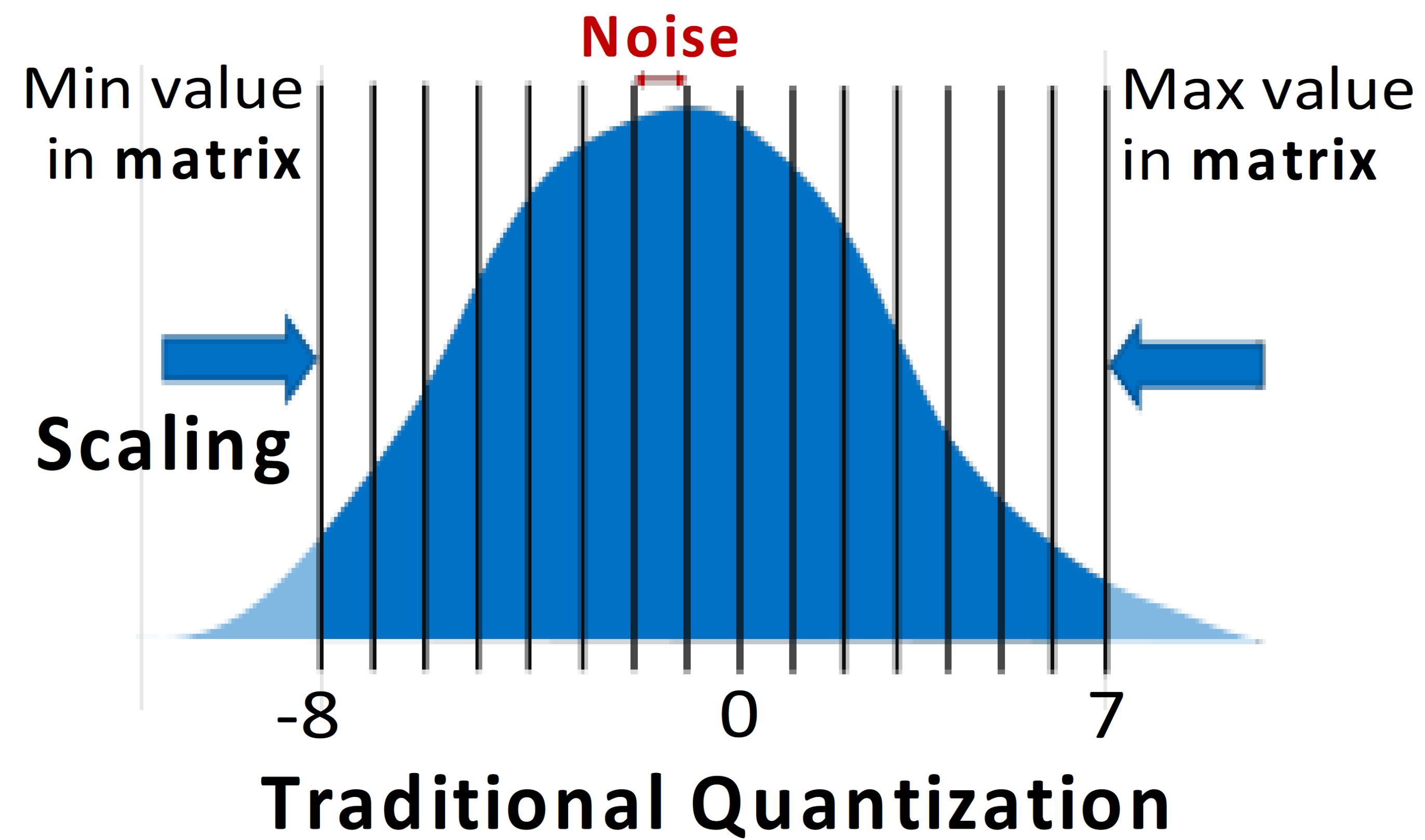
$$y_q(j) = \left( \sum_{i=0}^{vecsize-1} w_q(i) a_q(i) \right) s_w(j) s_a(j)$$



*Modified vector MAC unit for VS-Quant*

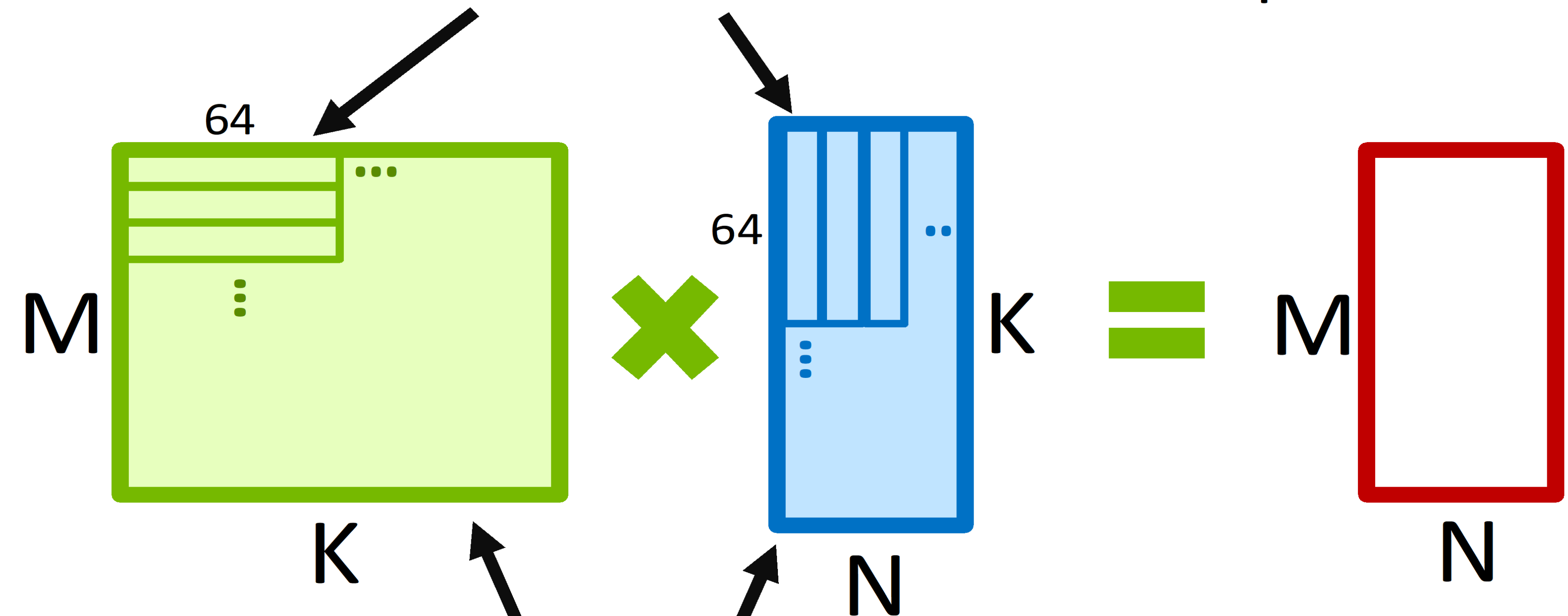
Works with either post-training quantization or quantization-aware retraining!

# INT4 Quantization



# VSQ Scale Factors

One scale factor for each 64-element input vector

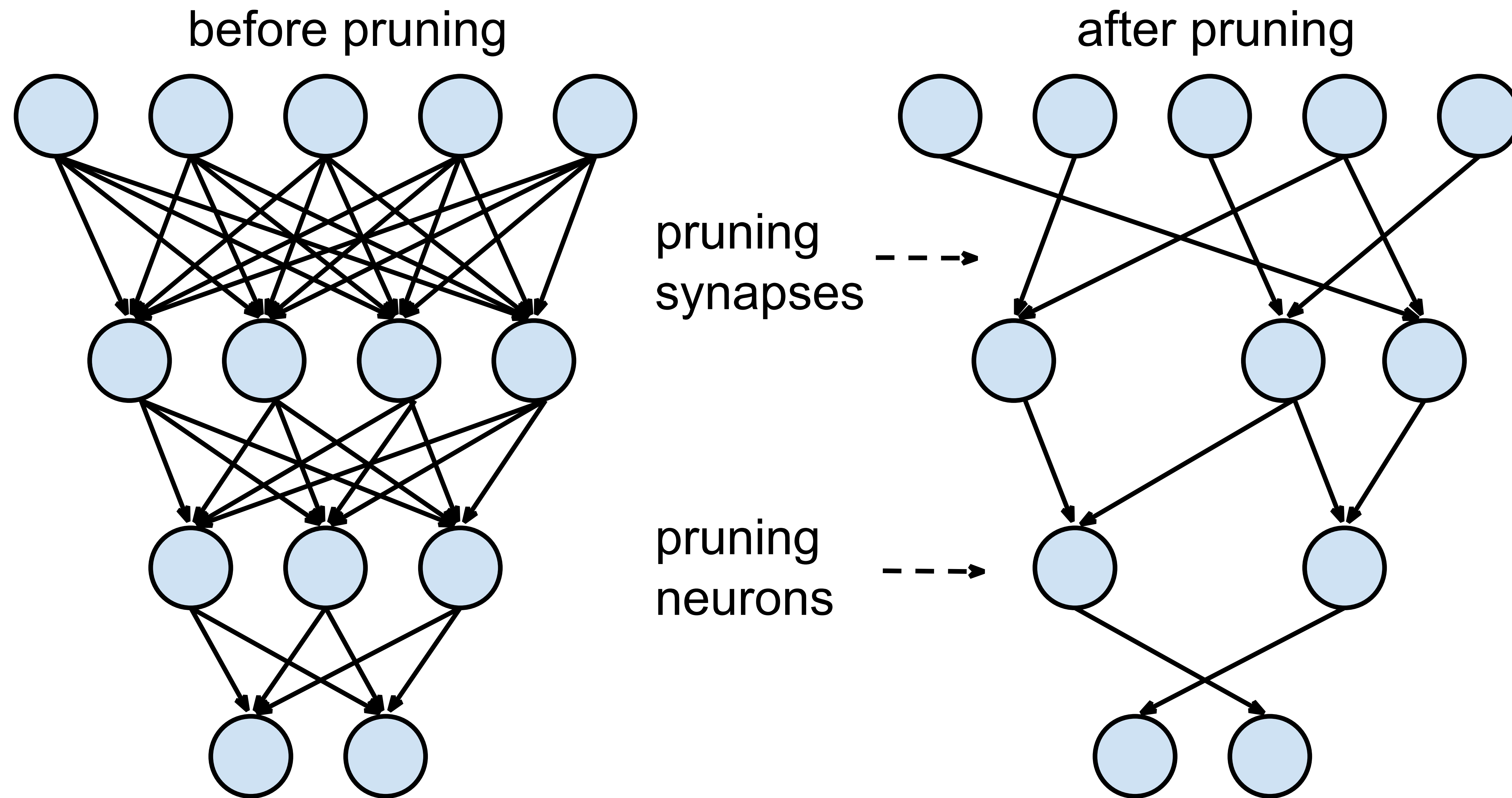


Second scale factor for each input matrix

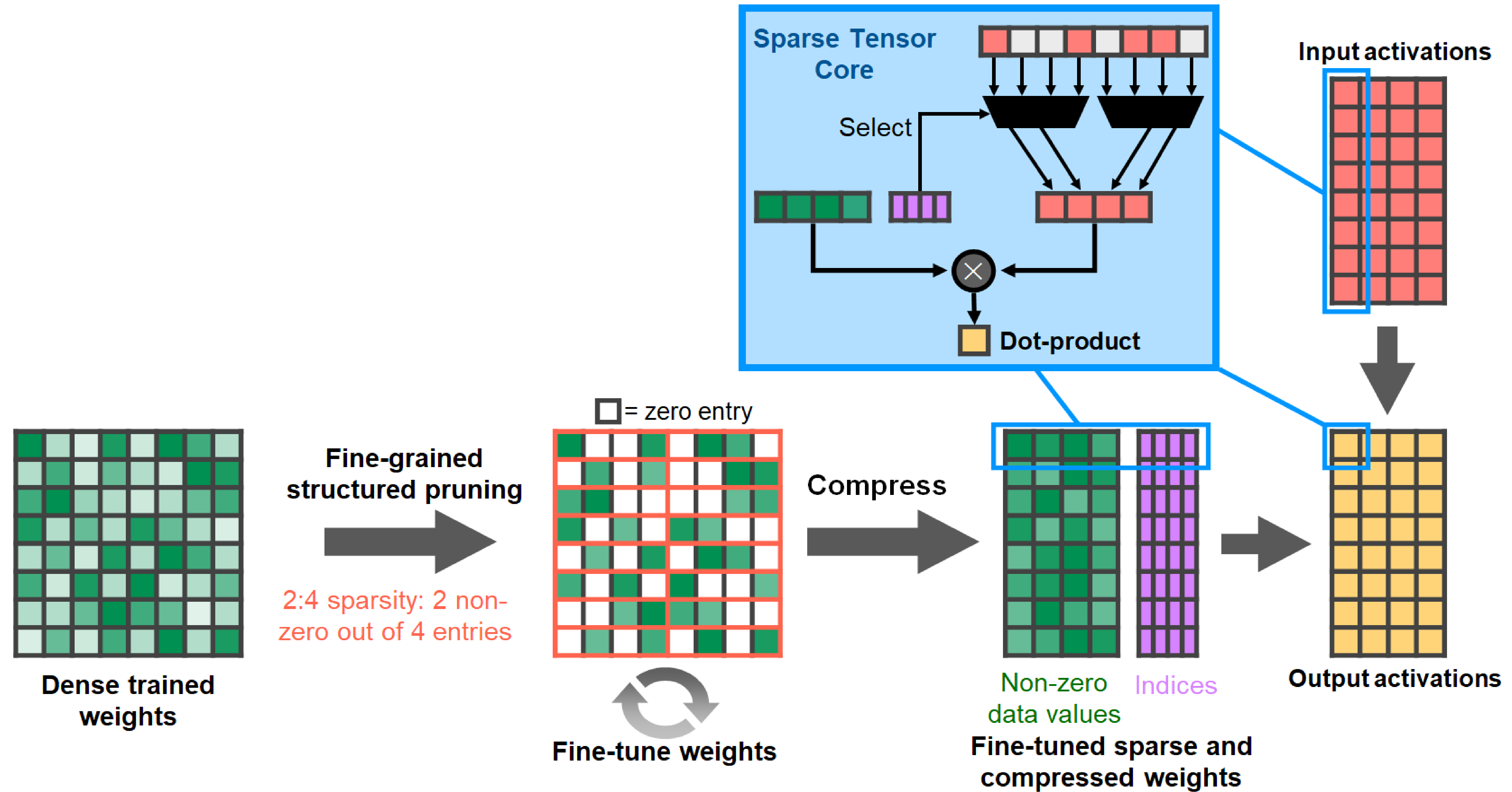
Traditional Quantization	VSQ
One scale factor per matrix	Two scale factors: one per vector, one per matrix
High quantization noise	Reduced quantization noise

**Sparsity**

# Pruning



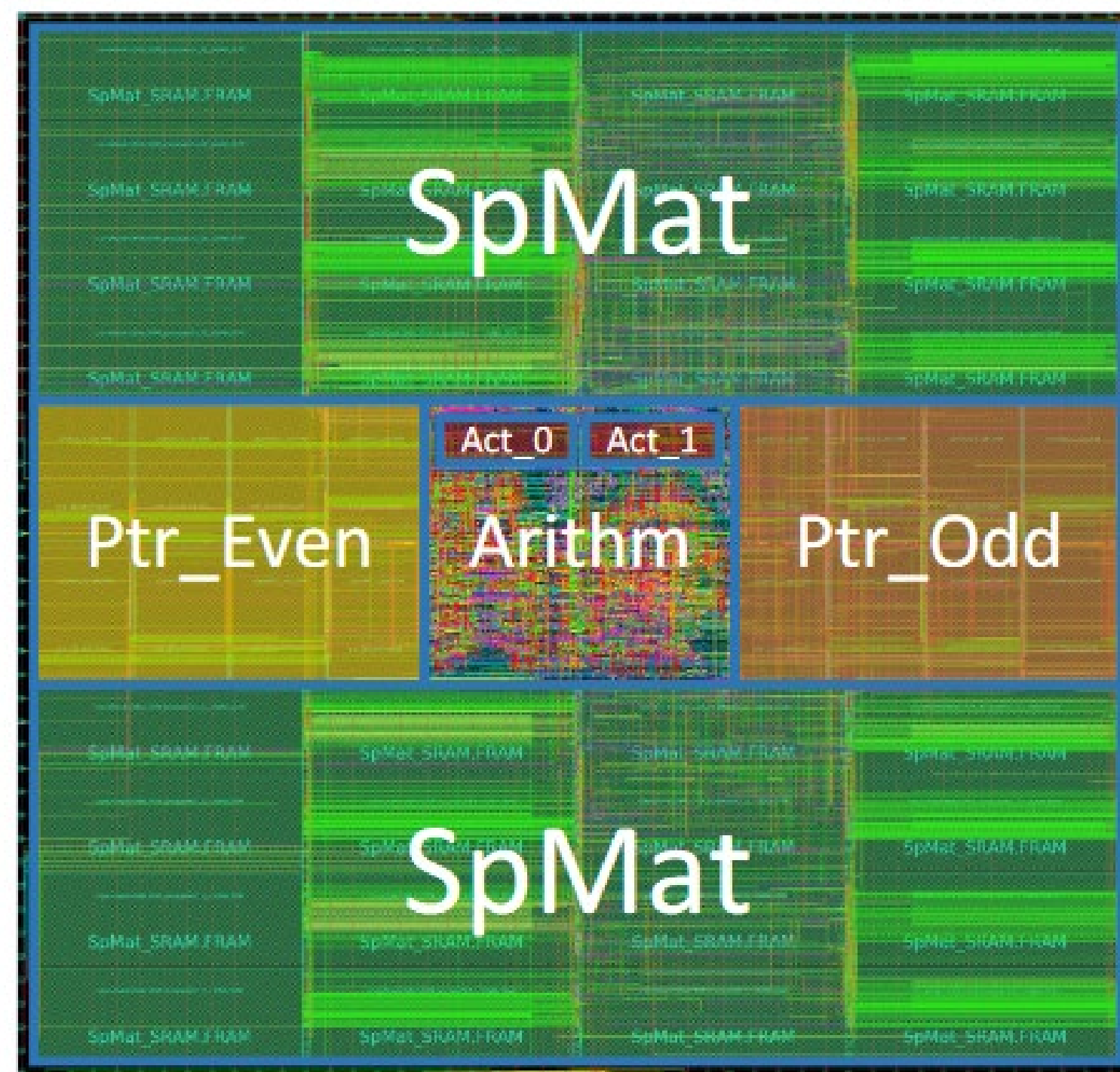
# Structured Sparsity



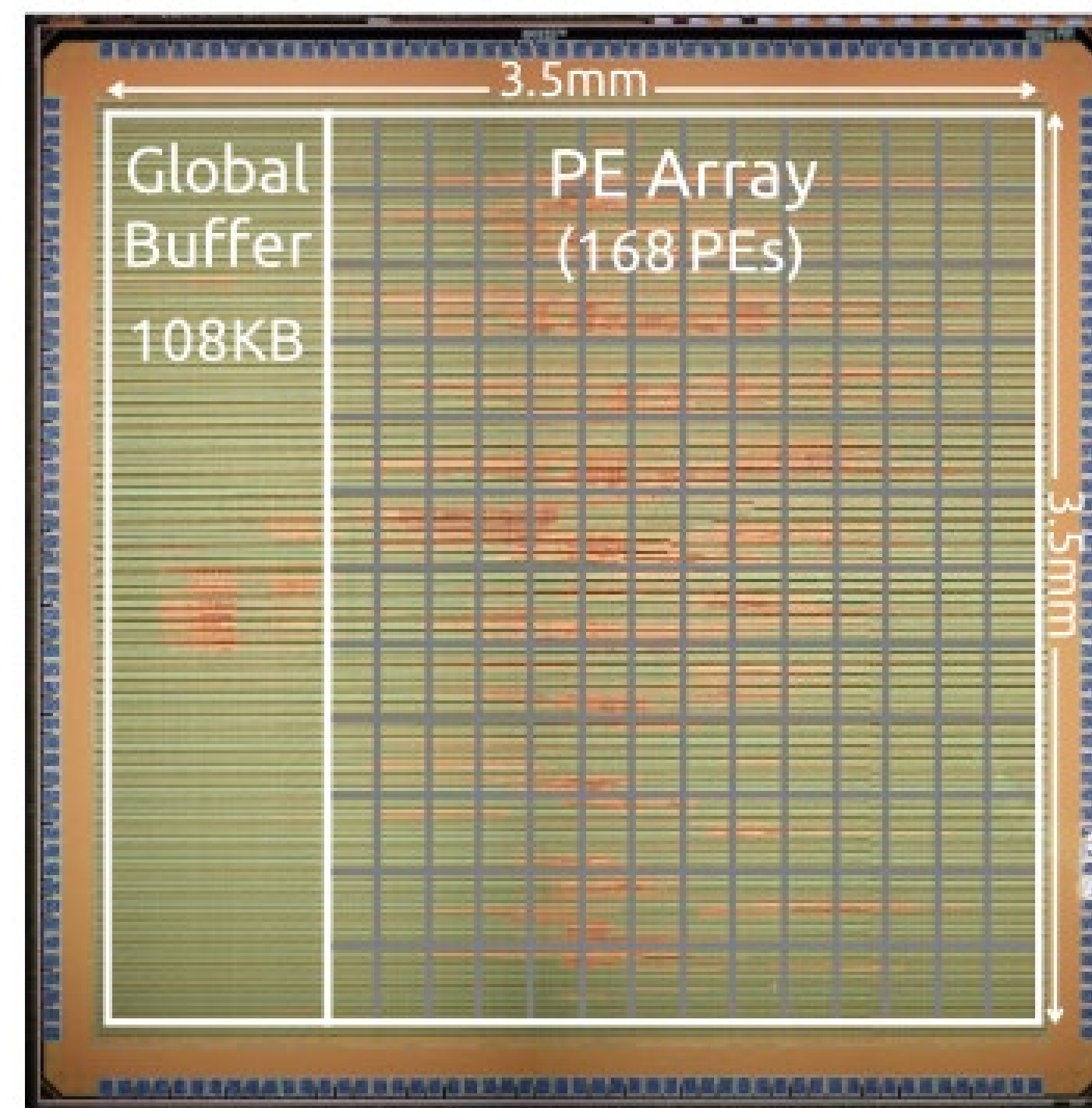
Mishra, Asit, et al. "Accelerating sparse deep neural networks." *arXiv preprint arXiv:2104.08378* (2021)

# Accelerators

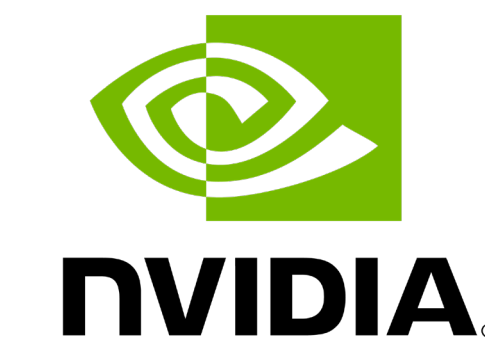
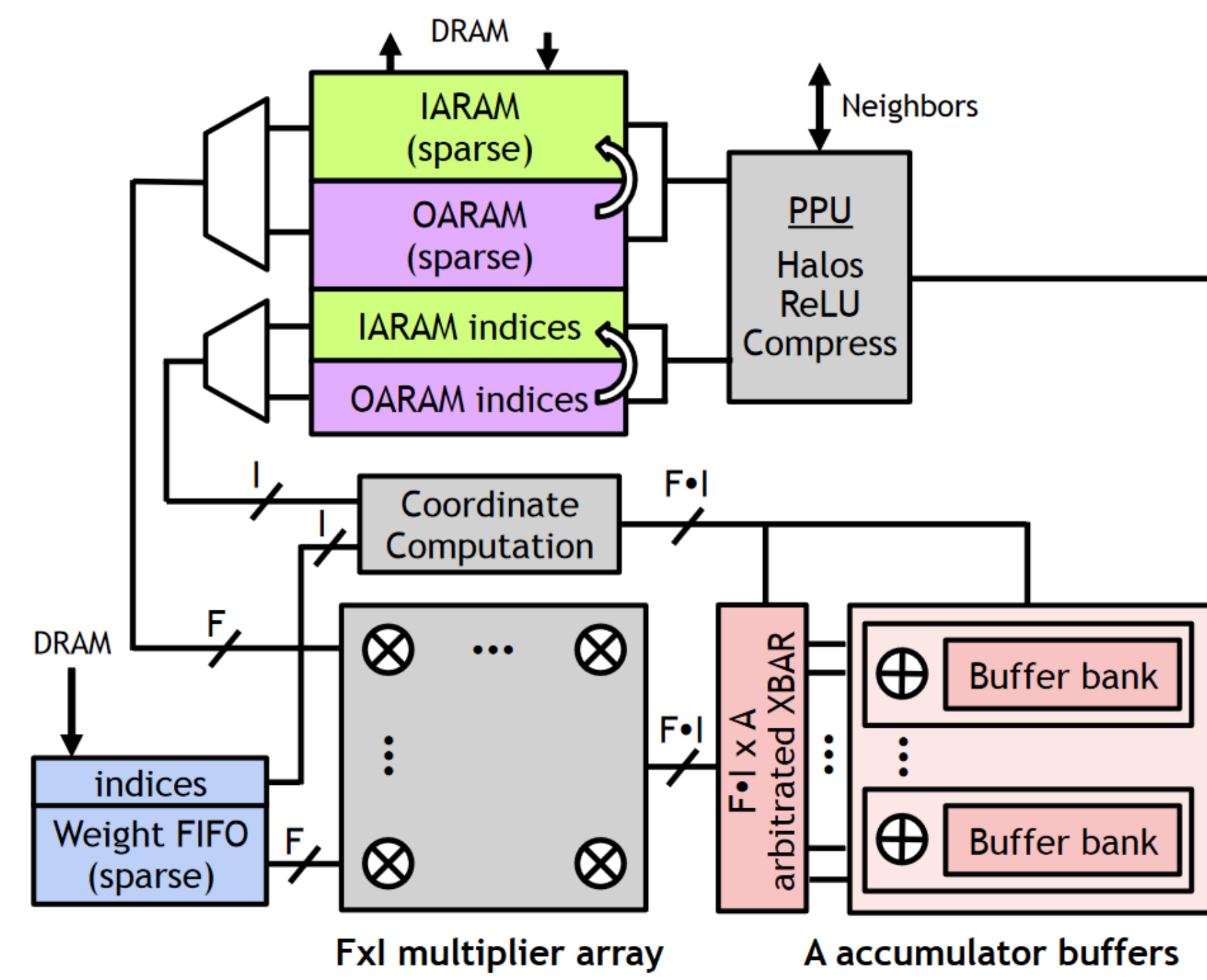
## EIE (2016)



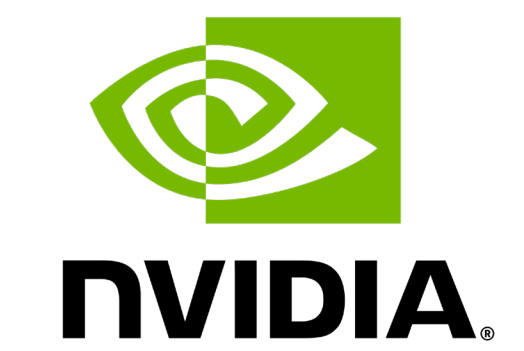
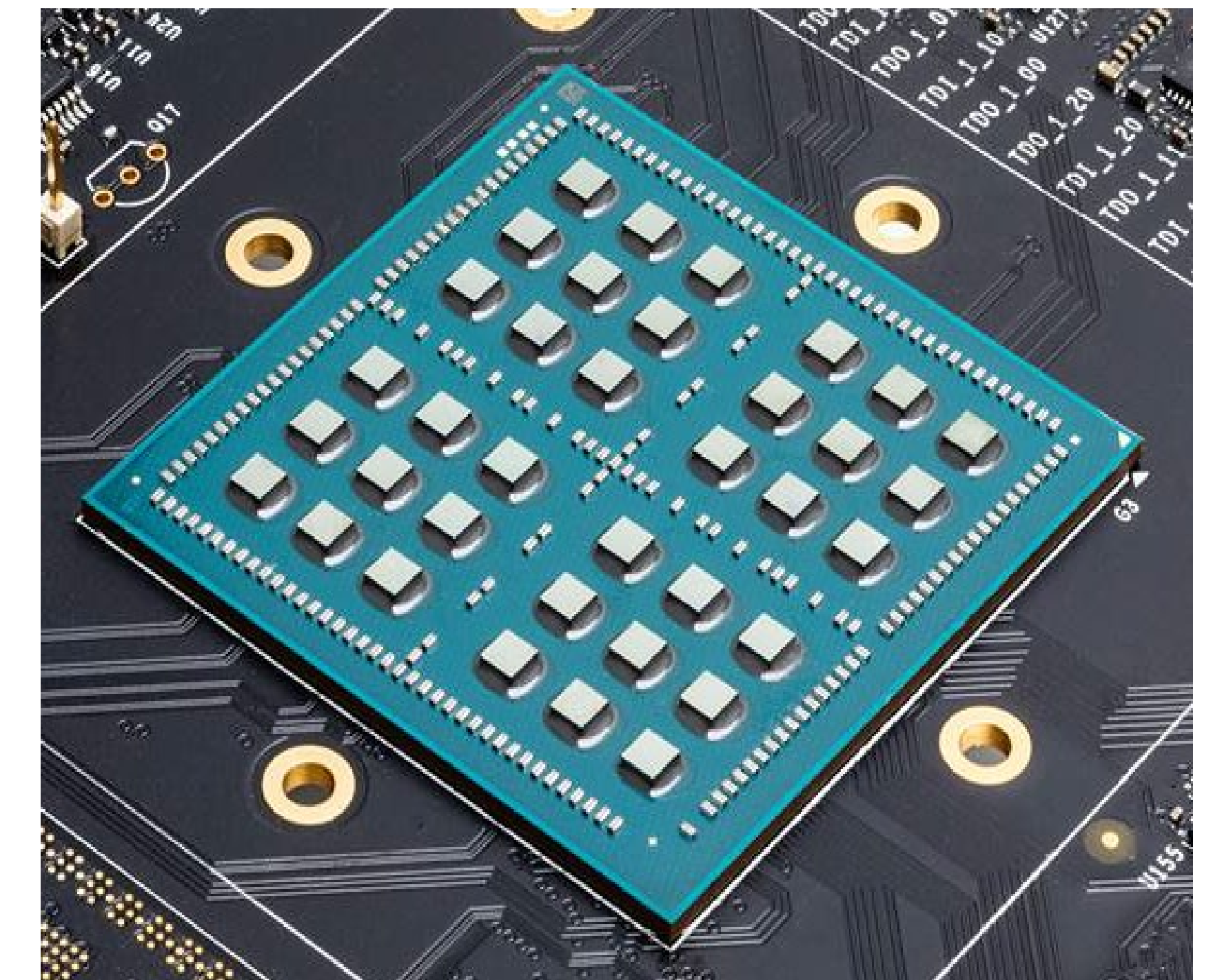
## Eyeriss (2016)



## SCNN (2017)



## Simba (2018)



# Accelerators Employ:

- Special **Data Types** and **Operations**
  - Do in 1 cycle what normally takes 10s or 100s – **10-1000x efficiency gain**
- Massive **Parallelism** – >1,000x, not 16x – with **Locality**
  - This gives performance, not efficiency
- Optimized **Memory**
  - High bandwidth (**and low energy**) for specific data structures and operations
- Reduced or Amortized **Overhead**
  - **10,000x efficiency gain** for simple operations
- Algorithm-Architecture **Co-Design**



# Fast Accelerators since 1985

- **Mossim Simulation Engine**: Dally, W.J. and Bryant, R.E., 1985. A hardware architecture for switch-level simulation. *IEEE Trans. CAD*, 4(3), pp.239-250.
- **MARS Accelerator**: Agrawal, P. and Dally, W.J., 1990. A hardware logic simulation system. *IEEE Trans. CAD*, 9(1), pp.19-29.
- **Reconfigurable Arithmetic Processor**: Fiske, S. and Dally, W.J., 1988. *The reconfigurable arithmetic processor* . ISCA 1988.
- **Imagine**: Kapasi, U.J., Rixner, S., Dally, W.J., Khailany, B., Ahn, J.H., Mattson, P. and Owens, J.D., 2003. Programmable stream processors. *Computer*, 36(8), pp.54-62.
- **ELM**: Dally, W.J., Balfour, J., Black-Shaffer, D., Chen, J., Harting, R.C., Parikh, V., Park, J. and Sheffield, D., 2008. Efficient embedded computing. *Computer*, 41(7).
- **EIE**: Han, S., Liu, X., Mao, H., Pu, J., Pedram, A., Horowitz, M.A. and Dally, W.J., 2016, June. EIE: efficient inference engine on compressed deep neural network, ISCA 2016
- **SCNN**: Parashar, A., Rhu, M., Mukkara, A., Puglielli, A., Venkatesan, R., Khailany, B., Emer, J., Keckler, S.W. and Dally, W.J., 2017, June. Scnn: An accelerator for compressed-sparse convolutional neural networks, ISCA 2017
- **Darwin**: Turakhia, Bejerano, and Dally, “Darwin: A Genomics Co-processor provides up to 15,000× acceleration on long read assembly”, ASPLOS 2018.
- **SATiN**: Zhuo, Rucker, Wang, and Dally, “Hardware for Boolean Satisfiability Inference,”

# Eliminating Instruction Overhead

Area is proportional to energy – all 28nm



16b Int Add, 32fJ

**OOO CPU Instruction – 250pJ (99.99% overhead, ARM A-15)**

# Cost of Operations

Relative Energy Cost

Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

1 10 100 1000 10000

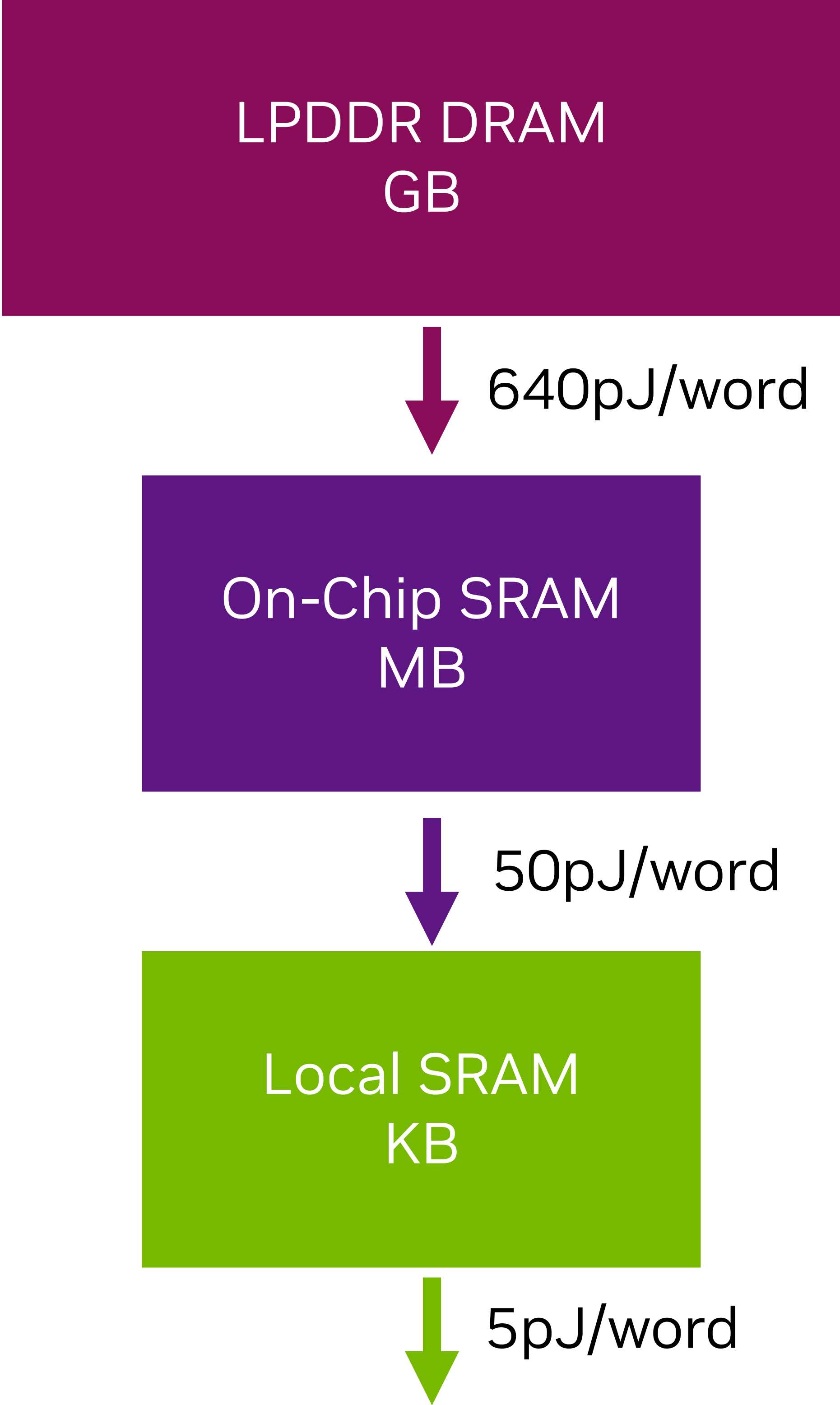
Relative Area Cost

Area ( $\mu\text{m}^2$ )
36
67
137
1360
4184
282
3495
1640
7700
N/A
N/A

1 10 100 1000

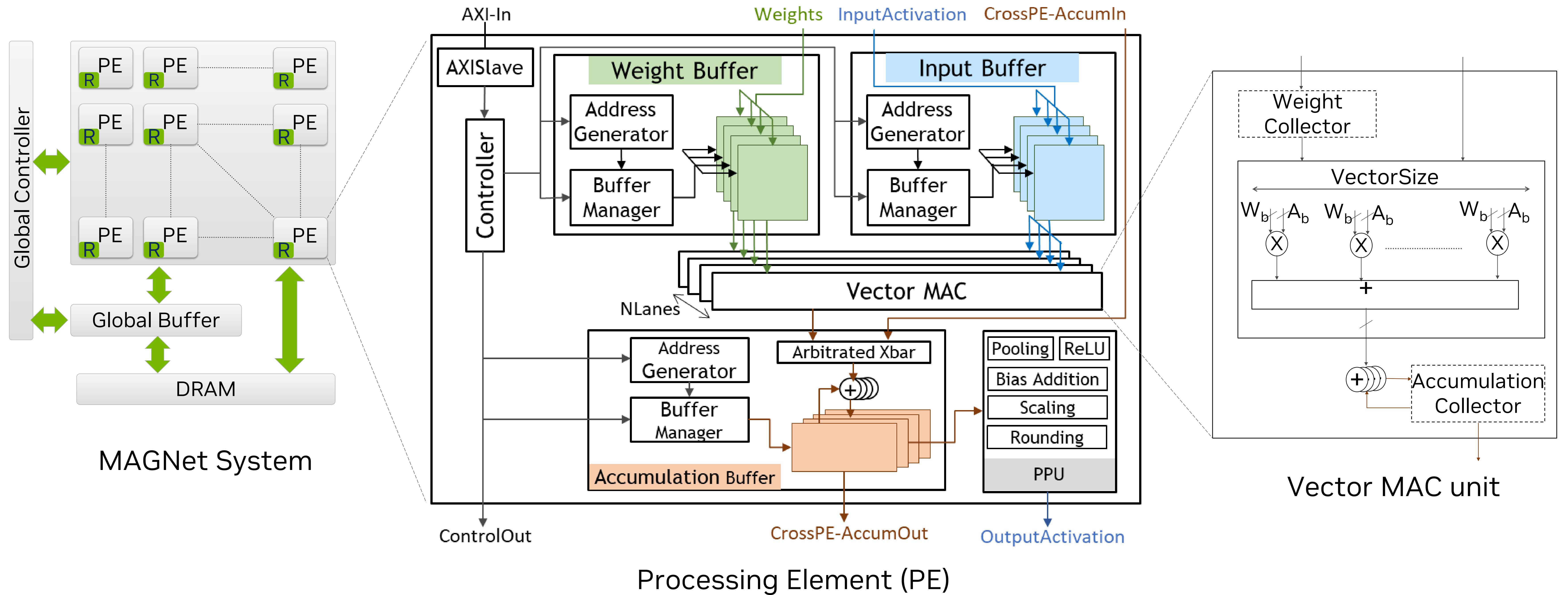
Energy numbers are from Mark Horowitz "Computing's Energy Problem (and what we can do about it)", ISSCC 2014  
 Area numbers are from synthesized result using Design Compiler under TSMC 45nm tech node. FP units used DesignWare Library.

# The Importance of Staying Local



# Magnet

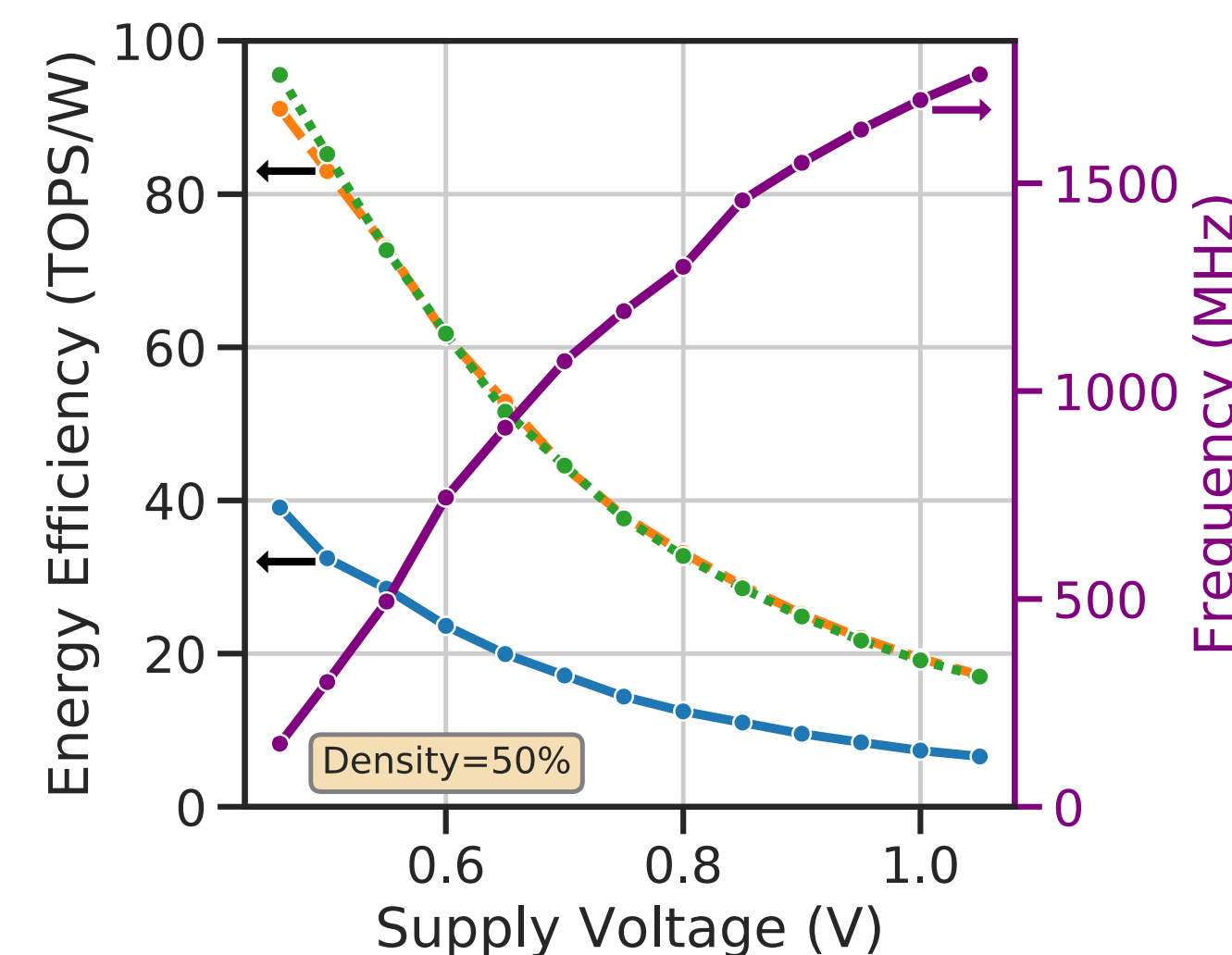
Configurable using synthesizable SystemC, HW generated using HLS tools



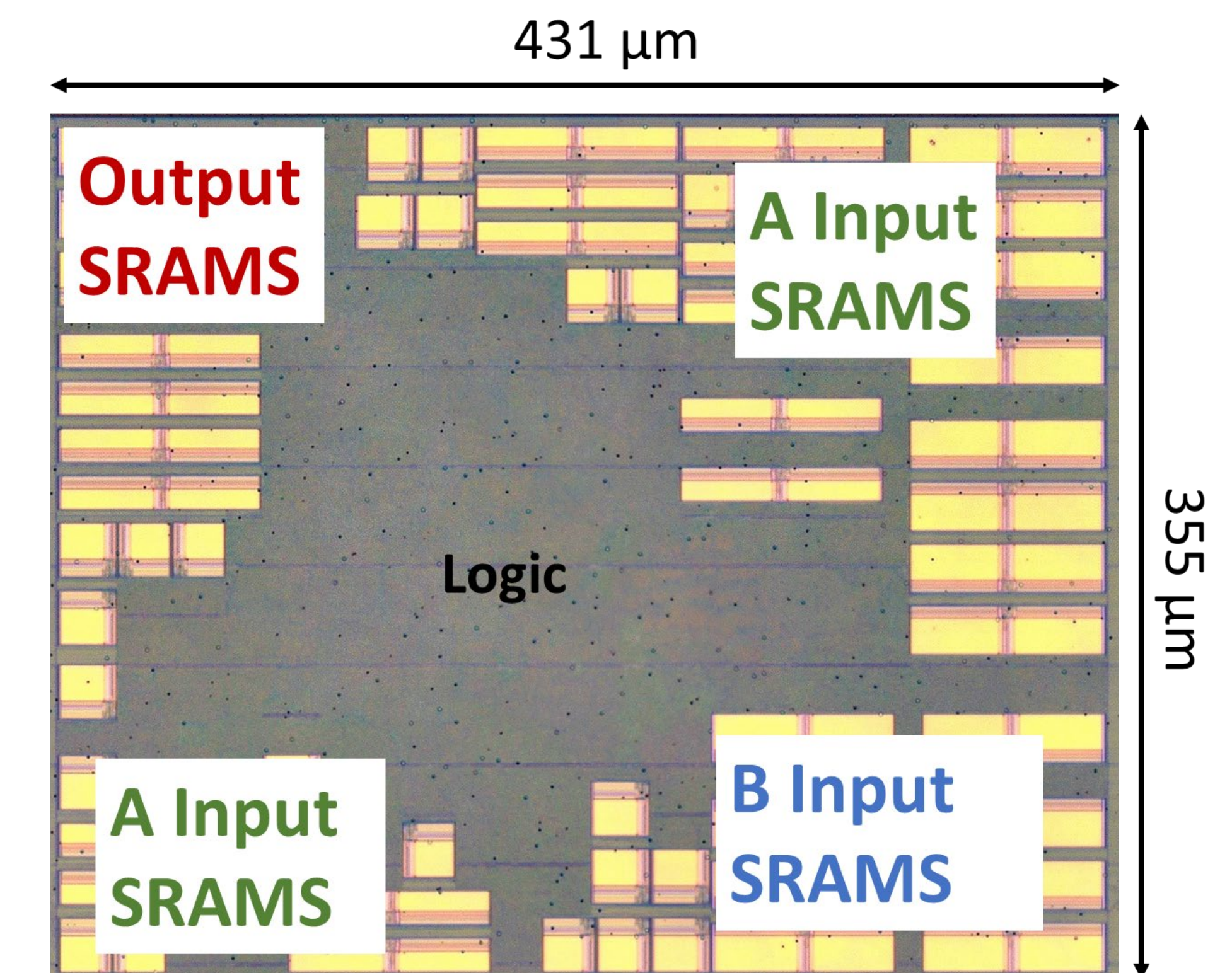
# Energy-efficient DL Inference accelerator

Transformers, VS-Quant INT4, TSMC 5nm

- Efficient architecture
  - Used MAGNet [Venkatesan et al., ICCAD 2019] to design a low-precision DL inference accelerator for Transformers
  - Multi-level dataflow to improve data reuse and energy efficiency
- Low-precision data format: VS-Quant INT4
  - Hardware-software techniques to tolerate quantization error
  - Enable low cost multiply-accumulate (MAC) operations
  - Reduce storage and data movement
- Special function units



- 95.6 TOPS/W with 50%-dense 4-bit input matrices with VSQ enabled at 0.46V
- 0.8% energy overhead from VSQ support with 50%-dense inputs at 0.67V



- TSMC 5nm
- 1024 4-bit MACs/cycle (512 8-bit)
- 0.153 mm<sup>2</sup> chip
- Voltage range: 0.46V – 1.05V
- Frequency range: 152 MHz – 1760 MHz

# Conclusion

# Conclusion

- Deep Learning was enabled by hardware and its progress is limited by hardware
- 1000x in last 10 years
  - Number representation, complex ops, sparsity
- Logarithmic numbers
  - Lowest worst-case error for a given number of bits
  - Can 'factor out' hard parts of an add
- Optimum clipping
  - Minimize MSE by trading quantization noise for clipping noise
- VS-Quant
  - Separate scale factor for each small vector – 16 to 64 scalars
- Accelerators – Testbeds for GPU 'cores'
  - Test chip validates concepts and measures efficiency
  - 95.6 TOPS/W on BERT with negligible accuracy loss

