MXCore: Rethinking Matrix Multiplication Accelerators for Microscaling (MX) Arithmetic

Jayanth Jonnalagadda[1]*

Paper ID: 336

Gamze İslamoğlu[1]*, Arpan Suravi Prasad[1], Angelo Garofalo[2], Francesco Conti[2], and Luca Benini[1,2]

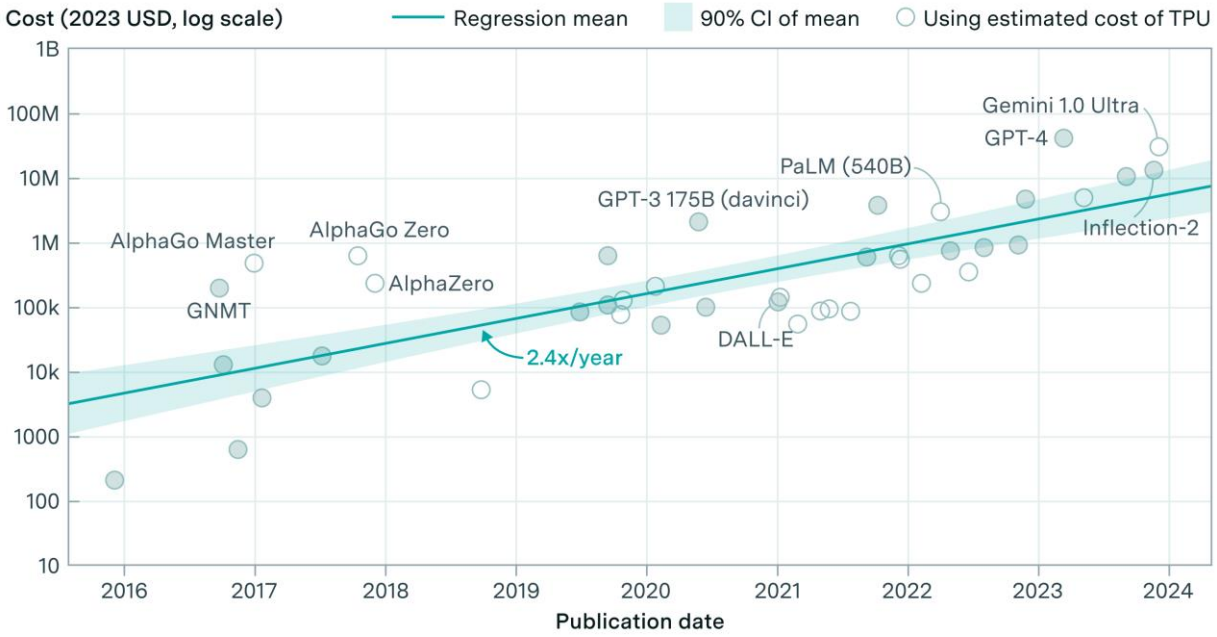
[1]Integrated Systems Laboratory (IIS), ETH Zürich, Switzerland

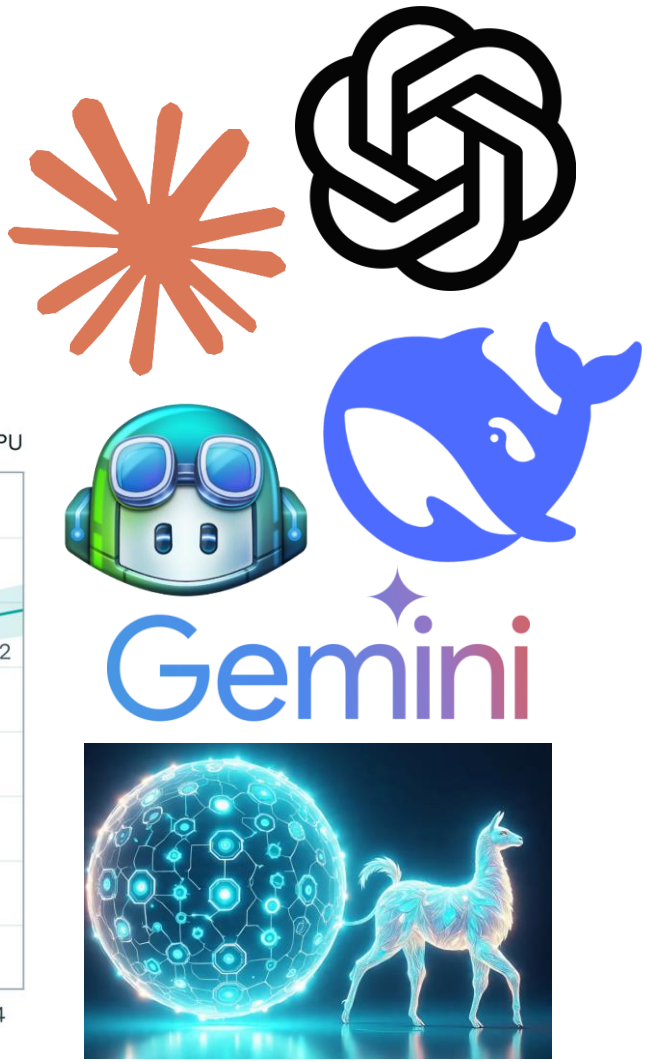
[2]DEI, University of Bologna, Italy

*Equal Contribution

Rapidly Rising Costs of Frontier AI

- Frontier AI models are rapidly growing.
  - Billions of parameters, peta-operation-level compute, and hundreds of gigabytes of memory.
  - Steep training cost increase of 2.4× per year.
  - Expected to cross a billion dollars by 2027.



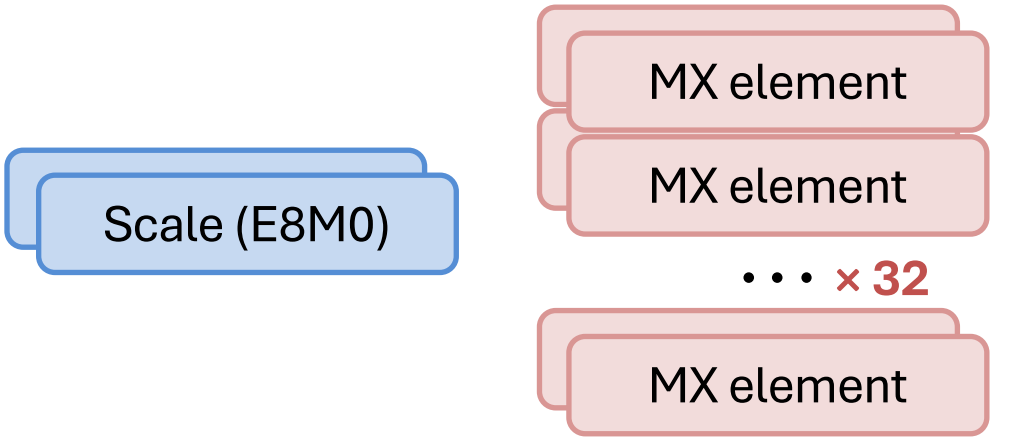


Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, Tamay Besiroglu, and David Owen. 2025. The rising costs of training frontier AI models. doi:10. 48550/arXiv.2405.21015.

ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Low-Precision AI is Promising

- Microscaling (MX) Data Formats.
  - Open Compute Project (2023).
  - Narrow bit-width (4 – 8 bit) formats.
    - MXFP4, MXFP6, MXFP8, and INT8.
  - Block-level scaling.
    - Wide dynamic range and fine-grained control.
  - Accuracies close to full precision.

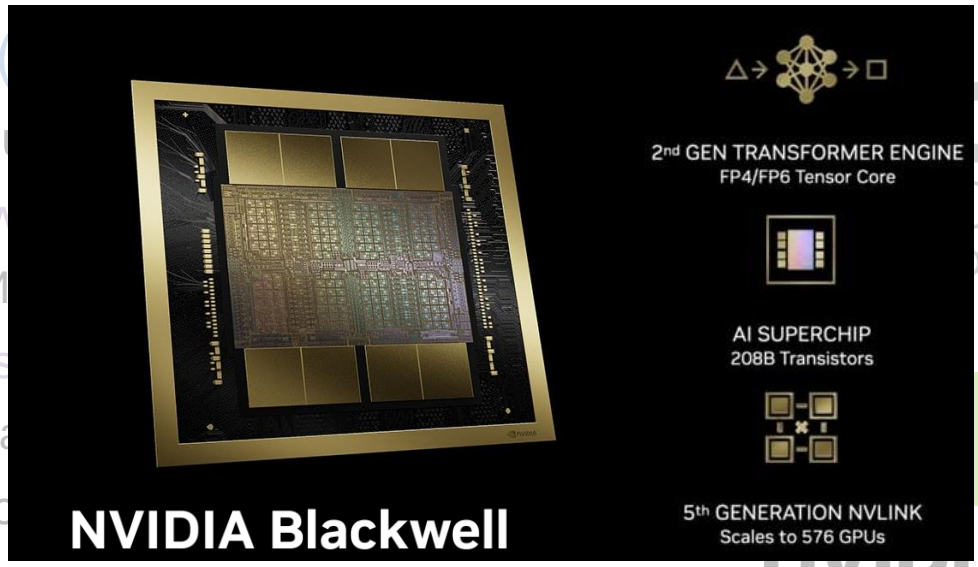


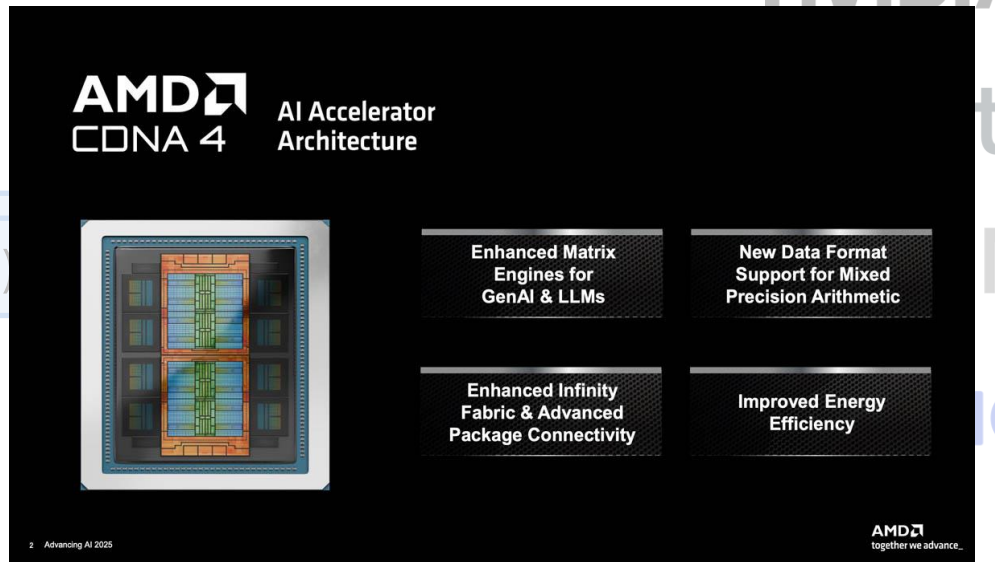
https://www.opencompute.org/documents/ocp-microscaling-formats-mx-v1-0-spec-final-pdf

Industry-leading MX Architectures

- Microscaling (
  - Open Compu
  - Narrow bit-w
    - MXFP4, M
  - Block-level s
    - Wide dyna
  - Accuracies c

Scale (E8M0)

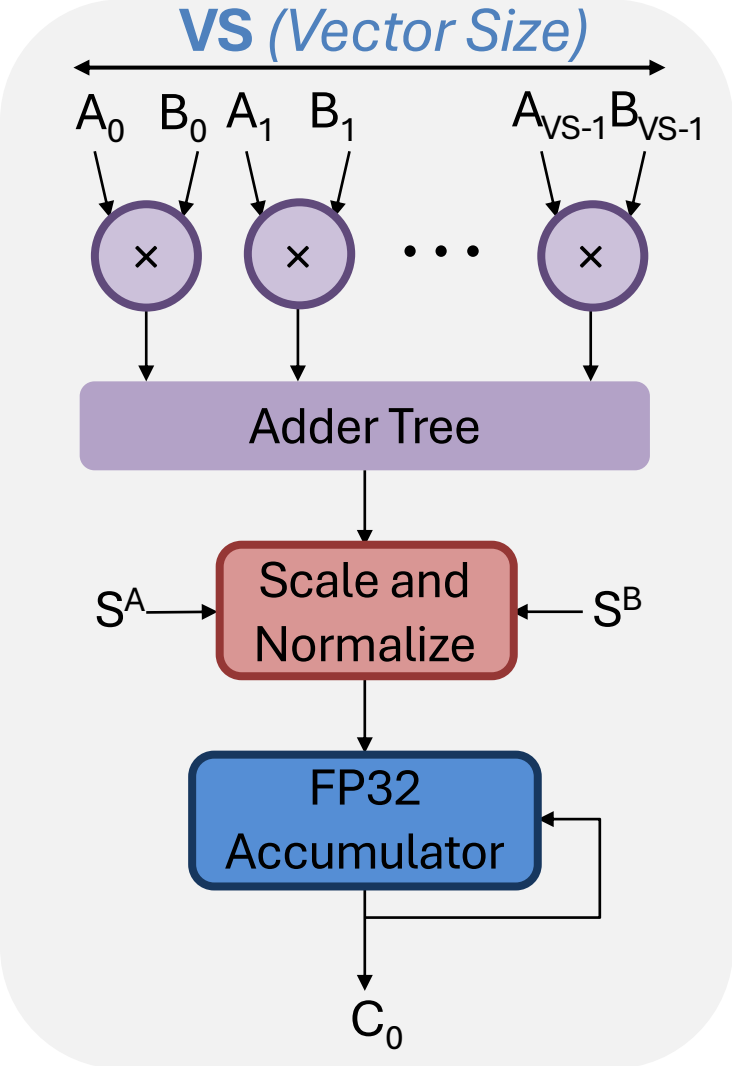




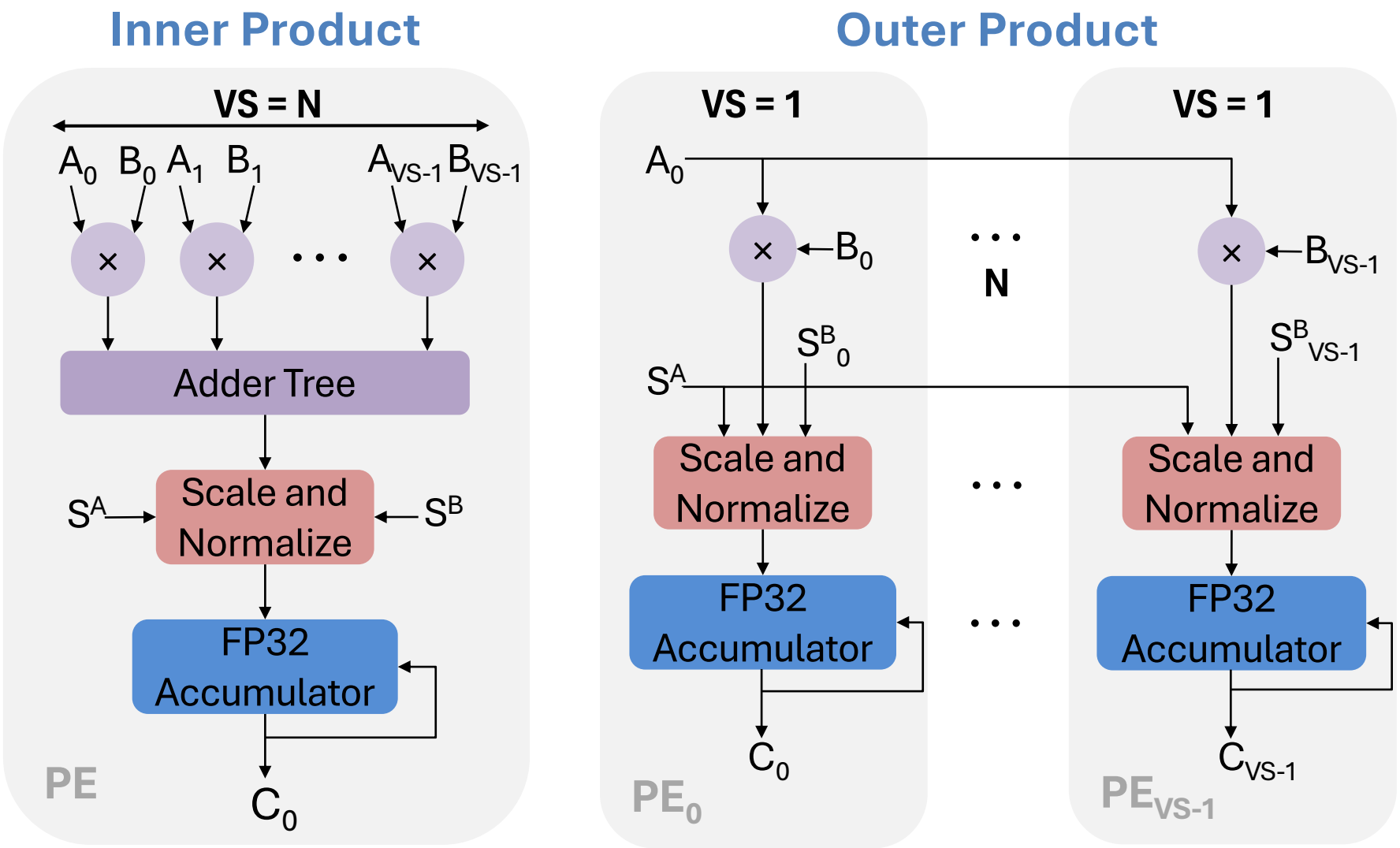
https://www.opencompute.org/documents/ocp-microscaling-formats-mx-v1-0-spec-final-pdf

Unique Aspects of MX Arithmetic

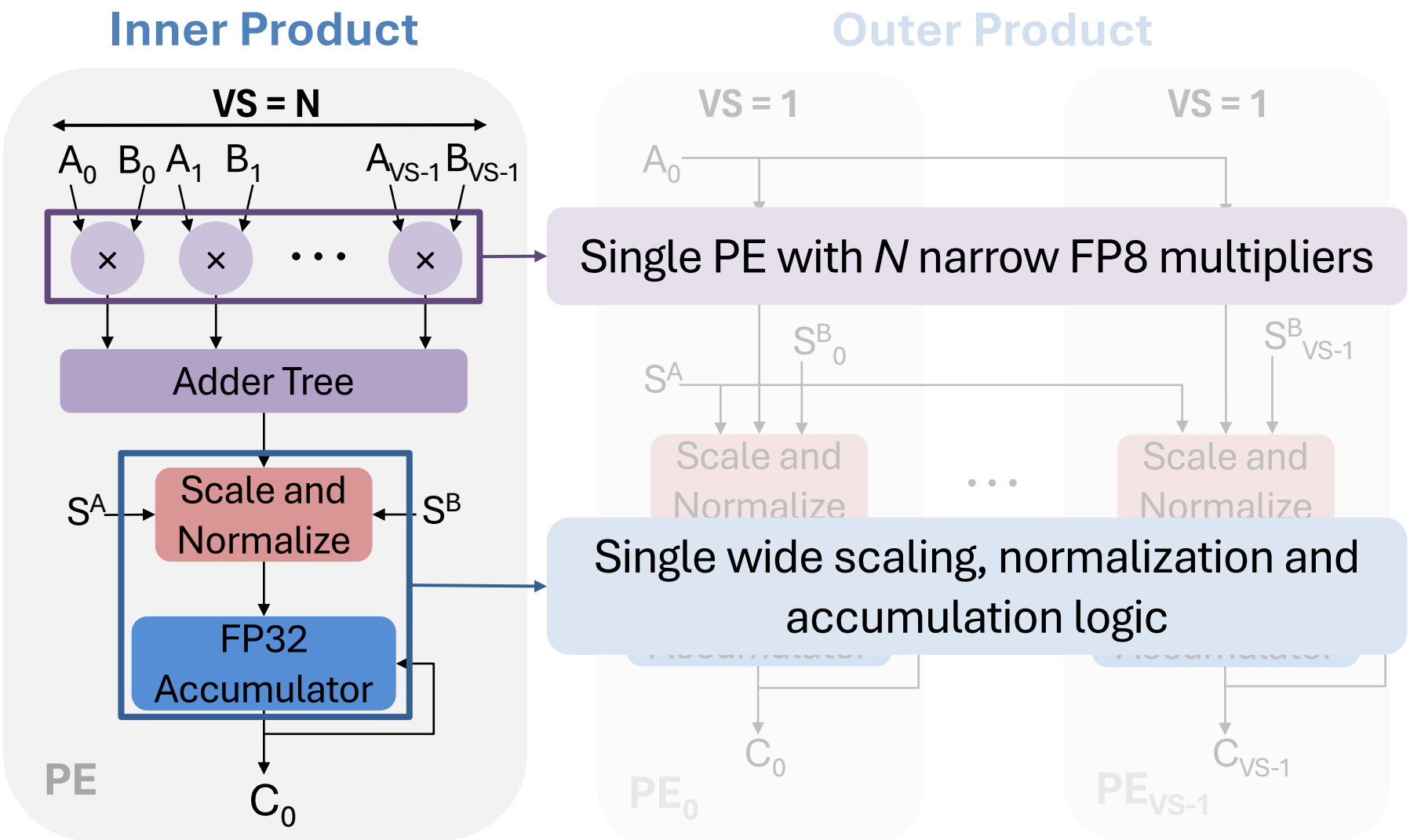
- Operand Precision Asymmetry.
  - Narrow multipliers.
    - 4 – 8 bit operands.
  - Wide accumulations.
    - Accumulation in FP32.
- Block-level Scaling.
  - Intra-block: directly accumulate.
  - Inter-block: scale before accumulate.



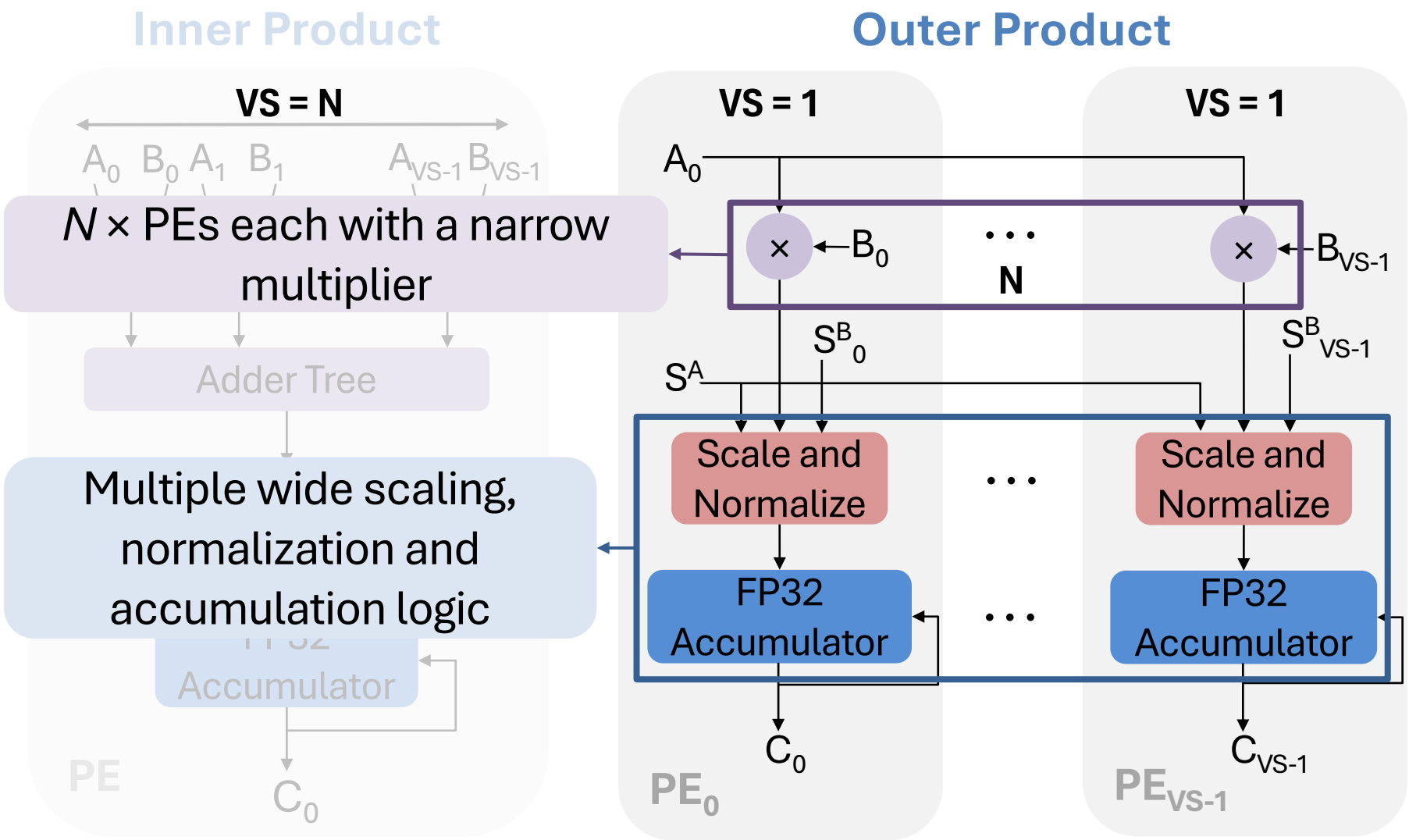
Inner vs. Outer Product for MX



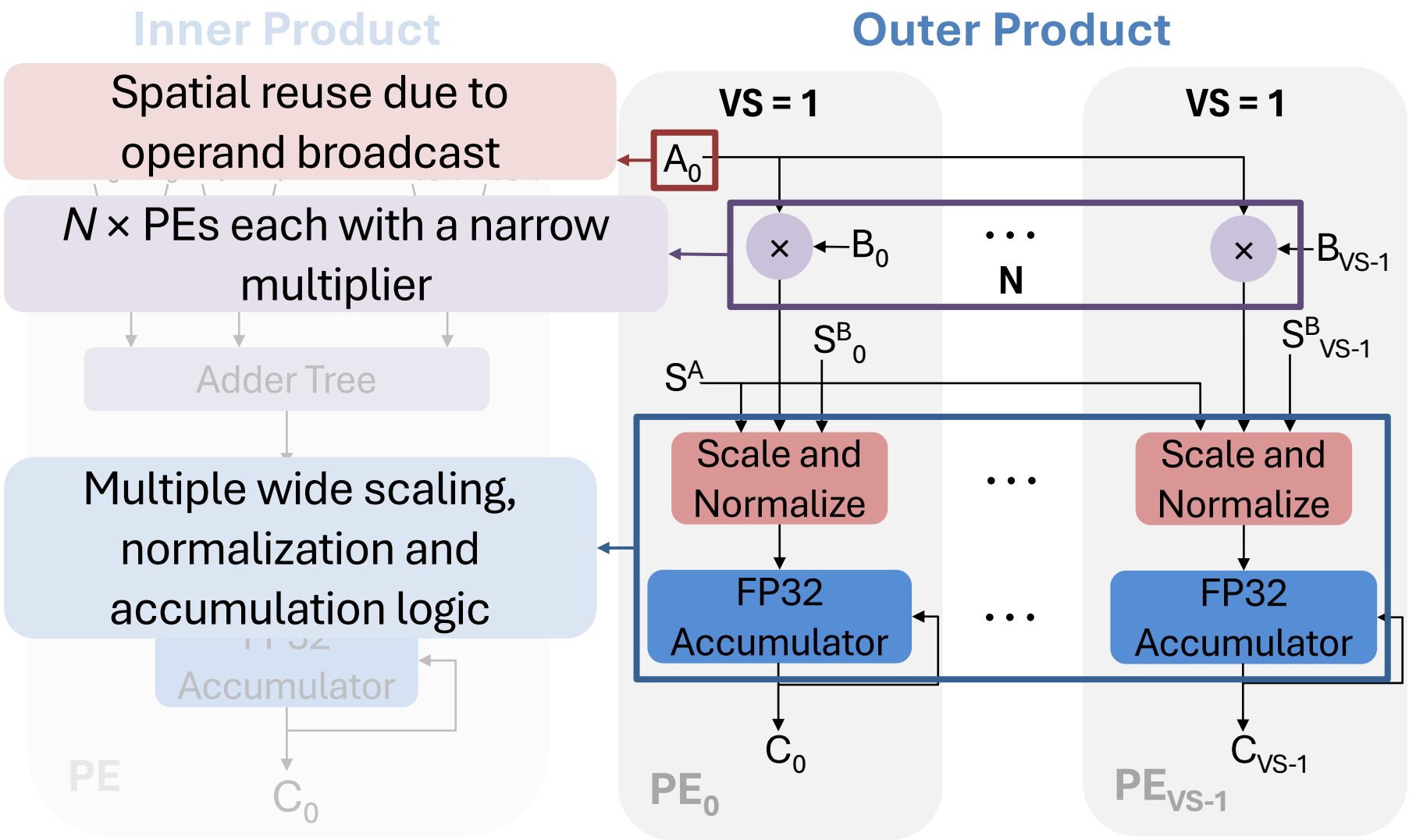
Inner vs. Outer Product for MX



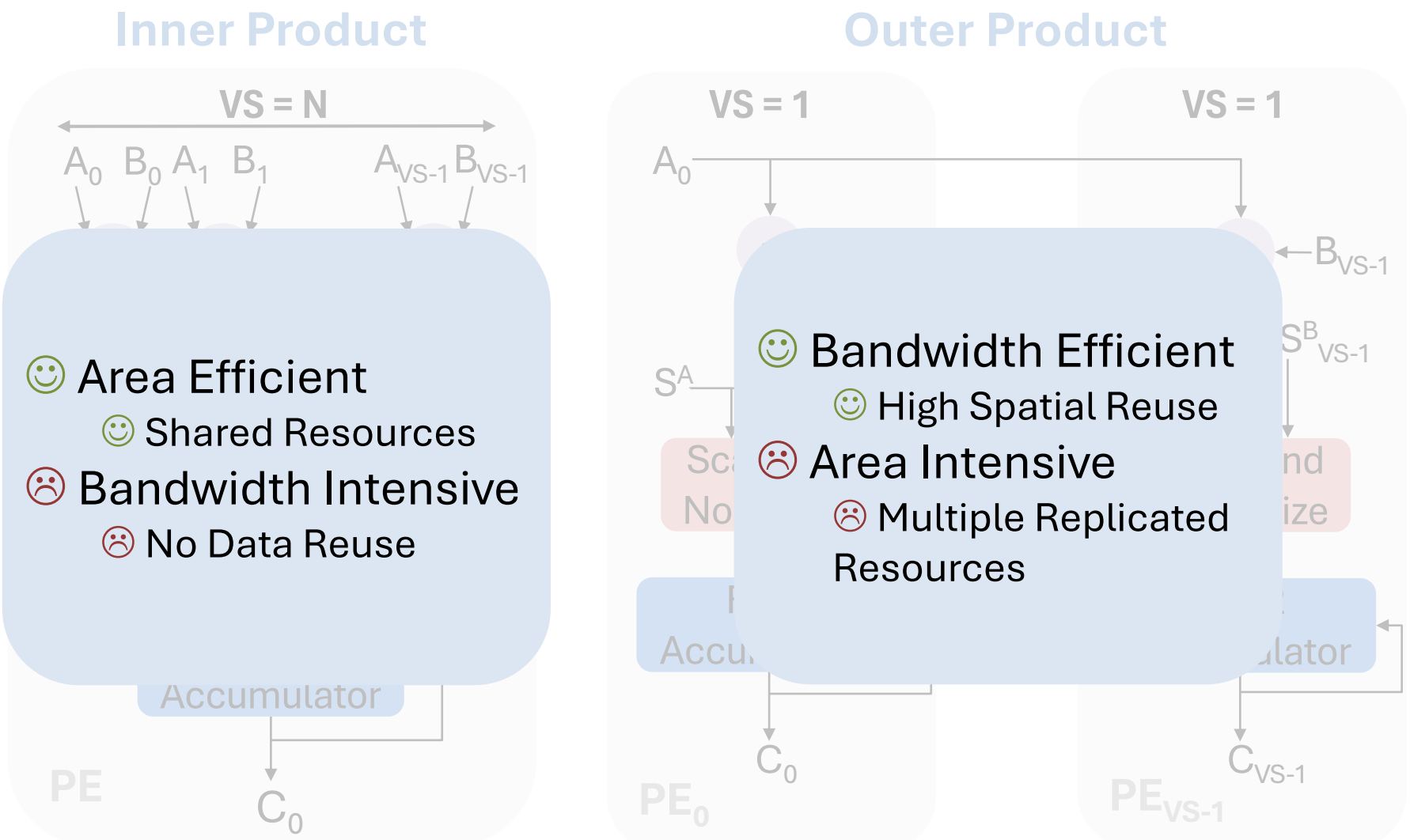
Inner vs. Outer Product for MX



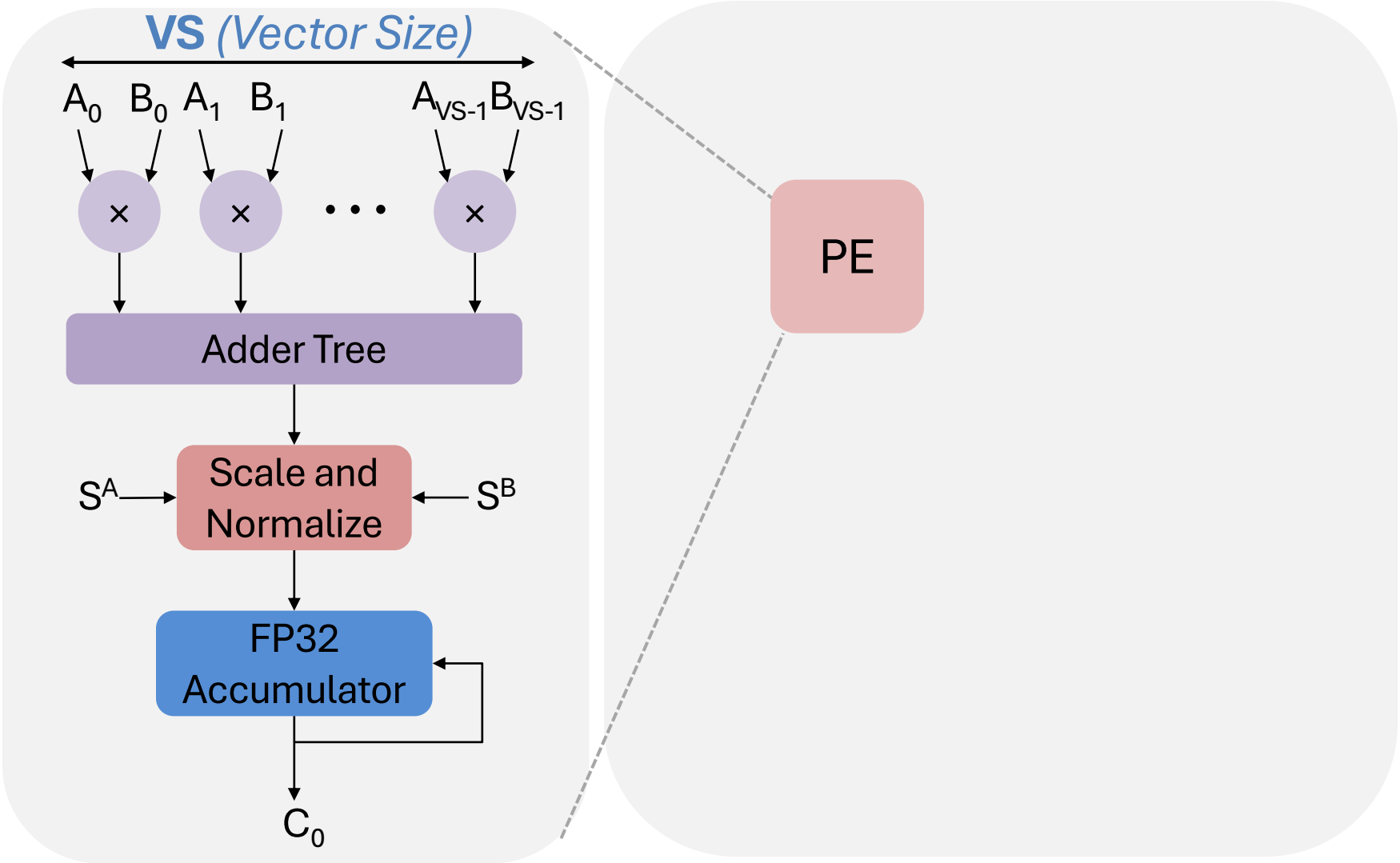
Inner vs. Outer Product for MX



Inner vs. Outer Product Trade-offs

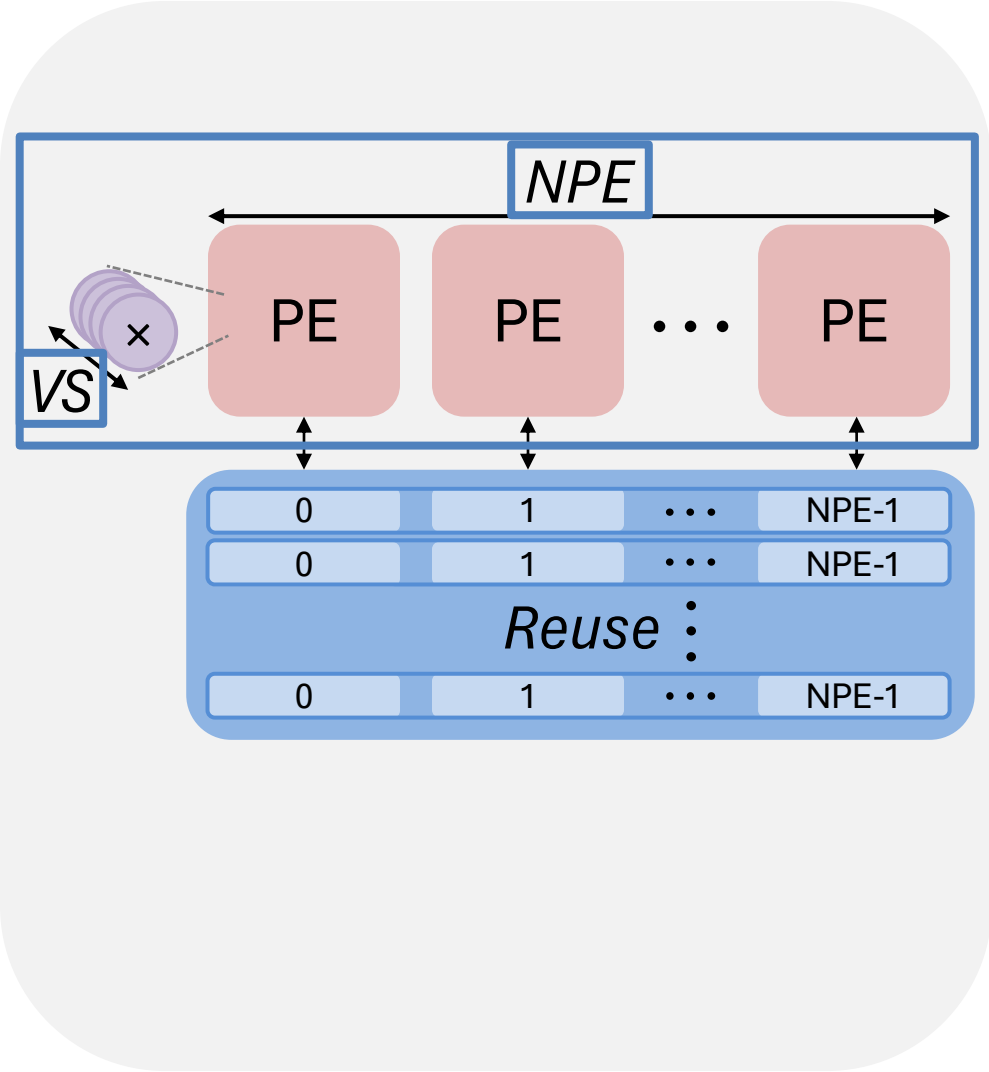


Architectural Framework of MXCore

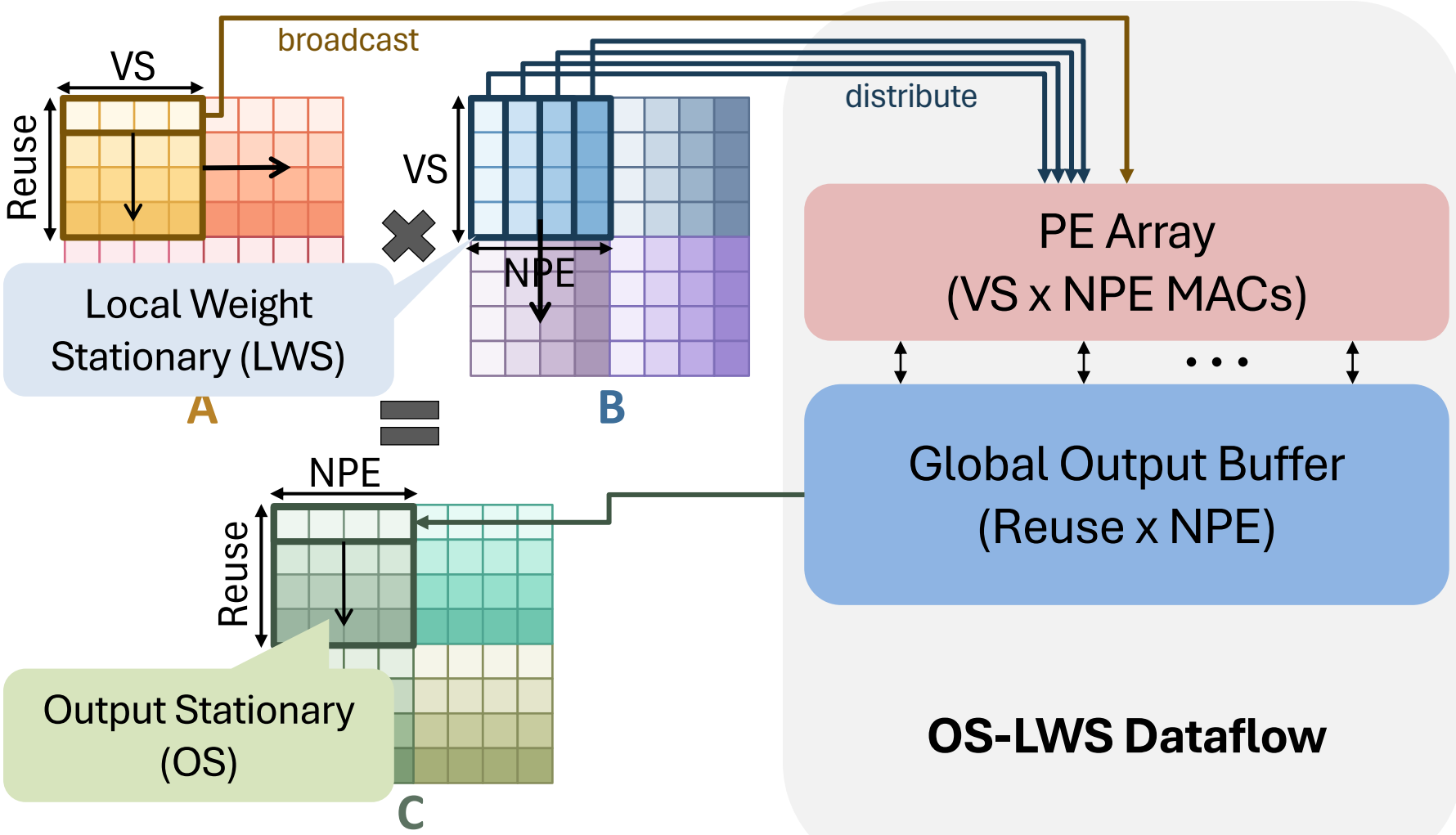


Architectural Framework of MXCore

- VS (Vector Size)
  - Inner product size.
  - Dot product length per PE.
- NPE (Number of PEs)
  - Outer product size.
  - Number of PEs in the array.
- Reuse
  - Output reuse factor.
  - Locally accumulated FP32 partial results.
- Peak Throughput
  - VS x NPE MACs/cycle.



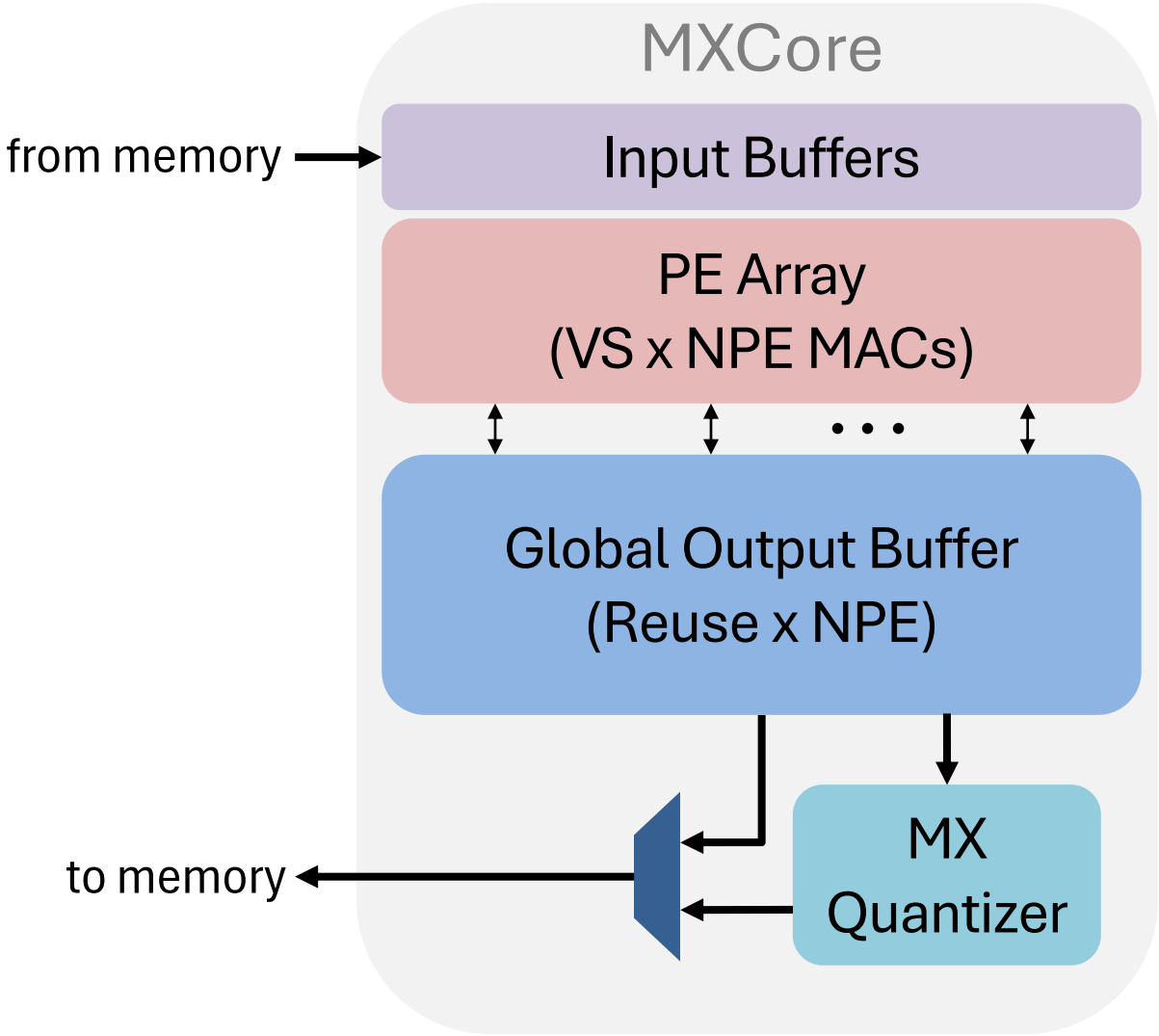
Workload Mapping and Dataflow



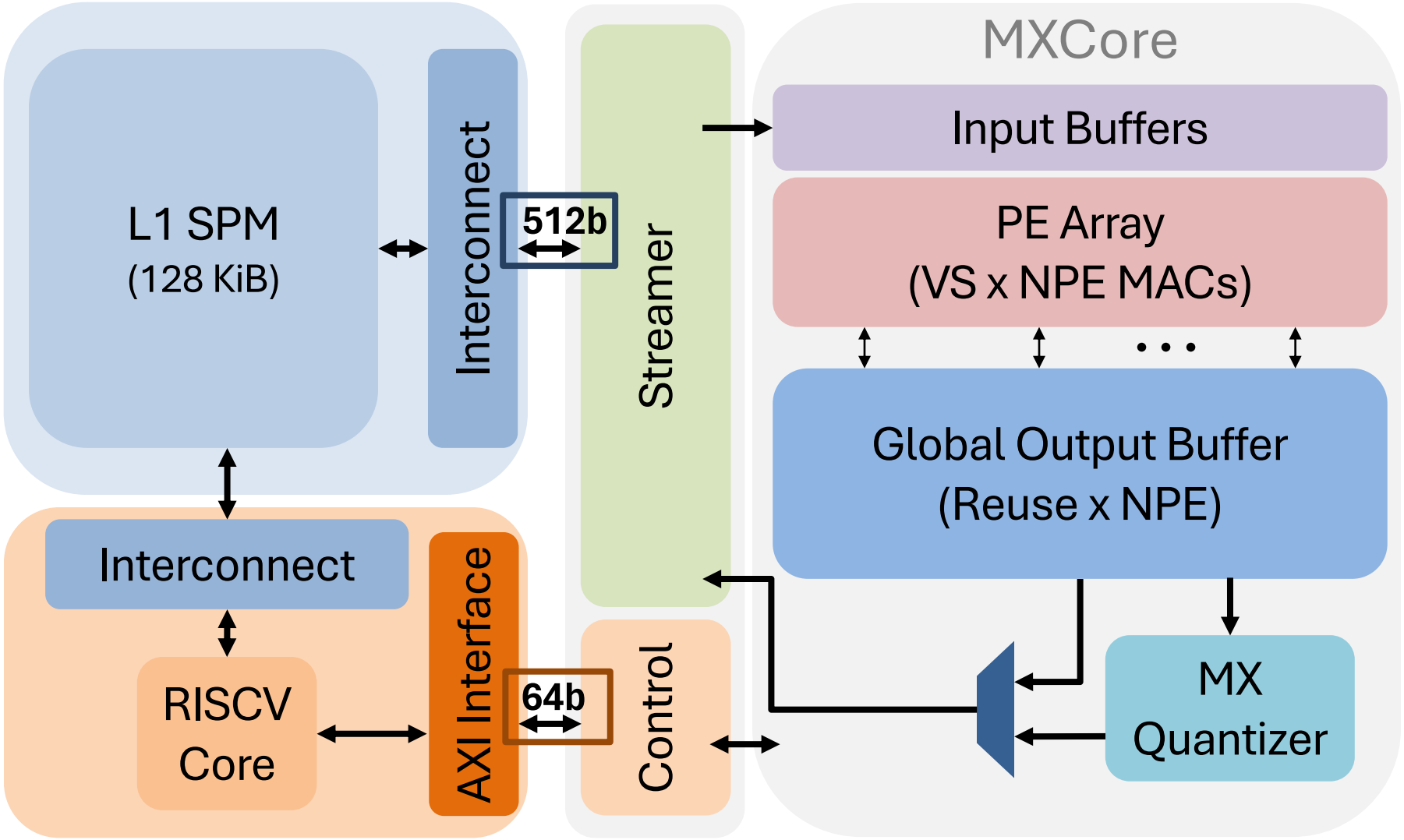
MXFP8 Matrix-Matrix Multiplication
(Scale matrices omitted for brevity)

ETH zürich
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Architecture of MXCore

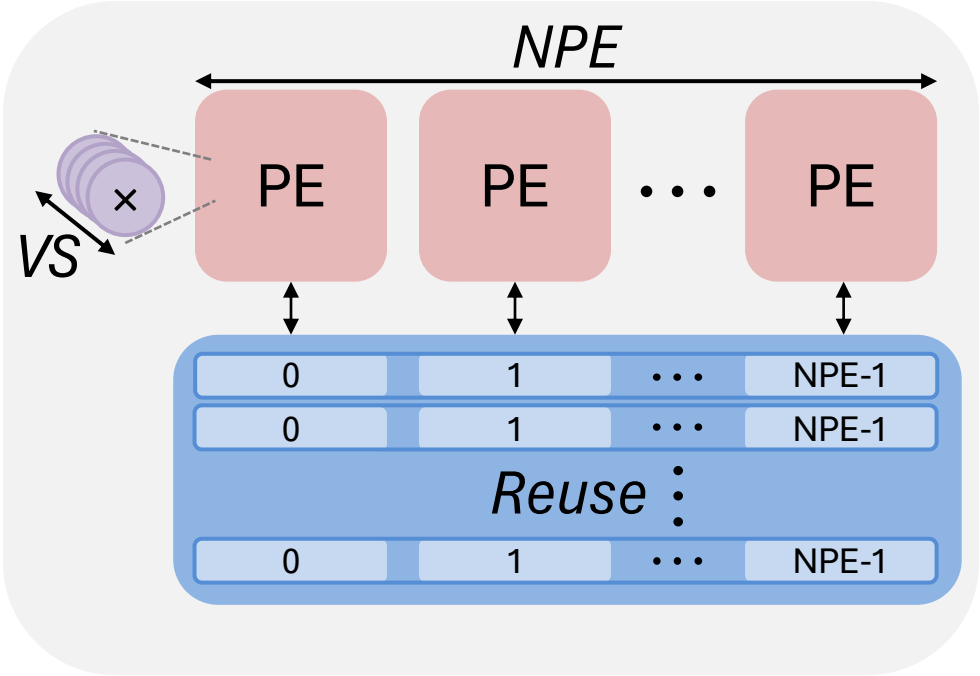


System-level Integration of MXCore

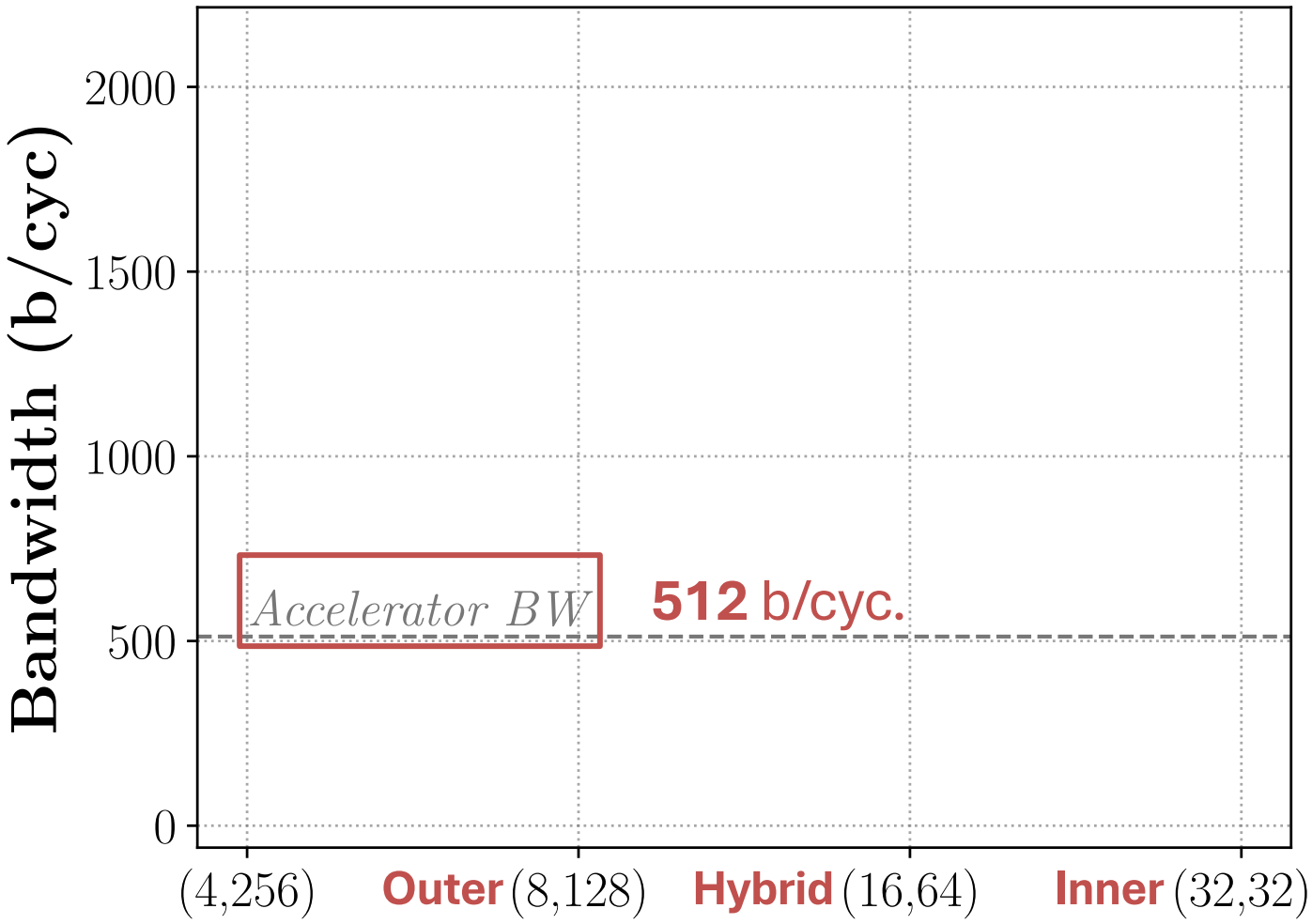


Target Configurations and Parameters

- Iso-Peak-Throughput
  - Fixed #MACs = 1024.
    - VS x NPE = 1024 MACs.
    - Representative of accelerators targeting edge- to mid-range inference platforms.
- Equi-boundary-bandwidth
  - Accelerator BW = 512 b/cyc.
- (VS, NPE)
  - (32, 32): Inner
  - (16, 64): Hybrid
  - (8, 128): Outer
- Reuse ∈ {32, 64, 128}

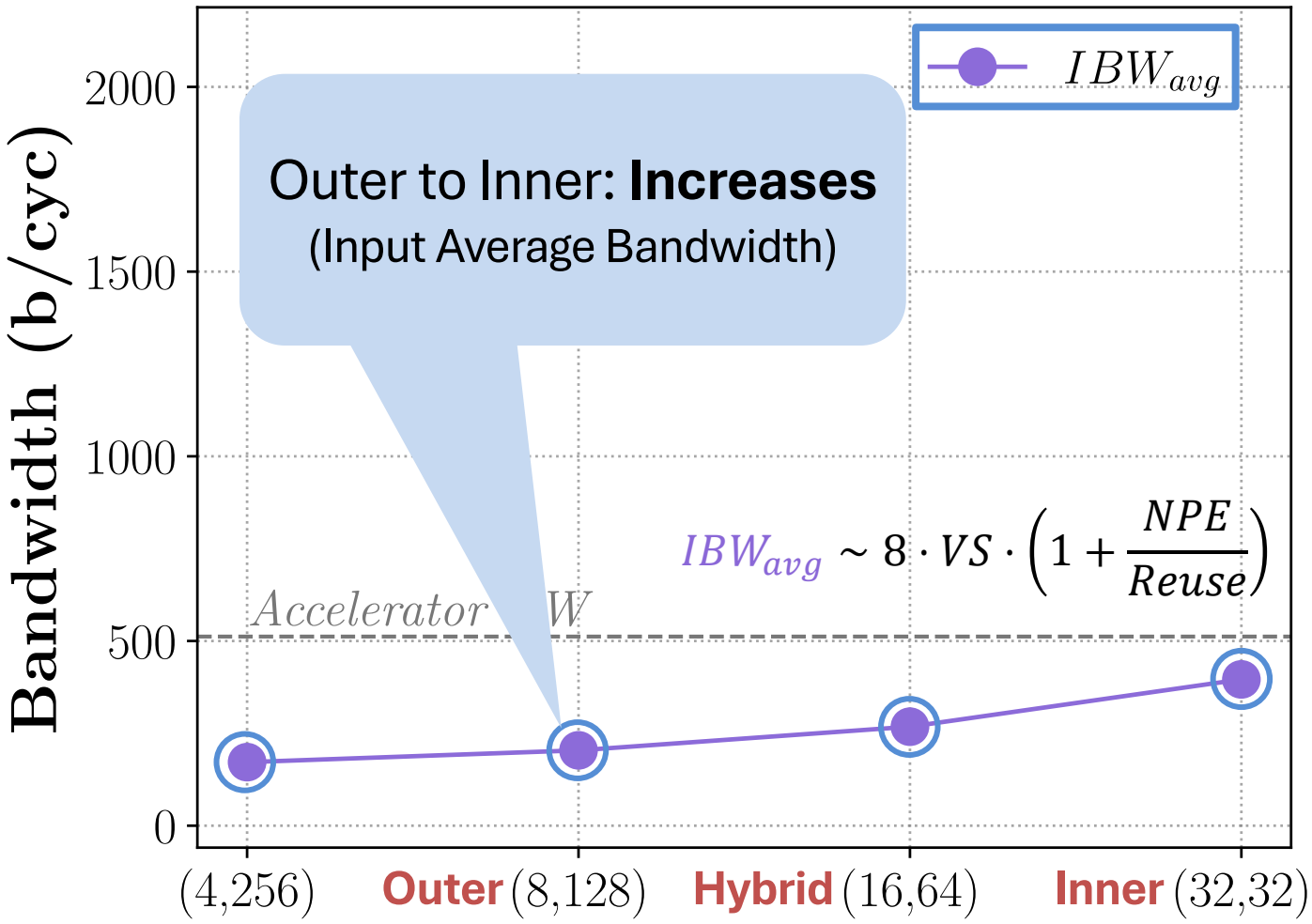


Bandwidth: Inner vs Outer



(Inner, Outer) sizes of MXCore: Inner x Outer = 1024 MACs

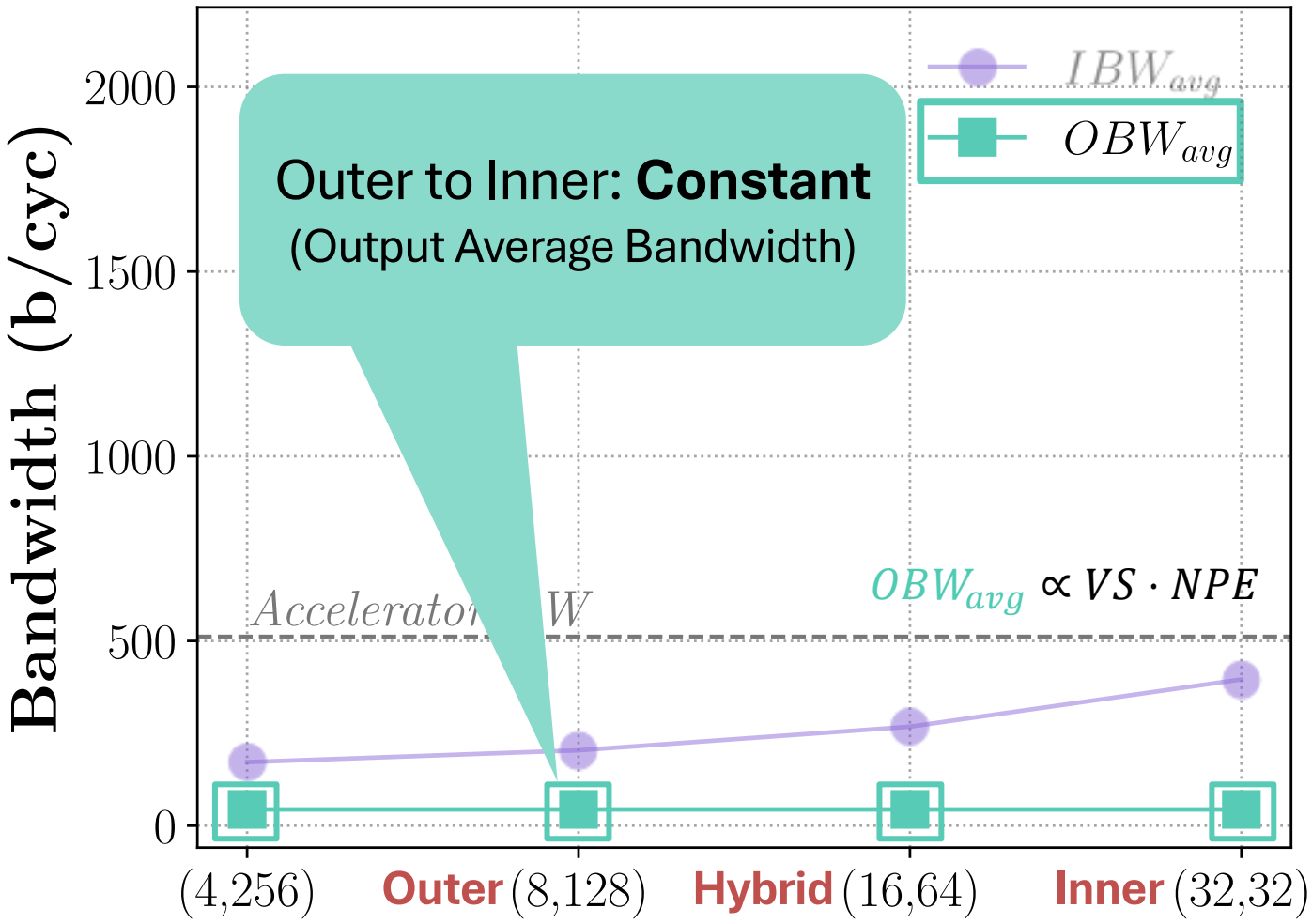
Bandwidth: Inner vs Outer



(Inner, Outer) sizes of MXCore: Inner x Outer = 1024 MACs

***Reuse* fixed at 64**

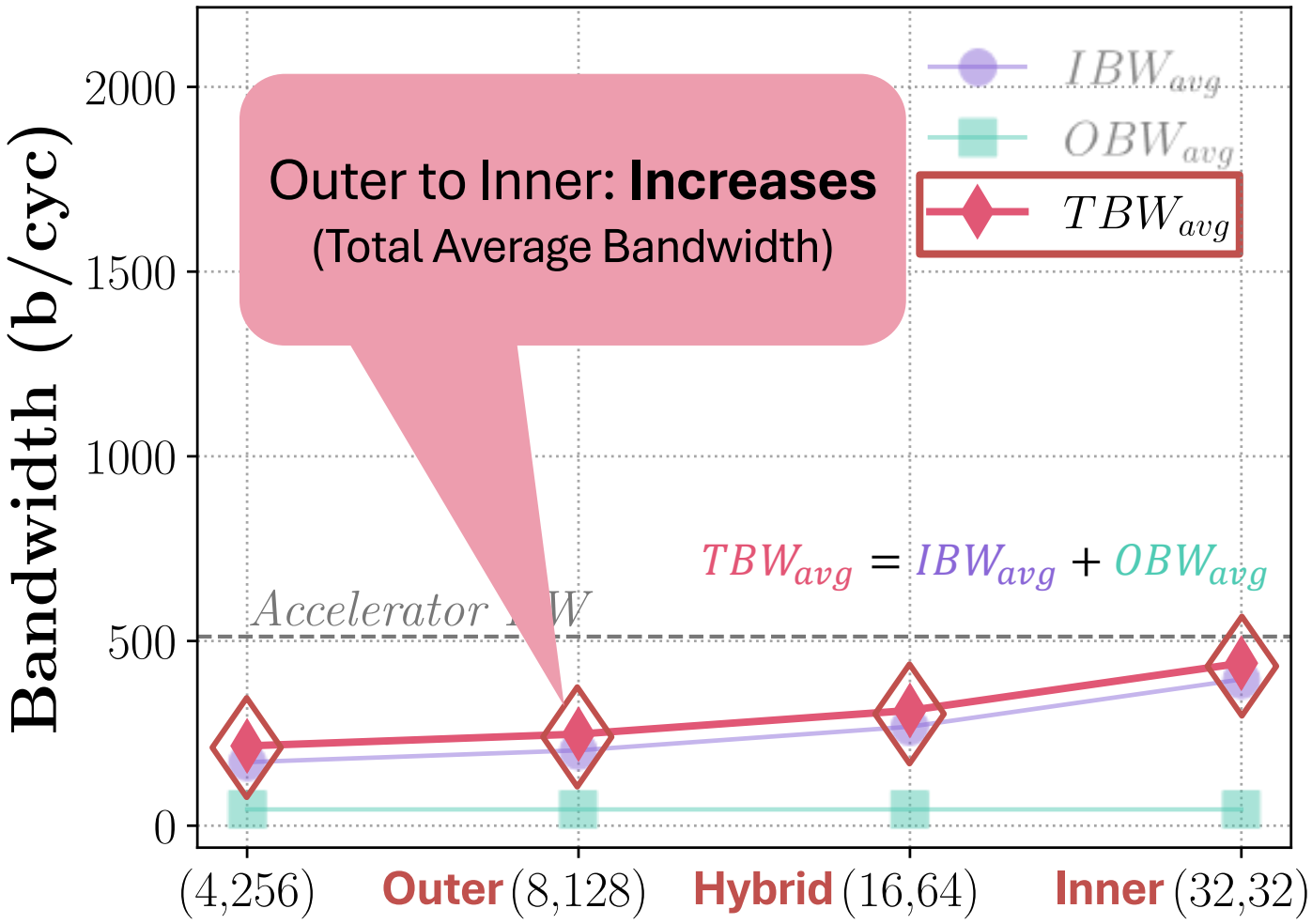
Bandwidth: Inner vs Outer



(Inner, Outer) sizes of MXCore: Inner x Outer = 1024 MACs

***Reuse* fixed at 64**

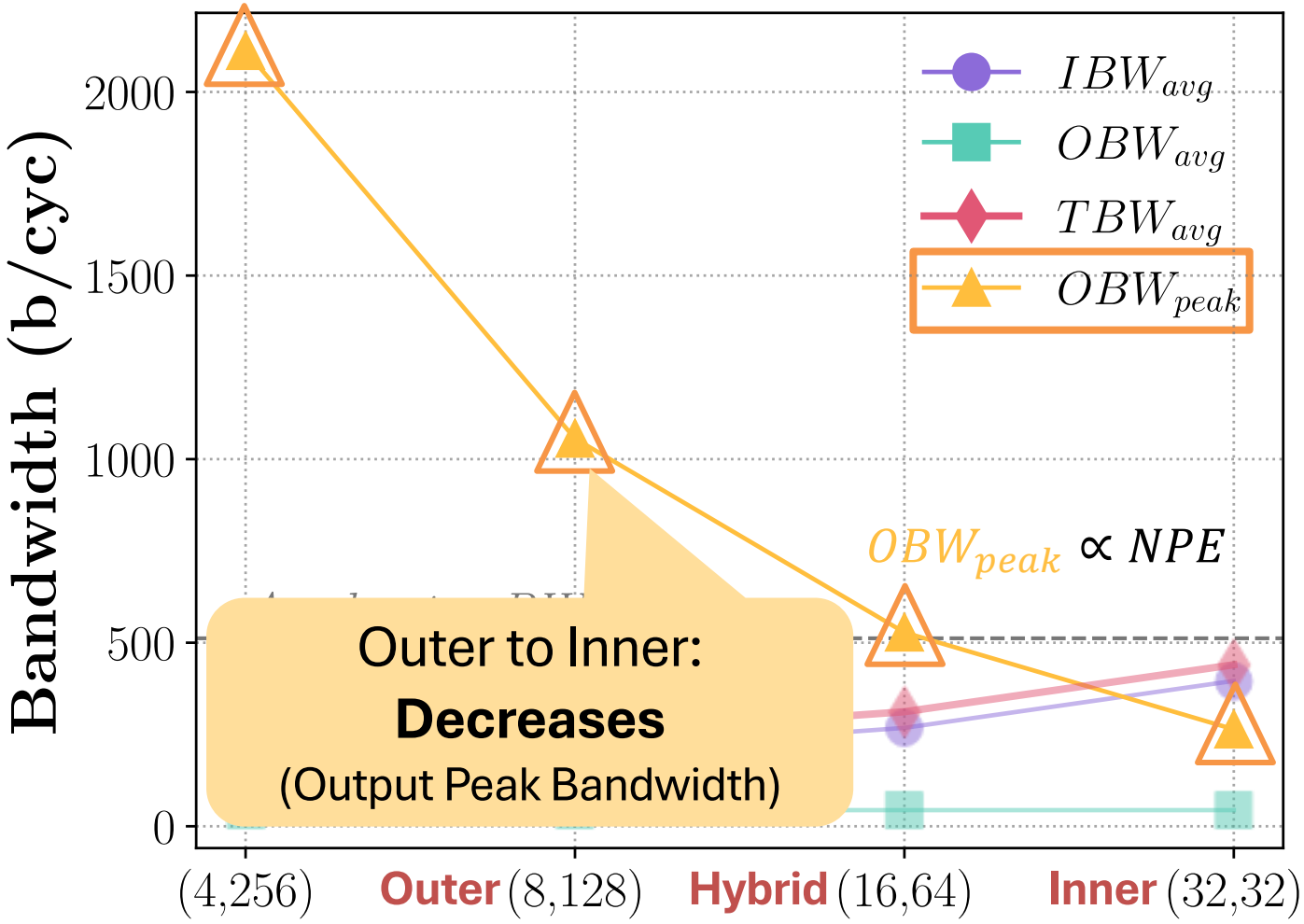
Bandwidth: Inner vs Outer



(Inner, Outer) sizes of MXCore: Inner x Outer = 1024 MACs

***Reuse* fixed at 64**

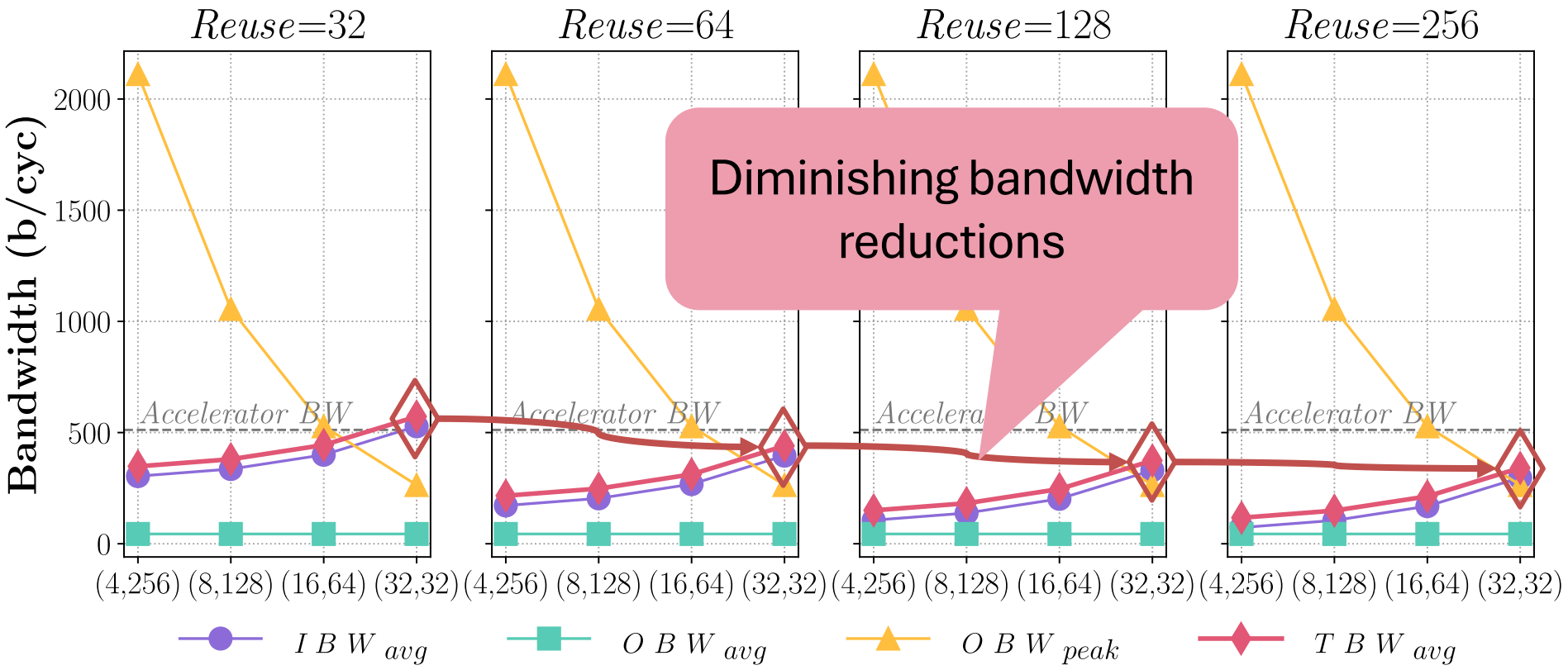
Bandwidth: Inner vs Outer



(Inner, Outer) sizes of MXCore: Inner x Outer = 1024 MACs

***Reuse* fixed at 64**

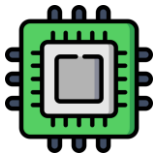
Bandwidth: Across Reuse



(Inner, Outer) sizes of MXCore: Inner x Outer = 1024 MACs

Across different *Reuse*

Physical Implementation of MXCore



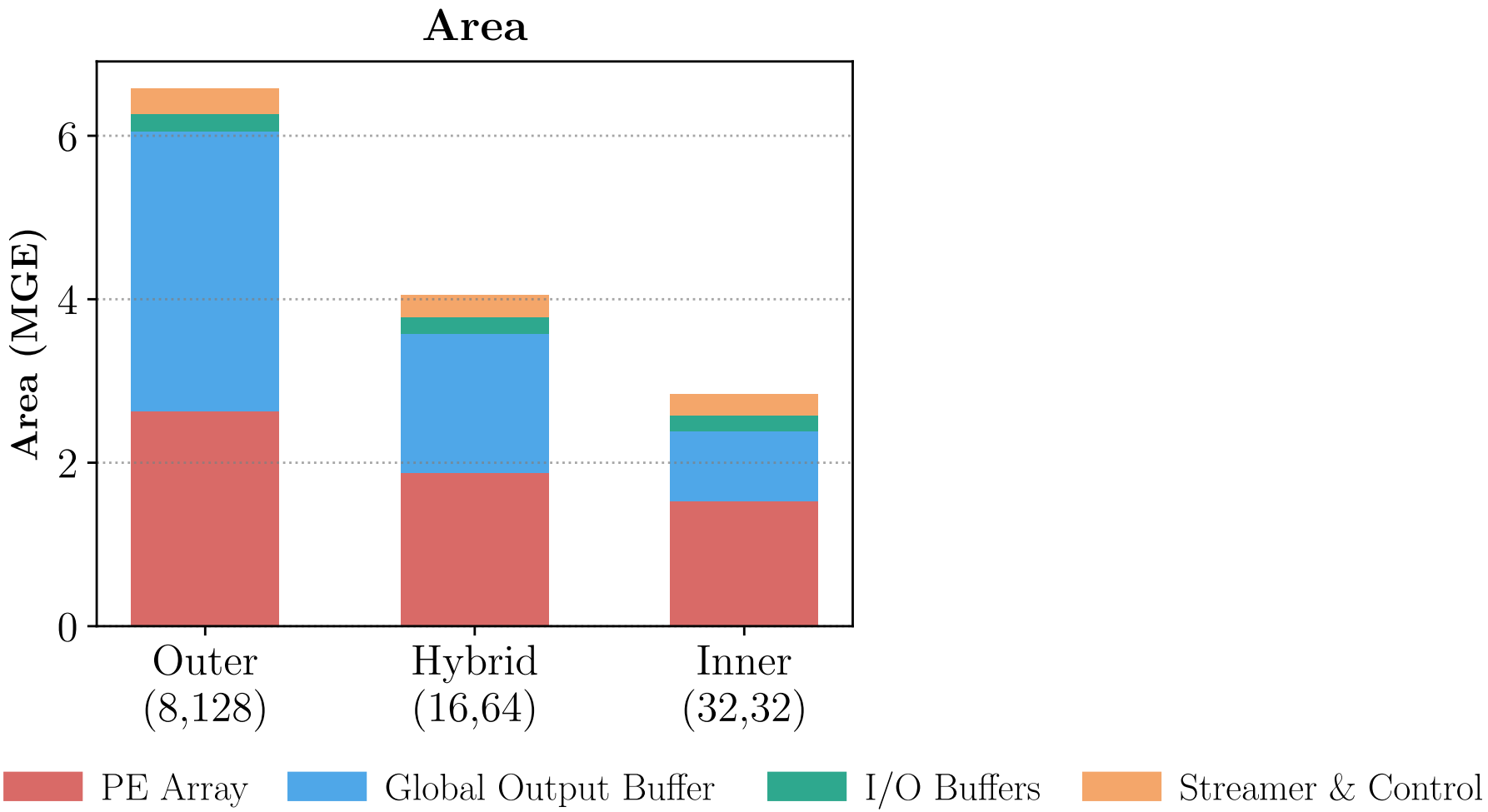
GlobalFoundries' 12nm FinFET technology

Target Frequency 0.8 GHz @ (SS/0.72V/125°C)

Outer to Inner: 0.98 – 1.07 GHz @ (TT/0.8V/25°C)

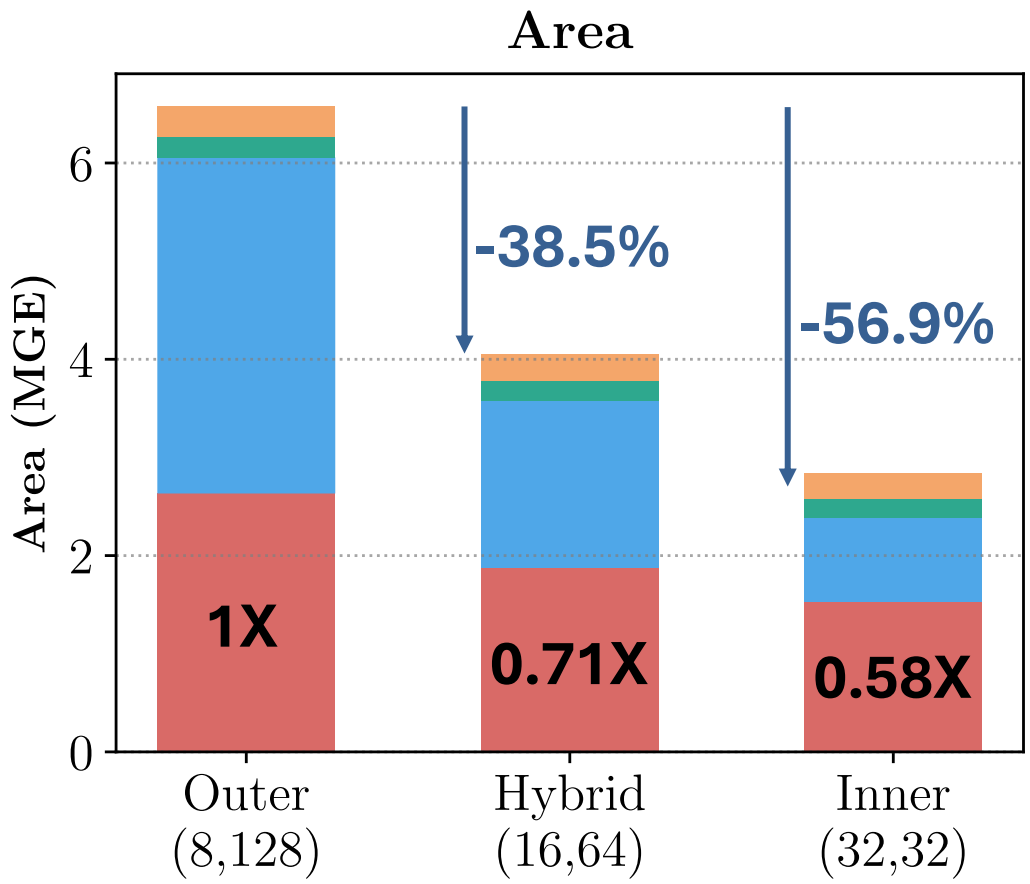
Post-layout simulations at **0.95 GHz** on *five* MXFP8-quantized samples extracted from *DeiT-Tiny*

Area Breakdown: Inner to Outer



Area Breakdown from Inner to Outer at a fixed *Reuse* of 64

Area Breakdown: Inner to Outer

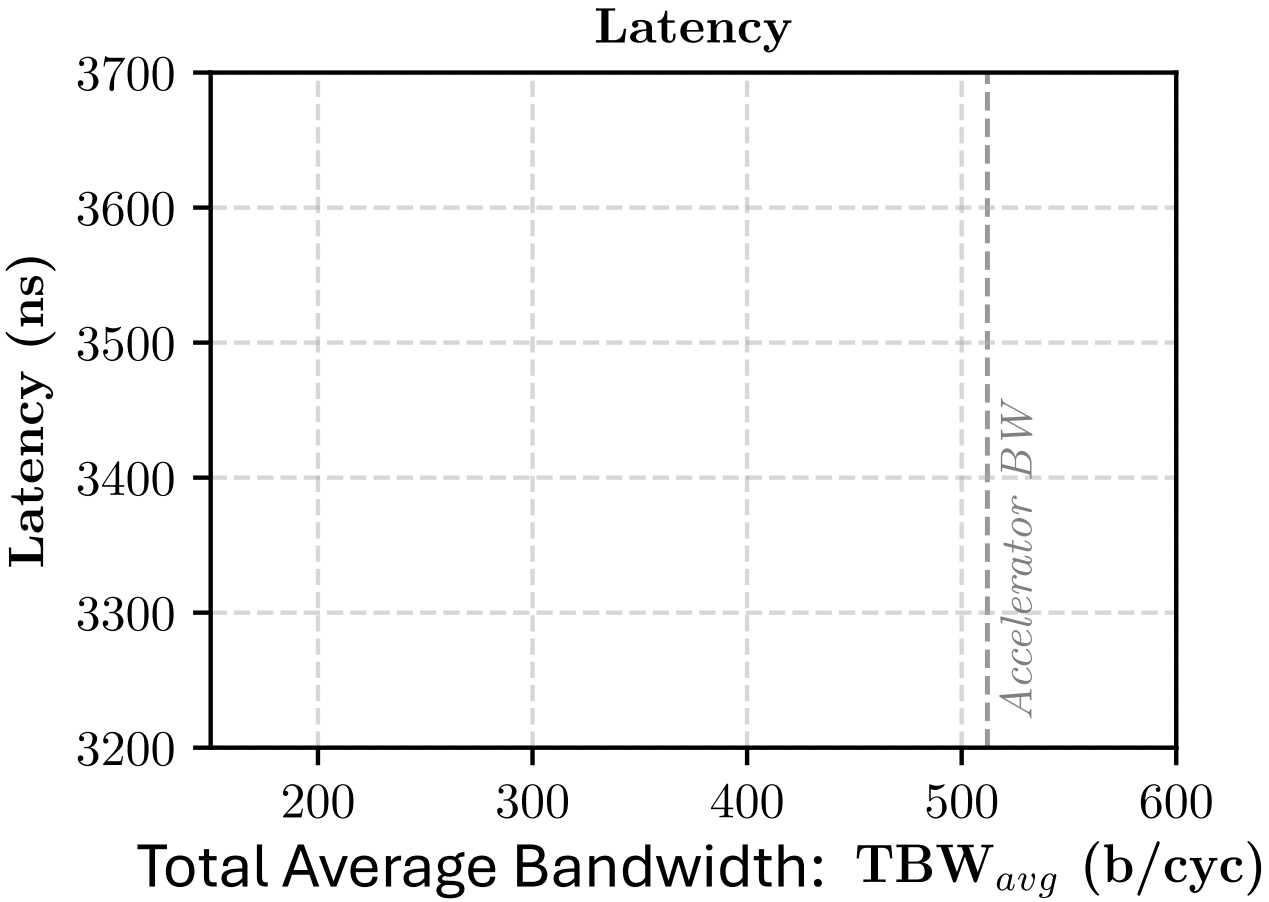


Observations

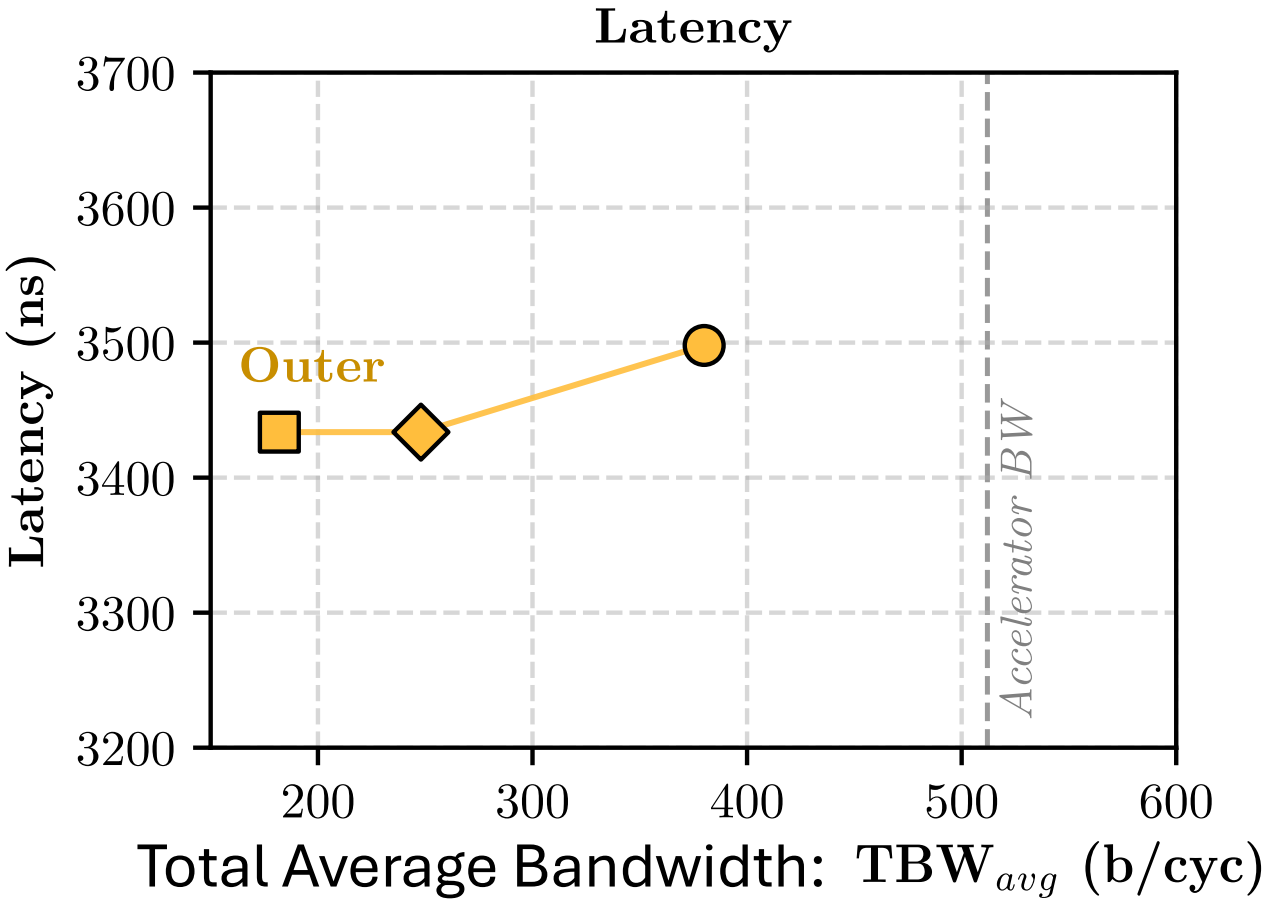
- *Inner* arrays favour area.
  - Increasing *VS* only adds narrow multipliers.
- *Outer* arrays are area-inefficient.
  - Increasing *NPE* replicates logic.

Area Breakdown from Inner to Outer at a fixed *Reuse* of 64

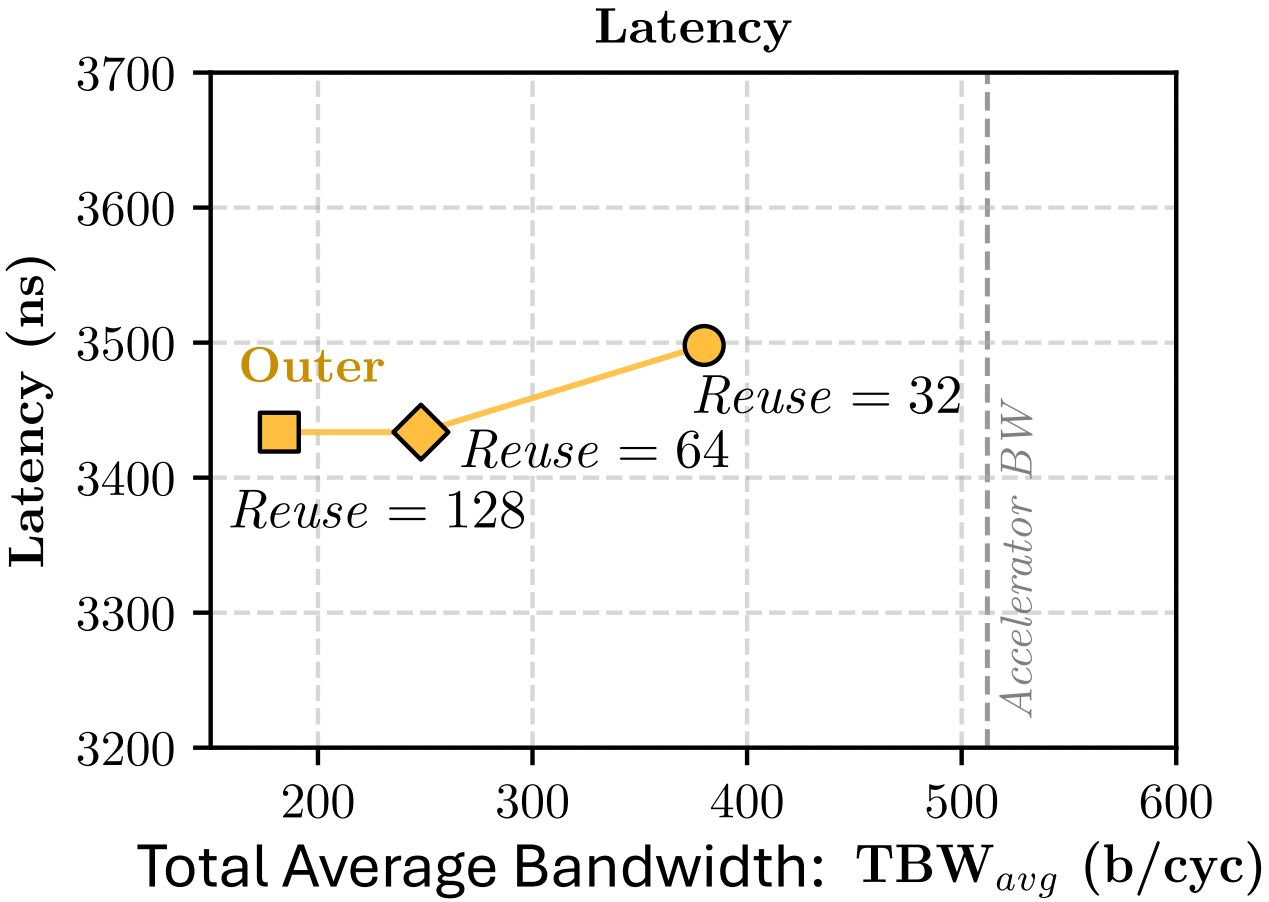
Pareto-optimal Designs: Latency



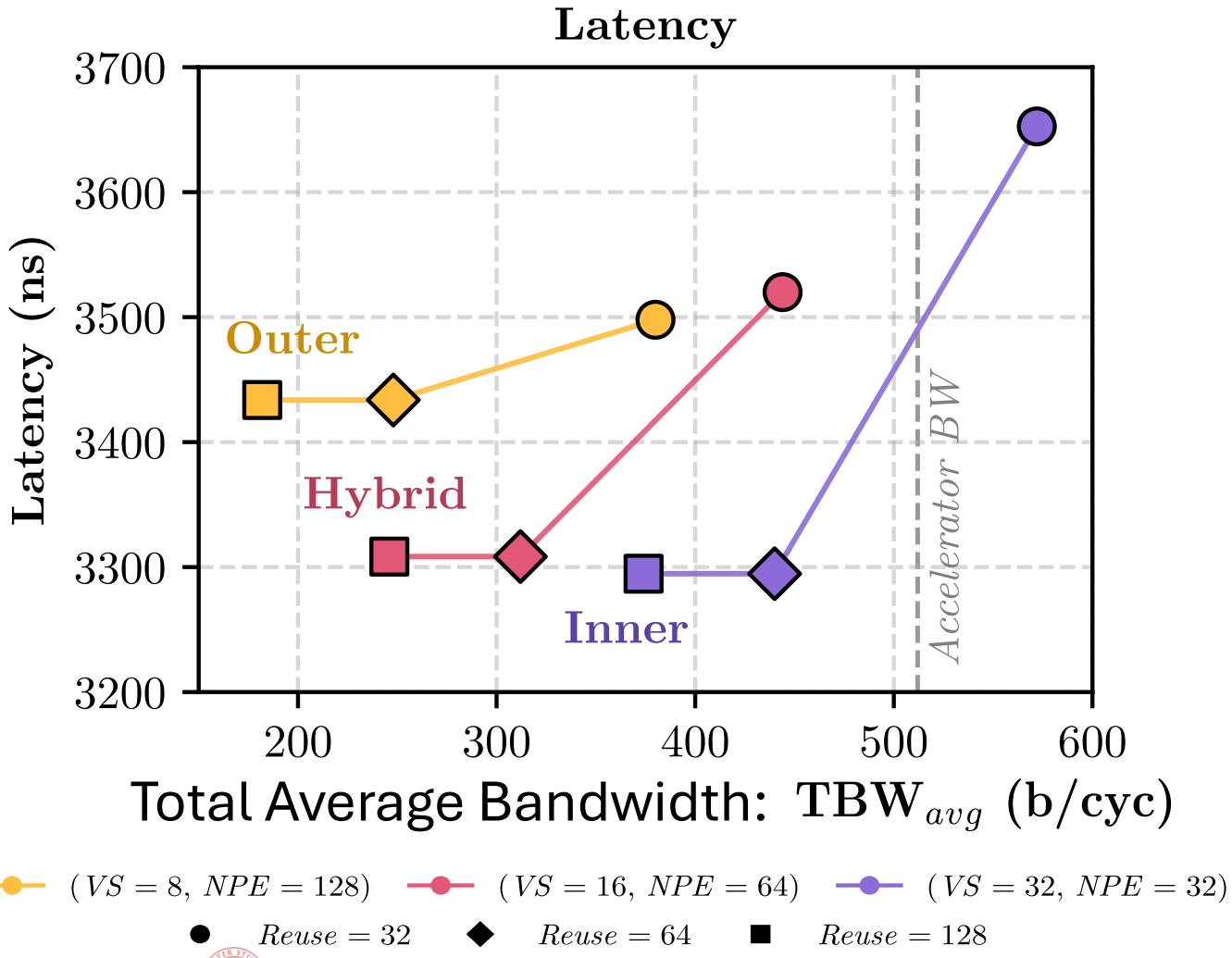
Pareto-optimal Designs: Latency



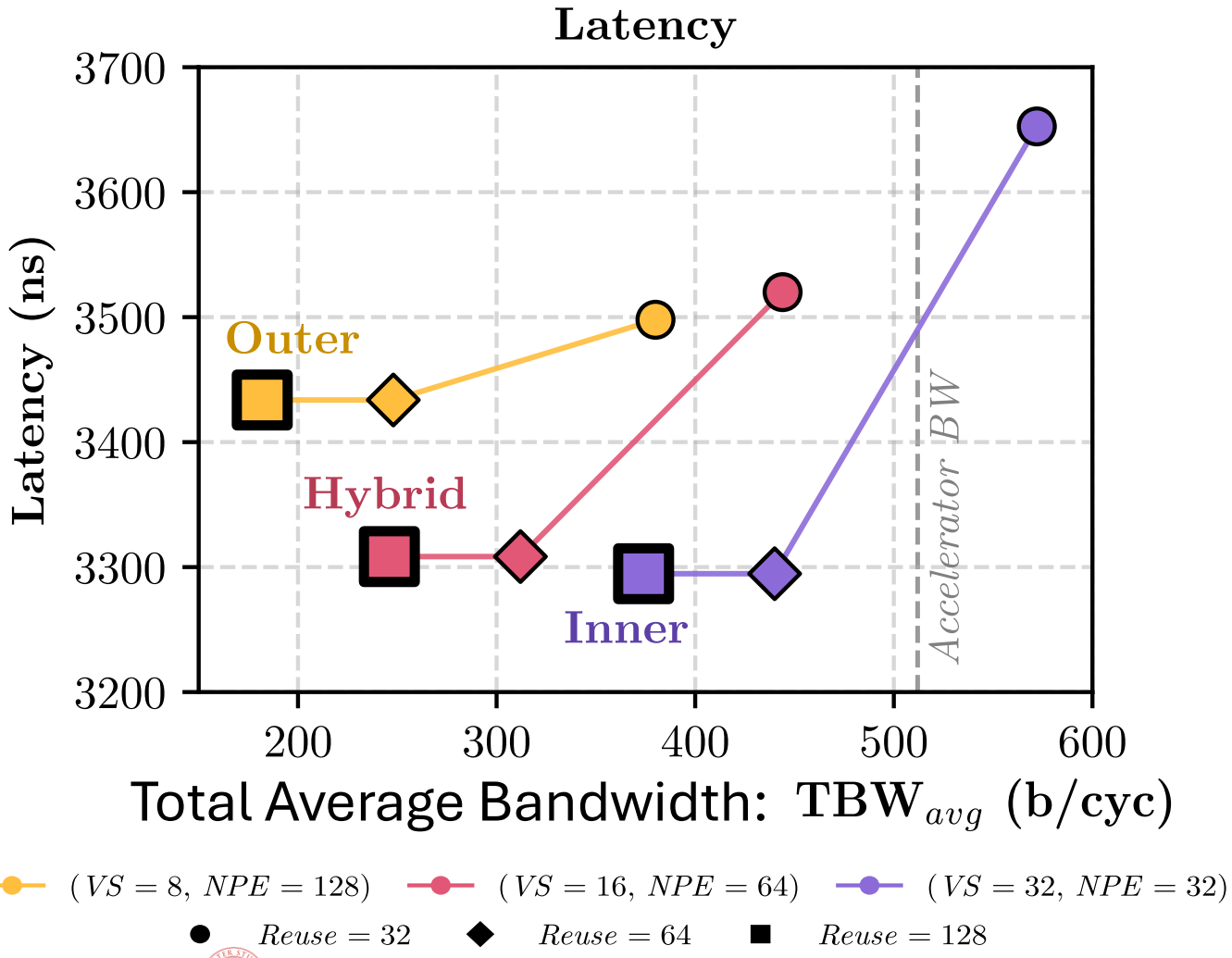
Pareto-optimal Designs: Latency



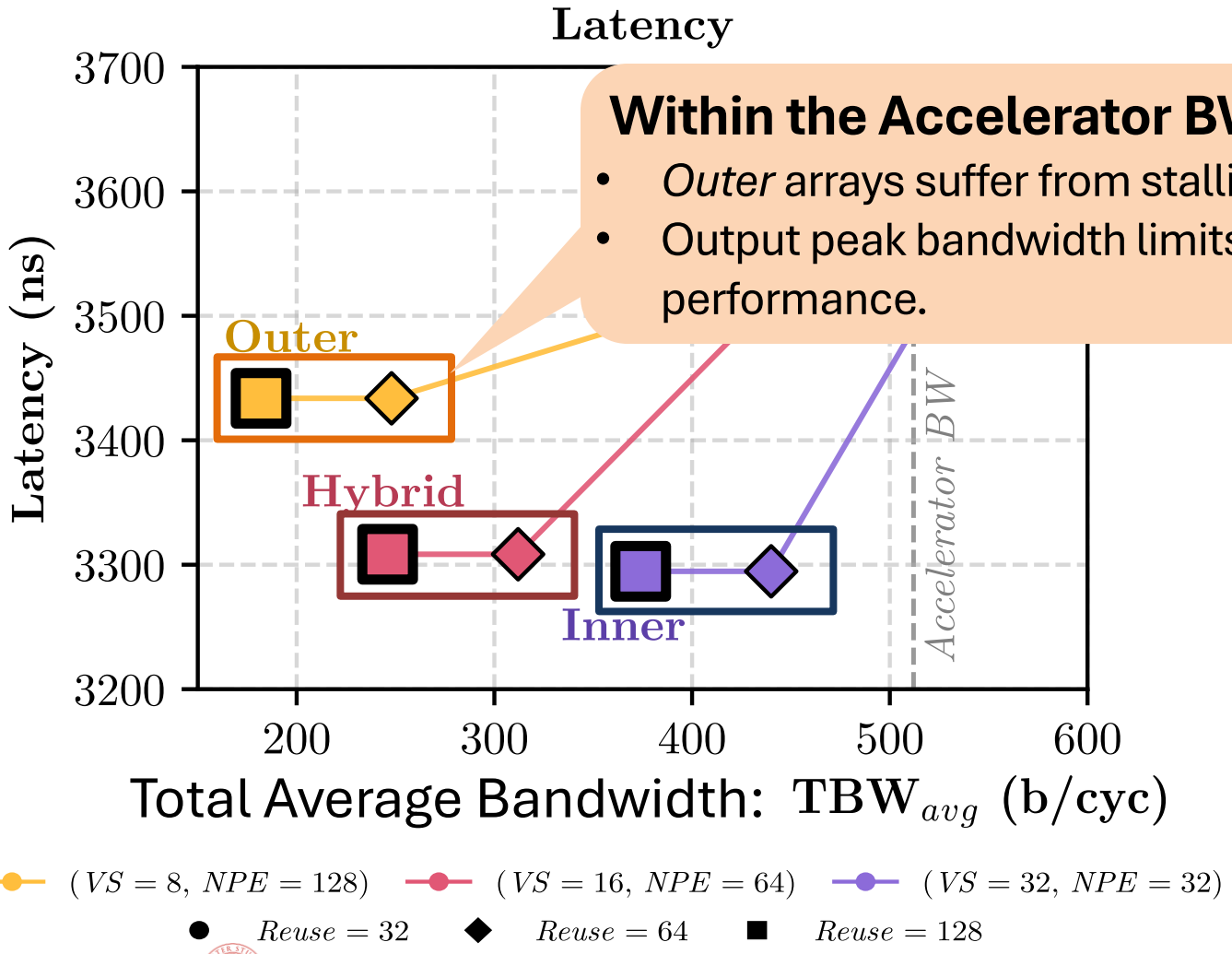
Pareto-optimal Designs: Latency



Pareto-optimal Designs: Latency



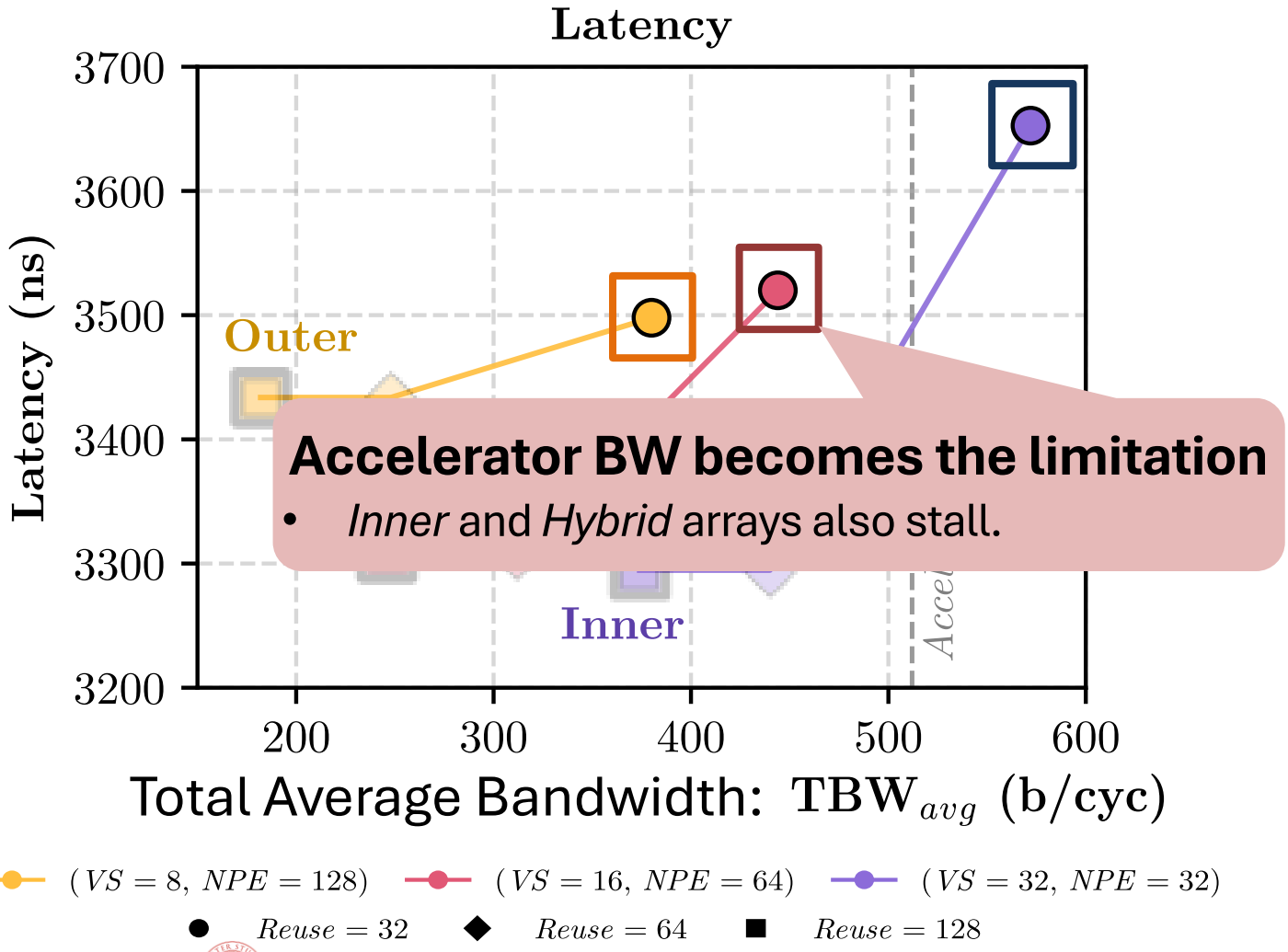
Pareto-optimal Designs: Latency



Within the Accelerator BW boundary

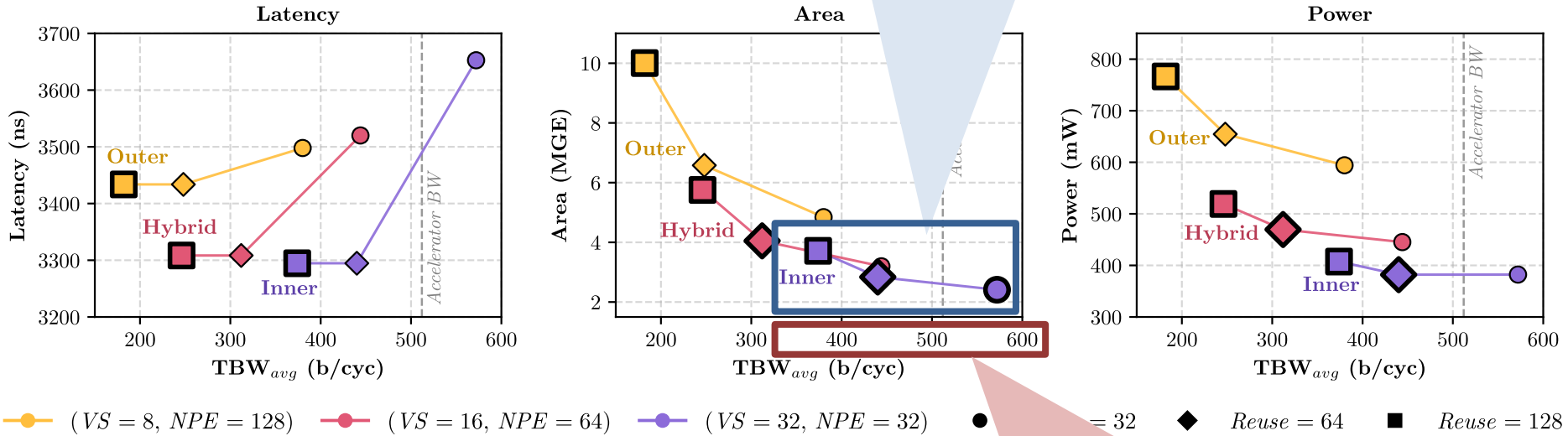
- *Outer* arrays suffer from stalling.
- Output peak bandwidth limits the performance.

Pareto-optimal Designs: Latency



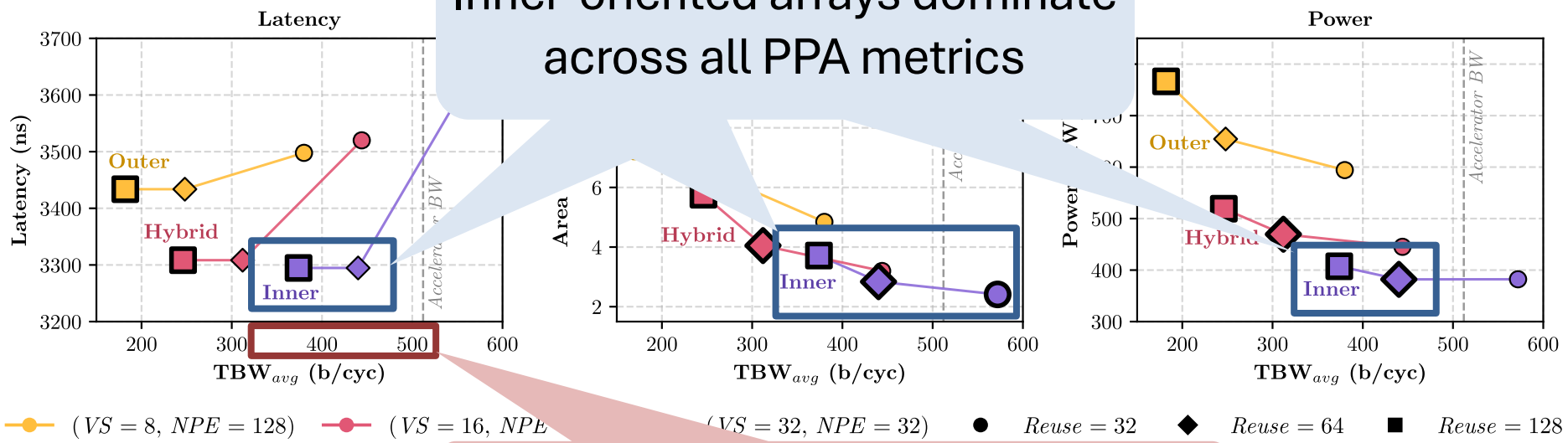
Pareto-optimal Designs: Trade-offs

Inner-oriented arrays offer superior area efficiency



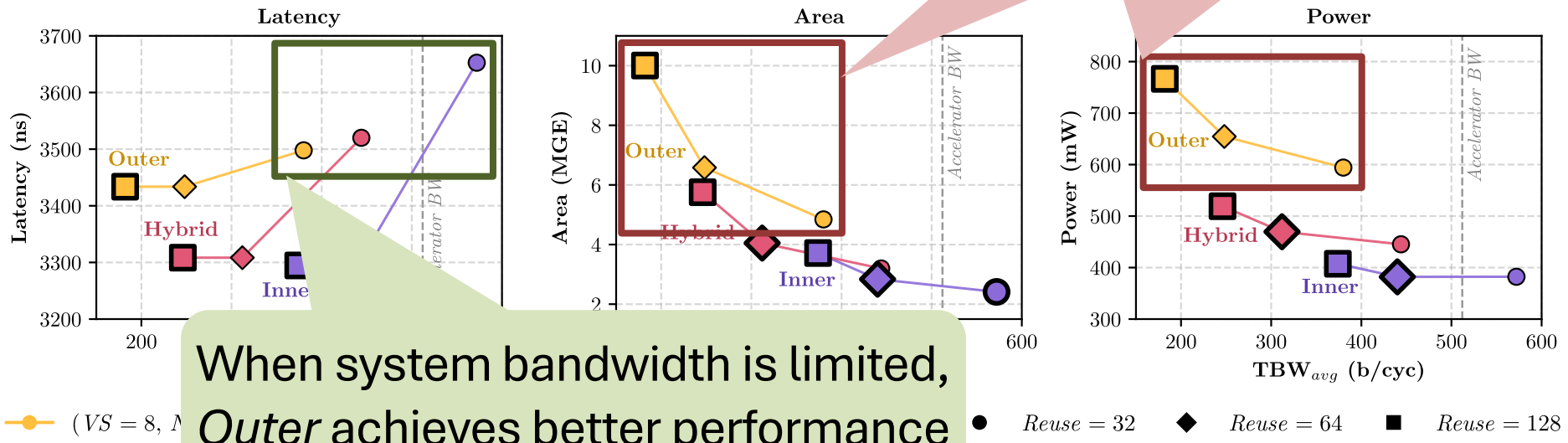
At the cost of bandwidth

Pareto-optimal Designs: Trade-offs



Pareto-optimal Designs: Trade-offs

At the expense of silicon and power



When system bandwidth is limited, *Outer* achieves better performance

Results: State-of-the-Art Comparison

Work	Tech. (nm)	Freq. (MHz)	Precision	Area Eff. GFLOPS/mm^2	Energy Eff. GFLOPS/W
Gemmini	22	210	INT8	59.4	106.1
OpenGEMM	16	200	INT2/4/8	329	4680
Cuyckens *et. al.*	16	400	All MX Formats	1469	943-1042
VEGETA	65	180	MXFP8	182/140	6460/5680
MXCore *Inner* (32, 32, 64)	12	950	MXFP8	3190	5093/4438 (8202/6742)

Results shown for E5M2/E4M3; Values in parantheses correspond to PE array efficiency.

Cuyckens *et. al.* 2025. Efficient Precision-Scalable Hardware for Microscaling (MX) Processing in Robotics Learning. doi:10.48550/arXiv.2505.22404
Nine *et. al.* 2025. Optimizing Sparse/Dense VEGETA Accelerator Performance with Microscaling Quantization. ISCAS. doi:10.1109/ISCAS56072.2025.11043957

Results: State-of-the-Art Comparison

Work	Tech. (nm)	Freq. (MHz)	Precision	Area Eff. GFLOPS/mm^2	Energy Eff. GFLOPS/W
Gemmini	22	210	INT8		106.1
OpenGEMM	16	200	INT2/4/8		4680
Cuyckens *et. al.*				1469	.2
VEGETA				182/140	30
MXCore *Inner* (32, 32, 64)	12	950	MXFP8	3190	5093/4438 (8202/6742)

48X

1.1X

No MX support
Widely used *SOTA* frameworks

Results shown for E5M2/E4M3; Values in parantheses correspond to PE array efficiency.

Cuyckens *et. al.* 2025. Efficient Precision-Scalable Hardware for Microscaling (MX) Processing in Robotics Learning. doi:10.48550/arXiv.2505.22404
Nine *et. al.* 2025. Optimizing Sparse/Dense VEGETA Accelerator Performance with Microscaling Quantization. ISCAS. doi:10.1109/ISCAS56072.2025.11043957

Results: State-of-the-Art Comparison

Work	Tech. (nm)	Freq. (MHz)	Precision	Area Eff. GFLOPS/mm^2	Energy Eff. GFLOPS/W
Cuyckens *et. al.*	16	400	All MX Formats	1469	943-1042
VEGETA	65	180	M	182/140	6460/5680
MXCore *Inner* (32, 32, 64)	12	950	MXFP8	3190	5093/4438 (8202/6742)

Results shown for E5M2/E4M3; Values in parantheses corr fficiency.

2.2X

7.2 – 7.9X

Cuyckens *et. al.* 2025. Efficient Precision-Scalable Hardware for Microscaling (MX) Processing in Robotics Learning. doi:10.48550/arXiv.2505.22404
Nine *et. al.* 2025. Optimizing Sparse/Dense VEGETA Accelerator Performance with Microscaling Quantization. ISCAS. doi:10.1109/ISCAS56072.2025.11043957

Results: State-of-the-Art Comparison

Work	Tech. (nm)	Freq. (MHz)	Precision	Area Eff. GFLOPS/mm^2	Energy Eff. GFLOPS/W
Cuyckens *et. al.*	16	400	All MX Formats	1469	943-1042
VEGETA	65	180	MXFP8	182/140	6460/5680
MXCore *Inner* (32, 32, 64)	12	950	MXFP8	3190	5093/4438 (8202/6742)

Results shown for E5M2 [illegible] espond to PE array efficiency.

1.7 – 2.2X
(After technology scaling)

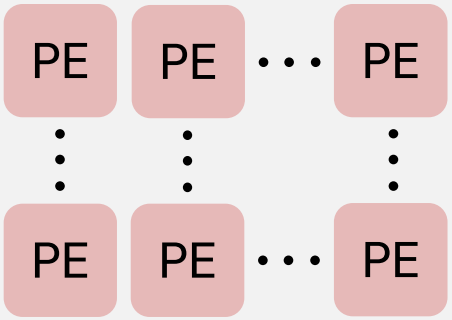
Cuyckens *et. al.* 2025. Efficient Precision-Scalable Hardware for Microscaling (MX) Processing in Robotics Learning. doi:10.48550/arXiv.2505.22404
Nine *et. al.* 2025. Optimizing Sparse/Dense VEGETA Accelerator Performance with Microscaling Quantization. ISCAS. doi:10.1109/ISCAS56072.2025.11043957

Conclusion

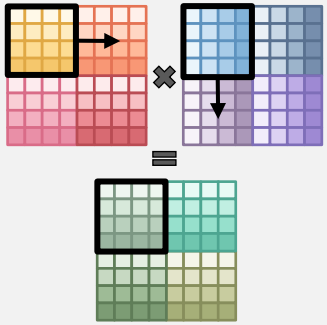
- MXCore: MX Matrix Multiplication Accelerator
  - Fully parametric and configurable.
  - System-level integration capabilities.
  - FP32-to-MX quantizer for output bandwidth reduction.
  - Pareto-optimal designs and PPA comparisons.
- Inner vs Outer Product Trade-offs
  - Outer to Inner: **58.6%** higher bandwidth demand.
  - Outer to Inner: **2.1X** higher energy efficiency.
  - Outer to Inner: **3.7X** higher area efficiency.
- State-of-the-Art Comparison
  - **2.2X** higher area efficiency.
  - **1.2 – 7.9X** higher energy efficiency.

Future Work

Architectural Explorations

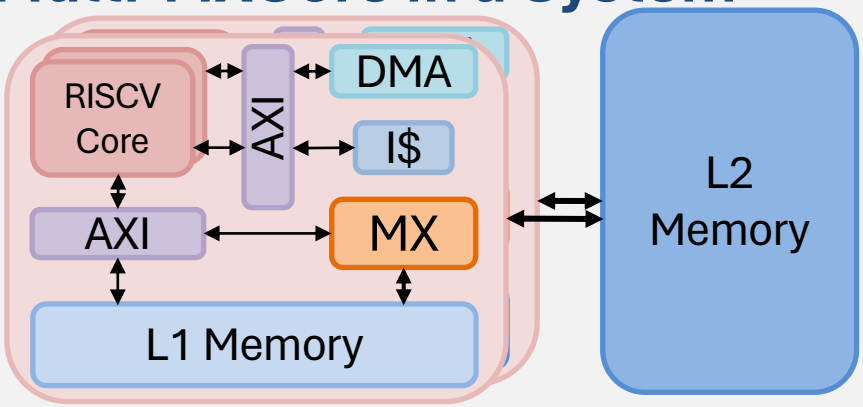


PE Organization



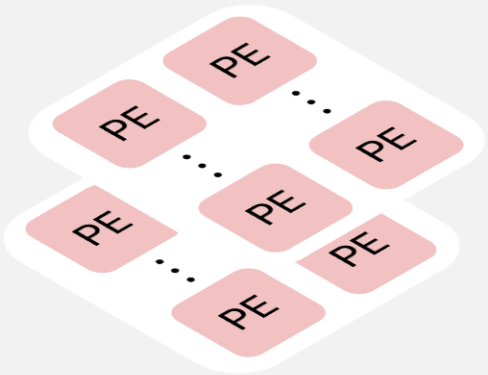
Dataflow

Multi-MXCore in a System

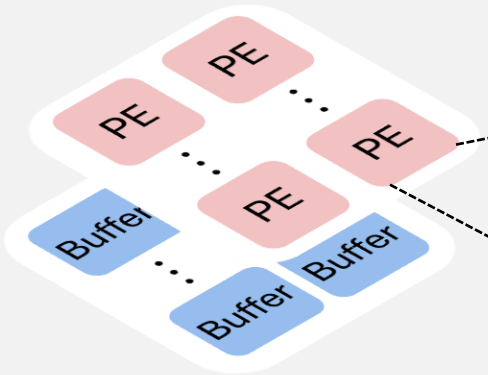


System-level Performance & Scalability

Novel Design Paradigms: Going 3D with MXCube



Compute Scalability



Shorter Interconnects

PE Optimization

THANK YOU