Silicon Photonic Accelerated Memory Pooling For Efficient Compute Resource Allocation

Zhenguo Wu and Keren Bergman
zw2542@columbia.edu

Lightwave Research Laboratory
Columbia University, New York, NY

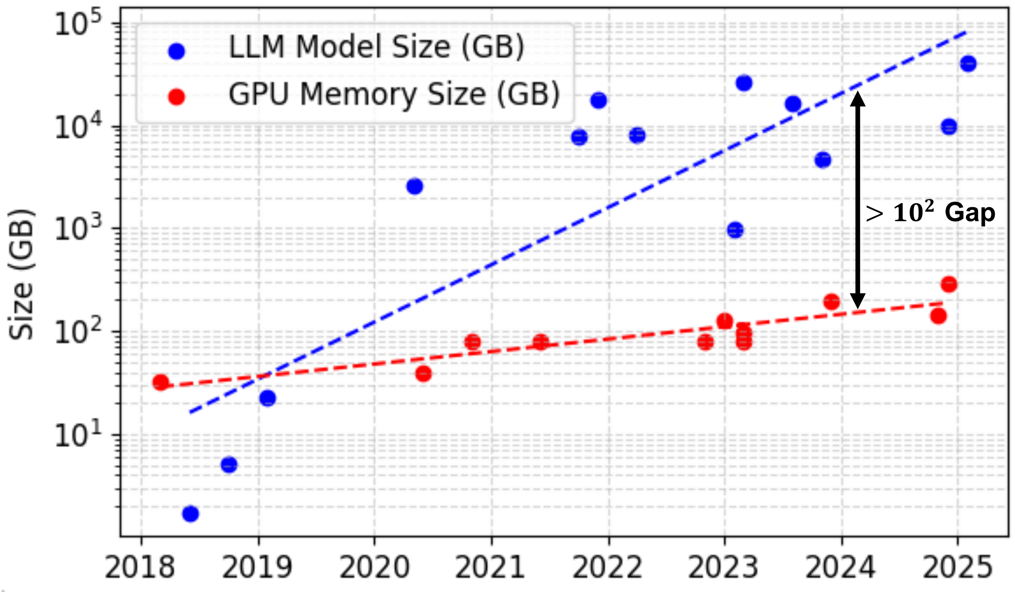
Aug 20th, 2025

COLUMBIA ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Memory Challenges in Scaling Large Language Models

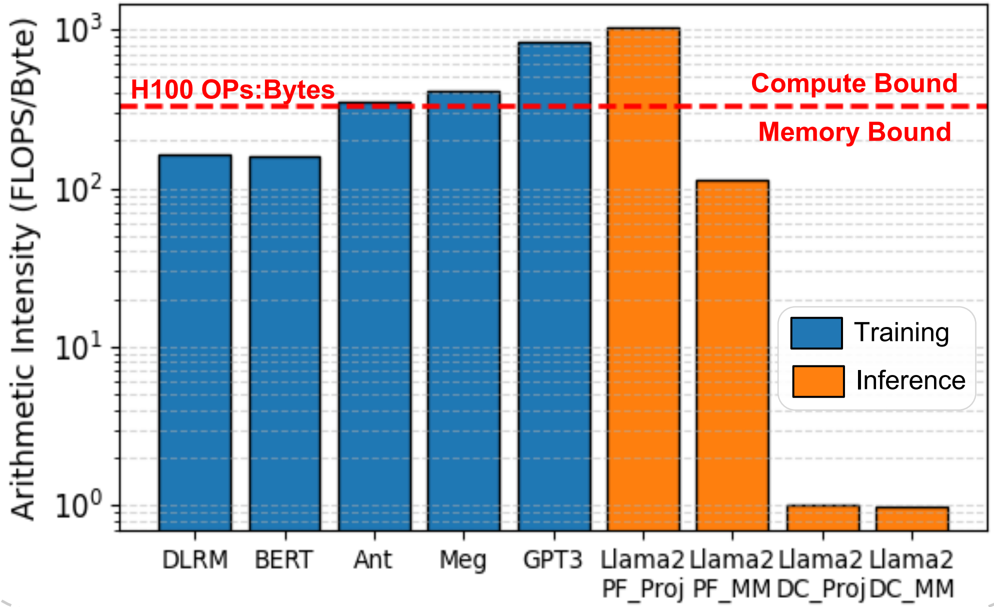
GPU Memory Limitations

- The growth rate of high-bandwidth memory (HBM) per GPU is much slower than the rapid scaling of Large Language Models (LLMs).
- The growing gap limits batch size scaling and adds distributed training/inference complexity.

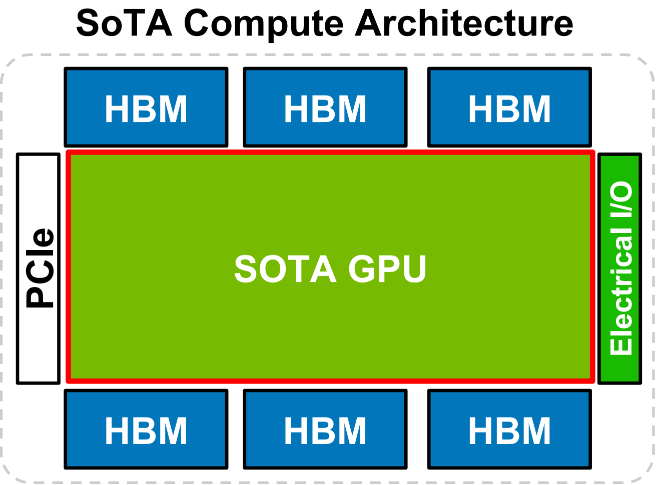


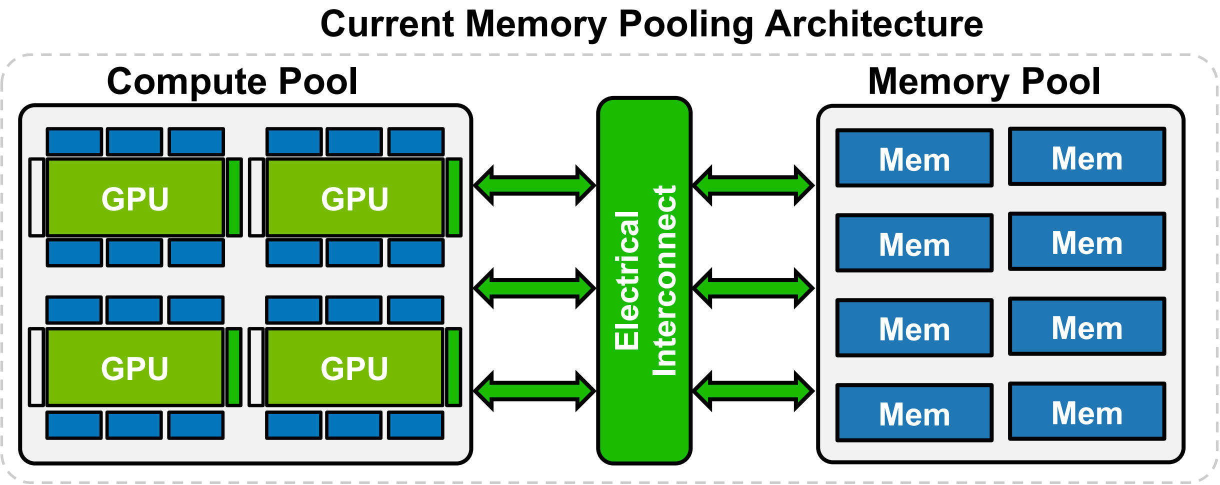
Arithmetic Intensity Variations

- LLM models exhibit diverse arithmetic intensities across workloads for training and inference, as well as between the prefill and decode stages of inference.
- Each workload (or stage) imposes unique demands on compute power and memory bandwidth.



Current Compute & Memory Pooling Architecture



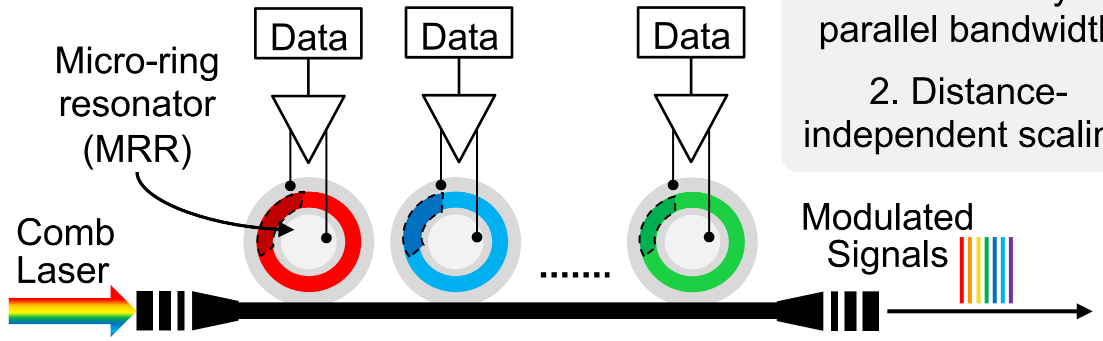


- ❖ **HBMs are placed around the periphery of the compute die using short-reach electrical I/Os.**
  - ➢ HBM’s wide I/O and high pin count requires short electrical traces, restricting it to areas near the compute die.
  - ➢ Compute die's periphery length restricts the number of HBMs that can be integrated locally.
  - ➢ Limited memory capacity scaling.

- ❖ **Memory Pooling: compute pool connects to memory pool via a high-speed electrical interconnect.**
  - ➢ Local HBMs serve as high-bandwidth memory suppliers (lower capacity).
  - ➢ Remote memory units (e.g., DDR, GDDR) act as capacity expanders via an electrically interconnected fabric.

Enabling Silicon Photonic Technologies

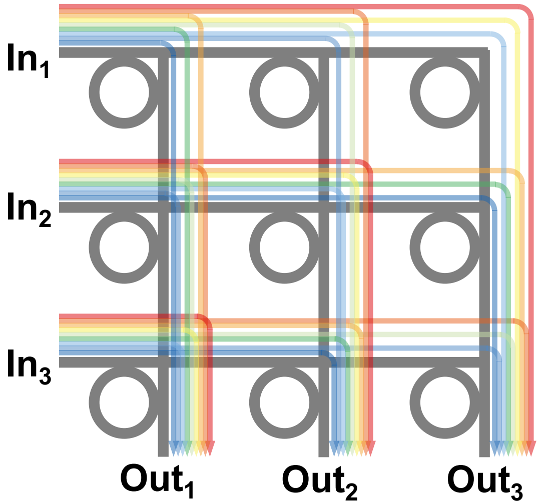
Embedded Photonic Transceiver

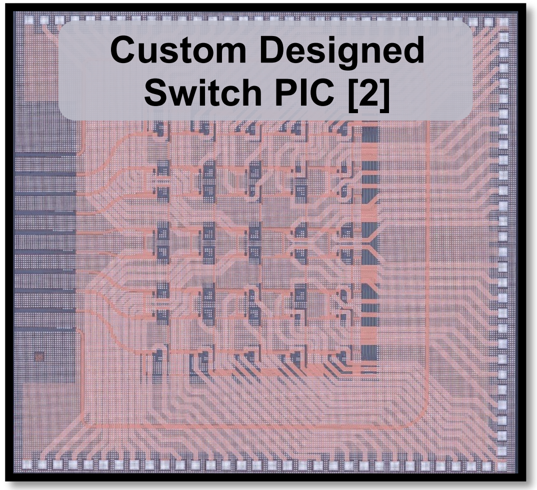


1. Massively parallel bandwidth

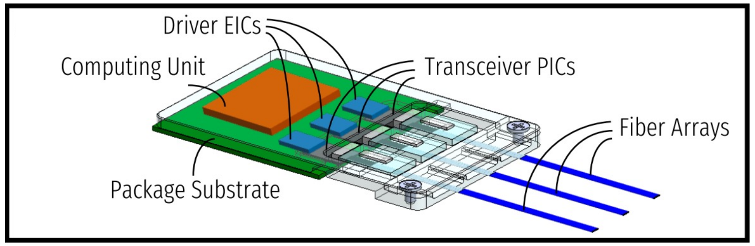
2. Distance-independent scaling

Optical Circuit Switches

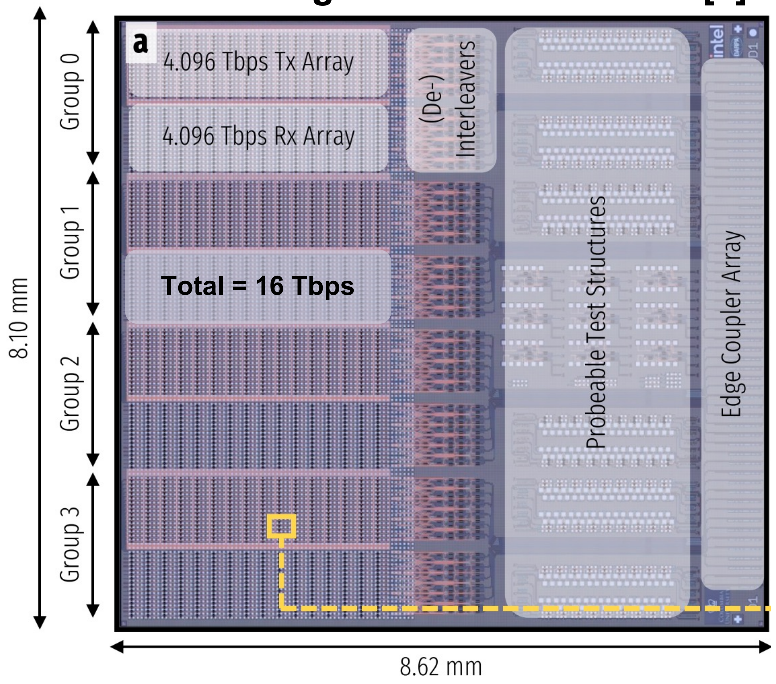




Co-Packaged/Embedded Photonic I/Os



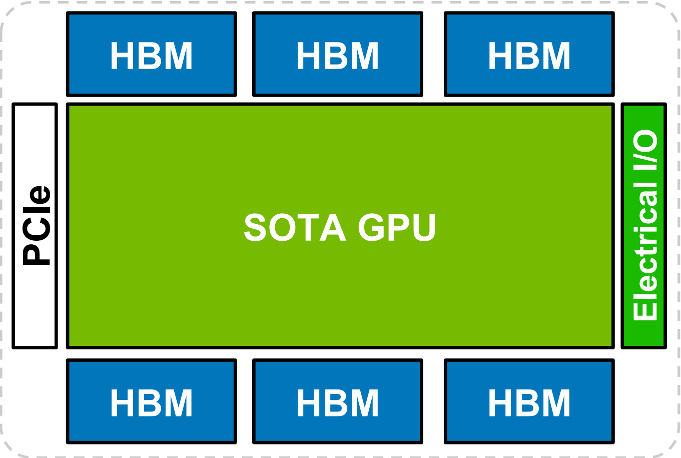
Custom Designed Transceiver PIC [1]

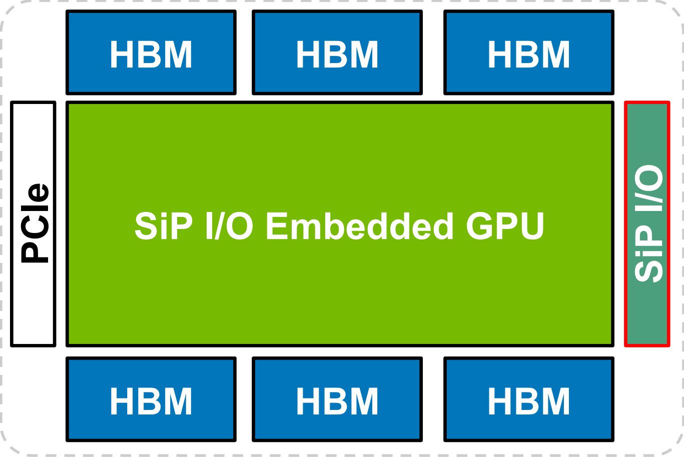


[1] Wang, Yuyang, et al. “Co-designed silicon photonics chip i/o for energy-efficient petascale connectivity.” IEEE Transactions on Components, Packaging and Manufacturing Technology. IEEE, 2024
[2] Dai, Liang Yuan, et al. "Ultra-scalable microring-based architecture for spatial-and-wavelength selective switching." 2023 IEEE Silicon Photonics Conference (SiPhotonics). IEEE, 2023.

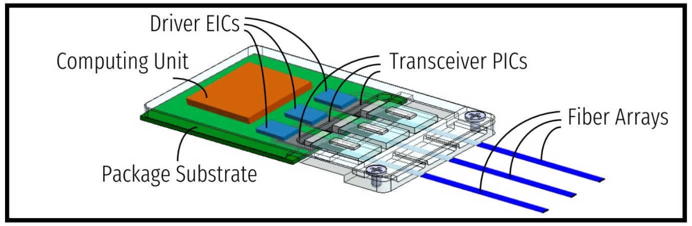
Expanding the Memory Pooling Design Space

Compute die's *shoreline width* is used a critical resource.

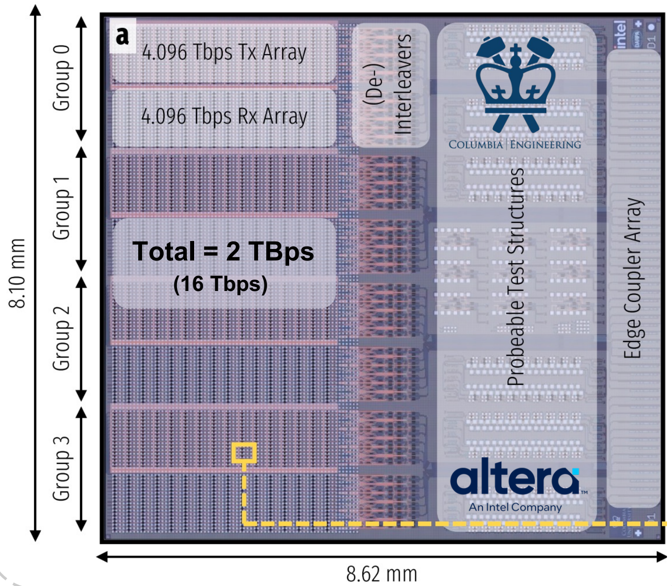


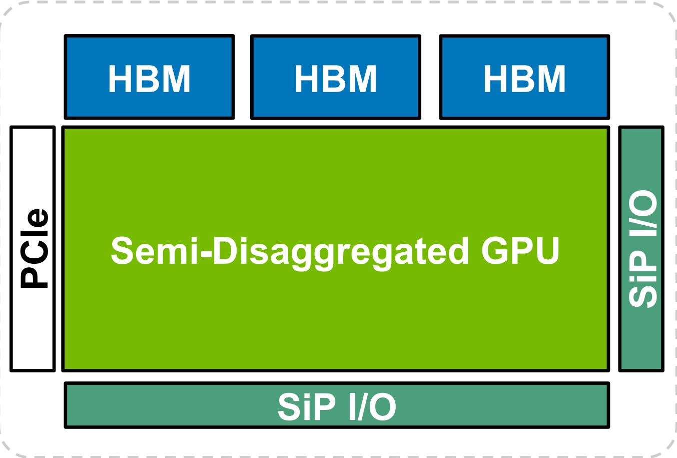


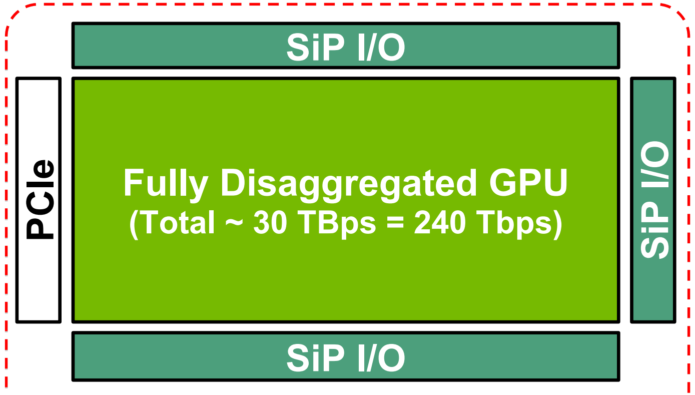
Conceptual Embedded SiP Chip I/O



Designed and fabricated transceiver PIC

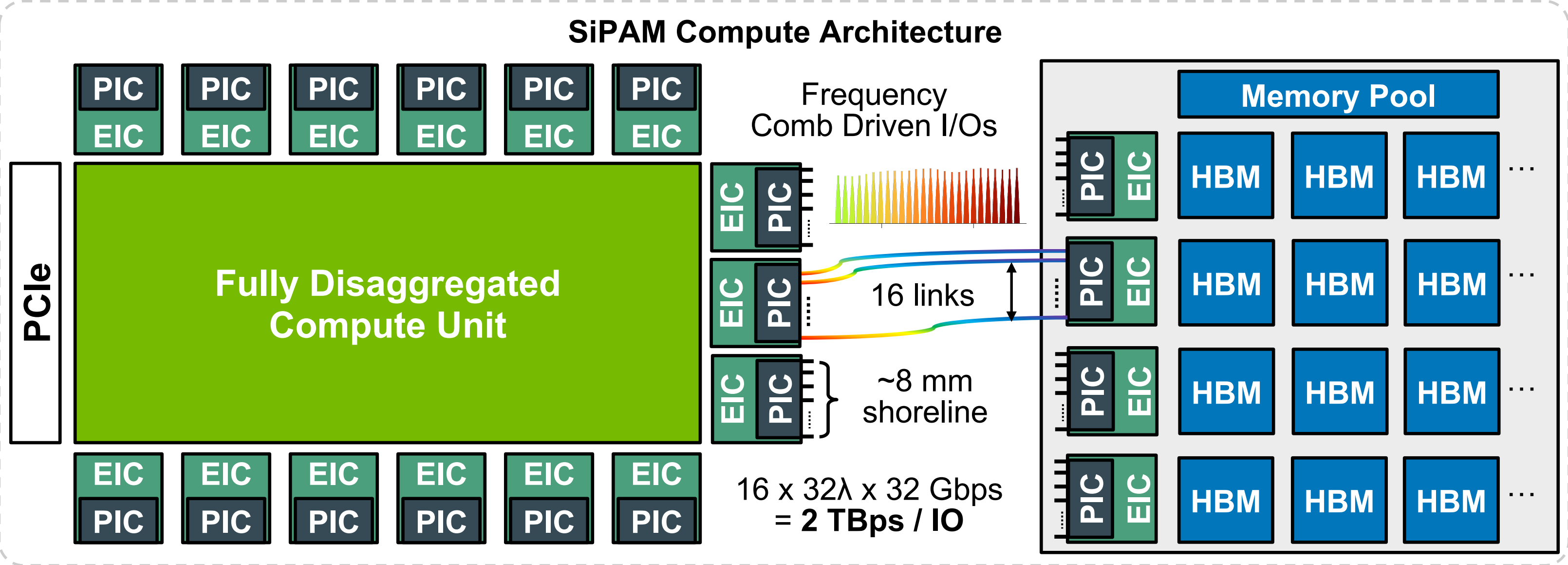


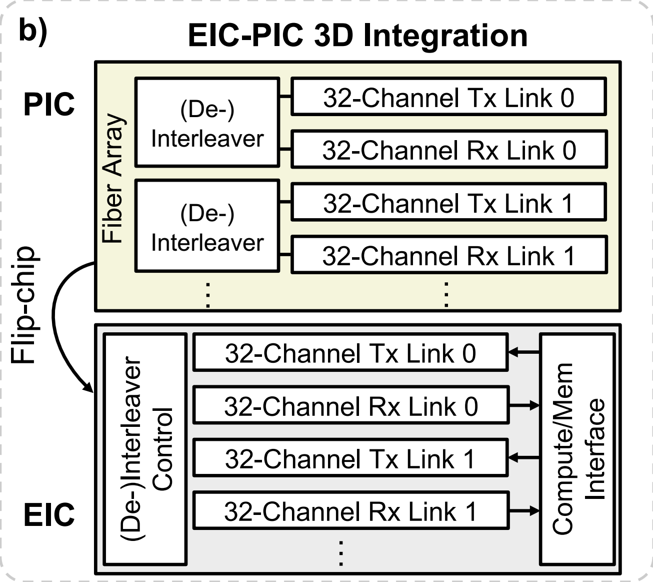




Unified high-bandwidth domain for memory and network communication

SiPAM: Silicon Photonic Accelerated Memory-Pooling

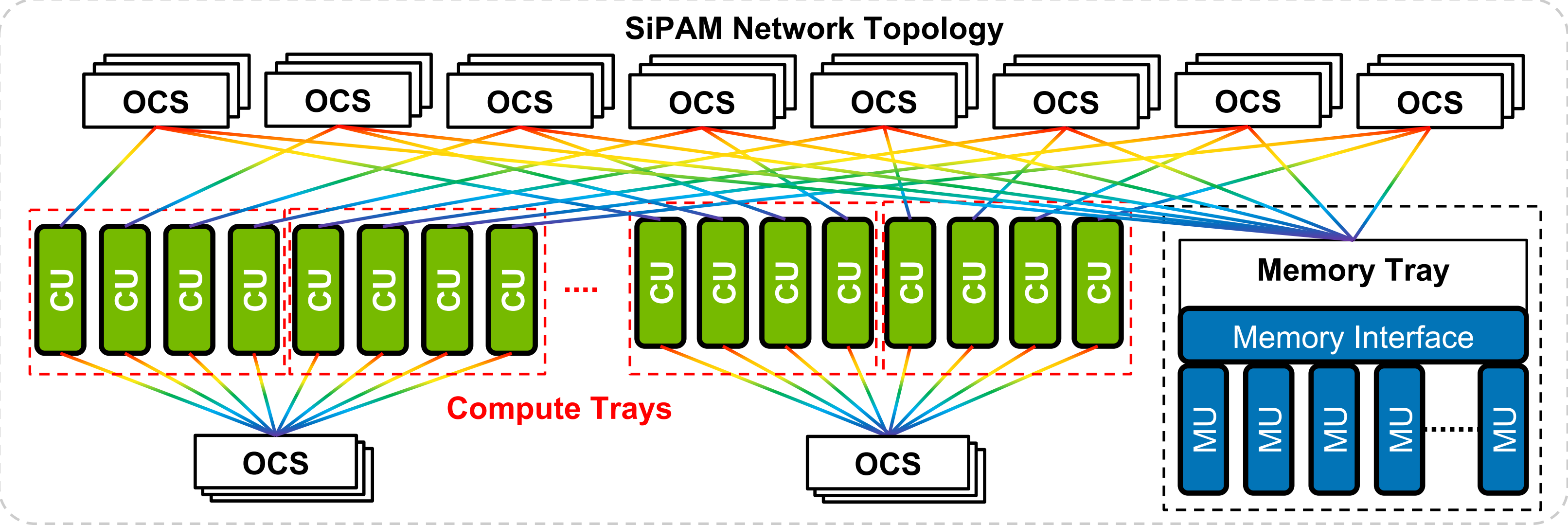

SiPAM Compute Architecture
PIC
EIC
PCIe
Fully Disaggregated Compute Unit
Frequency Comb Driven I/Os
16 links
~8 mm shoreline
16 x 32λ x 32 Gbps = 2 TBps / IO
Memory Pool
HBM
PIC
EIC

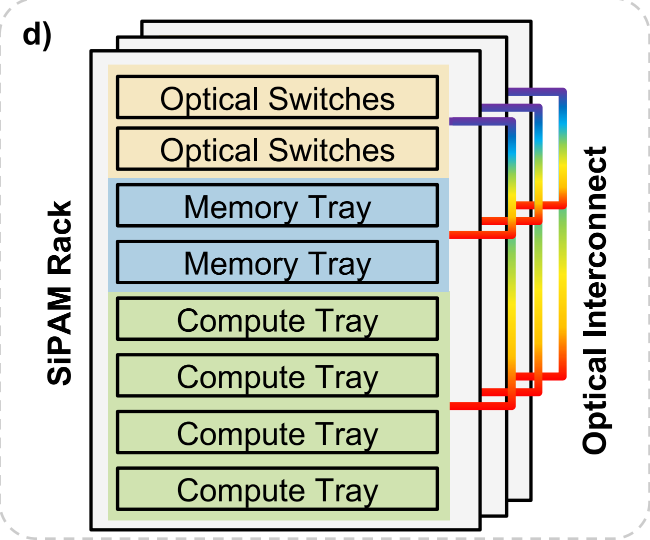


- ❖ SiP I/Os: 3D integrated EIC and PIC through flip-chip bonding
  - ➢ Each SiP I/O : 16 links × 32 λ / link × 32 Gbps / λ = 2 TBps
- ❖ Number of integratable I/Os: $N_{IO} = \lfloor W_D / W_{IO} \rfloor$
  - ➢ W_D = Available compute die shoreline width
  - ➢ W_{IO} = Edge width per SiP I/O
- ❖ Memory Pool: Optically Connected Multi-Stack HBM [3]
  - ➢ Multiple HBMs connect to a single SiP I/O chiplet
- ❖ Number of integratable MUs / IO: $N_m = \lfloor B_{IO} / B_m \rfloor$
  - ➢ B_{IO} = SiP I/O bandwidth
  - ➢ B_m= MU bandwidth

[3] Ou, Y., Zhang, H., Rovinski, A., Wentzlaff, D., & Batten, C. (2025). Optically Connected Multi-Stack HBM Modules for Large Language Model Training and Inference. IEEE Computer Architecture Letters.

SiPAM: Silicon Photonic Accelerated Memory-Pooling





- Each SiP I/O can be flexibly allocated for high-speed memory access or network communication.
  - One-shot reconfiguration per workload.
- SiPAC's physical design: replaces electrical packet switches (EPS) with optical circuit switches (OCS) in a BCube topology
  - Intra-rack resource disaggregation model [4] for a bounded increase in memory latency.
- CXL is a promising memory semantic interconnect technology:
  - Increased memory latency can be mitigated by increasing CXL bandwidth when the memory system is fully loaded [5].

[4] Michelogiannakis, George, et al. "Efficient intra-rack resource disaggregation for HPC using co-packaged DWDM photonics." 2023 IEEE International Conference on Cluster Computing (CLUSTER). IEEE, 2023.
[5] Cho, A., Saxena, A., Qureshi, M., & Daglis, A. (2024, November). COAXIAL: A CXL-centric memory system for scalable servers. In SC24.

Optimization Methodology

Goal: Determine the optimal configuration for **compute power**, **memory bandwidth**, and **capacity** for each workload.

Hardware

- **Compute Intensity (CI):** # of *required* FLOPs per byte of data loaded to keep cores active.

$$CI = \frac{\text{Peak FLOPs/s}}{\text{Bandwidth}_{\text{mem}}} = \frac{\text{FLOPs}}{\text{Byte}}$$

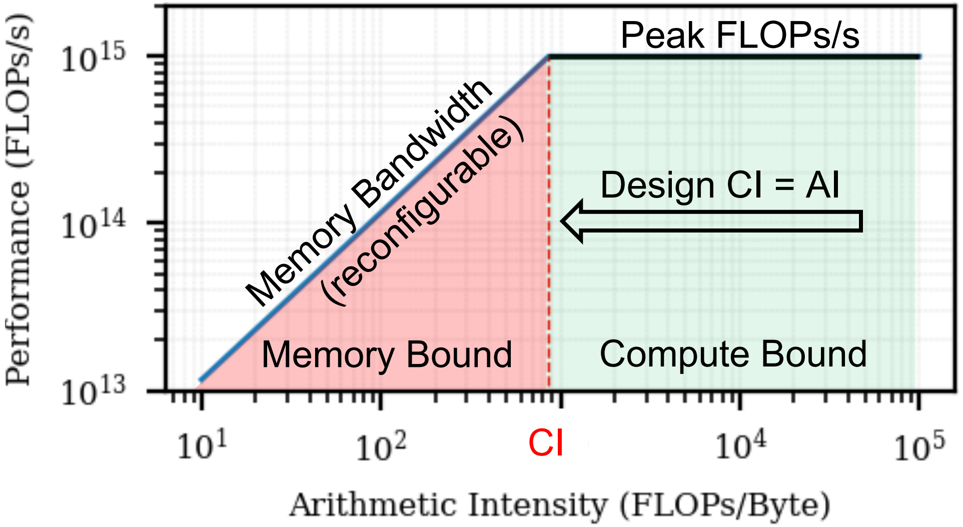
Workload

- **Arithmetic Intensity (AI):** # of *actual* FLOPs performed for each byte of data loaded.

$$AI = \frac{\text{FLOPs}}{\text{Byte}}$$

System Performance $= min \begin{cases} \text{Peak FLOPs/s} \\ \text{Bandwidth}_{\text{mem}} \times \text{AI} \end{cases}$

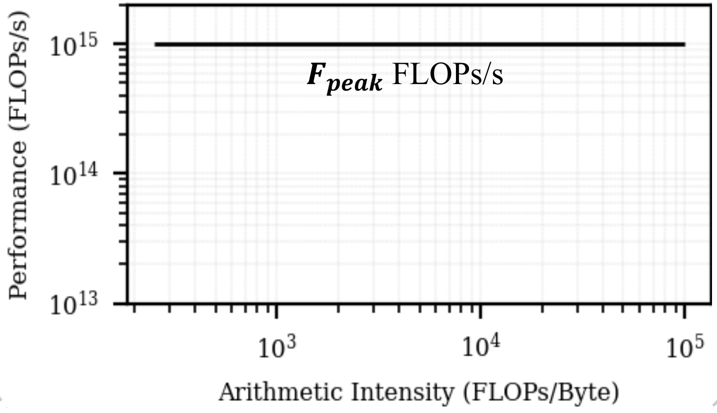
Roofline Model



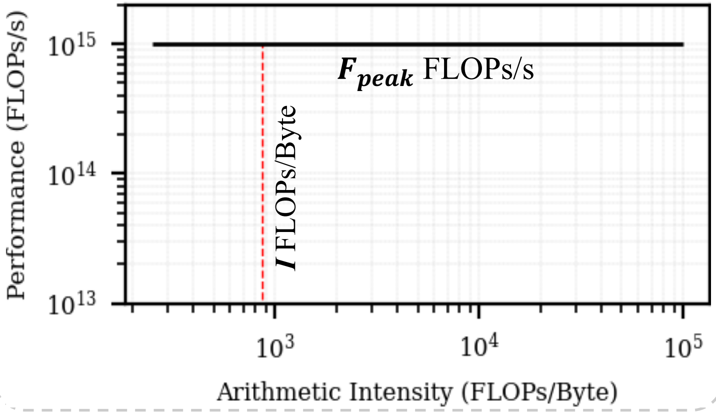
Optimization Methodology

Goal: Determine the optimal configuration for **compute power**, **memory bandwidth**, **capacity** for each workload.

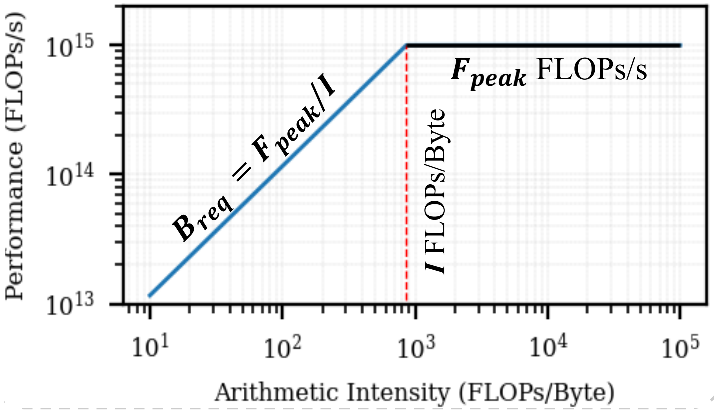
(1) Peak Compute FLOPs



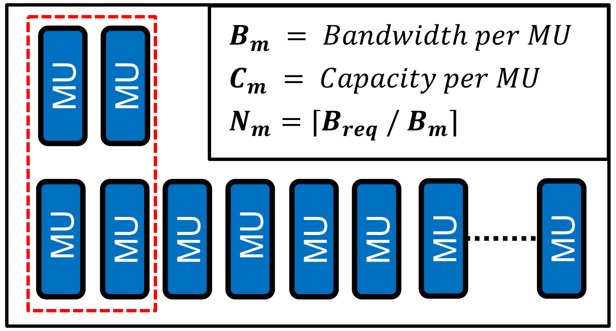
(2) Workload Arithmetic Intensity



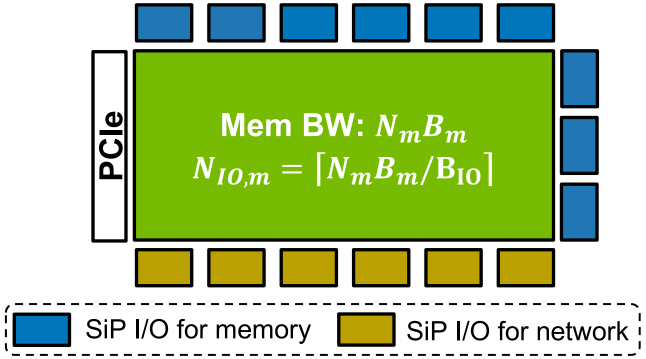
(3) Required Memory Bandwidth



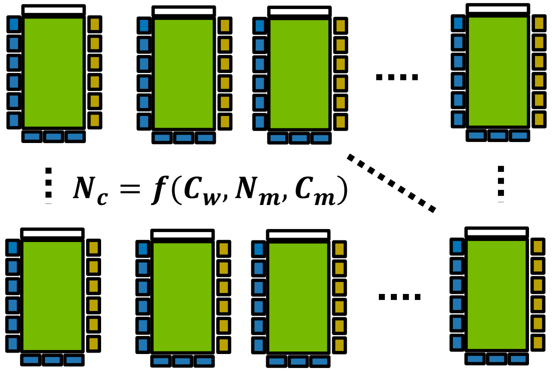
(4) Number of memory units to satisfy memory bandwidth requirement



(5) Distribution of SiP I/Os for memory and network communication



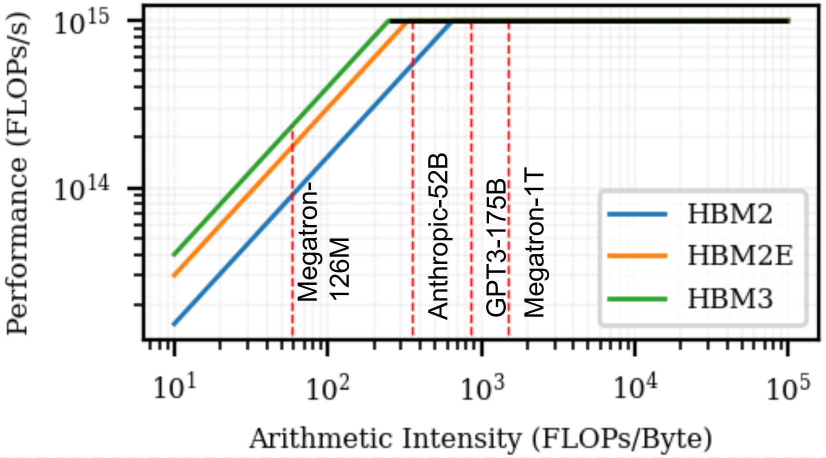
(6) Number of CUs + parallelisms



Evaluation Setup – Calculon [6]

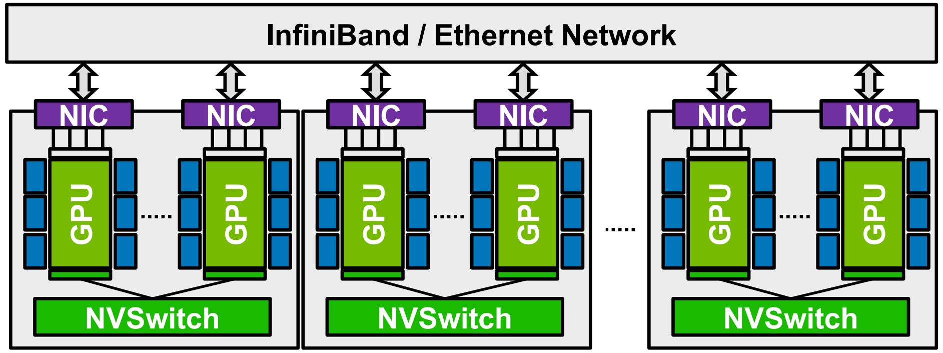
Workload & Configurations

- **Arithmetic Intensity:** profiled using Calculon
- **Capacity Requirement:** profiled using Calculon
- **Baseline Configuration:**
  - NVLink (scale-up) + InfiniBand (scale-out)
- **SiPAM Configuration:**
  - SiPAC network
  - Optimized # GPUs, memory capacity and bandwidth



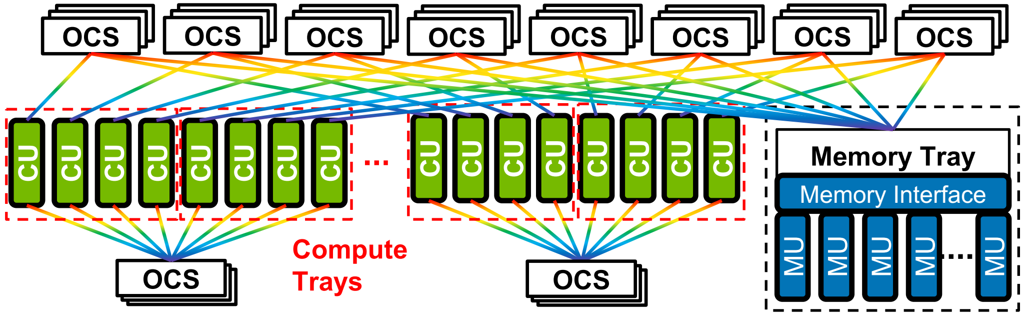
Baseline Configuration:

- NIC: up to **800 Gbps** / GPU



- NVL Domain: up to 72 GPUs

SiPAM Configuration:



[6] Isaev, Mikhail, et al. "Calculon: a methodology and tool for high-level co-design of systems and large language models." SC 2023.

Evaluation Setup – Calculon [6]

Hardware – Nvidia GPU Based

Single CU	FP16 TFLOPs	Mem Cap (GB)	Mem BW (TBps)
Nvidia A100	312	40	1.5
Nvidia H100	1000	80	3
Nvidia B100	3500	192	8
SiPAM*	3500	Up to 720	Up to 30

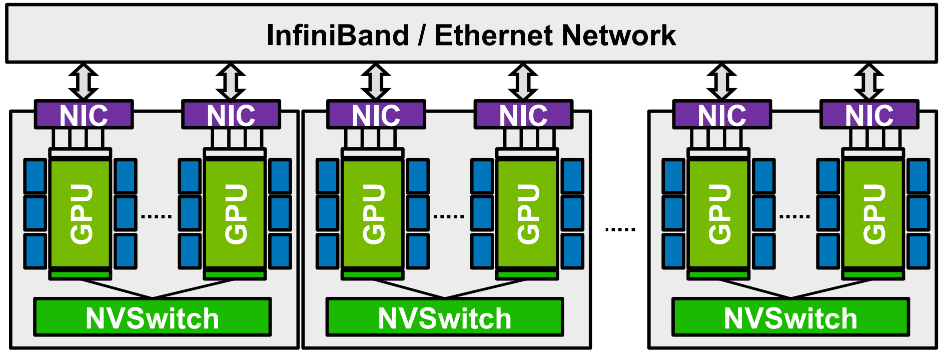
- ❖ **Cluster Size:** up to 1024 GPUs

Cluster of 1024 CUs	FP16 PFLOPs	Mem Cap (TB)	Mem BW (PBps)
Nvidia A100	320	41	1.5
Nvidia H100	1024	82	3.1
Nvidia B100	3584	197	8.2
SiPAM*	3584	Up to 737	Up to 31

* Assuming B100 as CU and HBM3E as MU

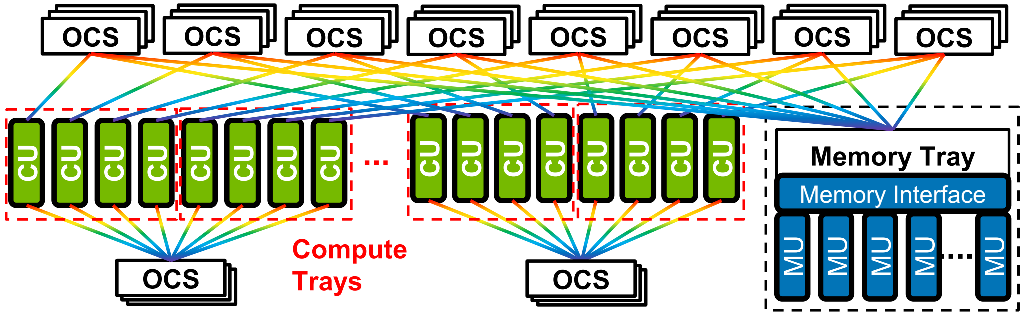
Baseline Configuration:

- NIC: up to **800 Gbps** / GPU



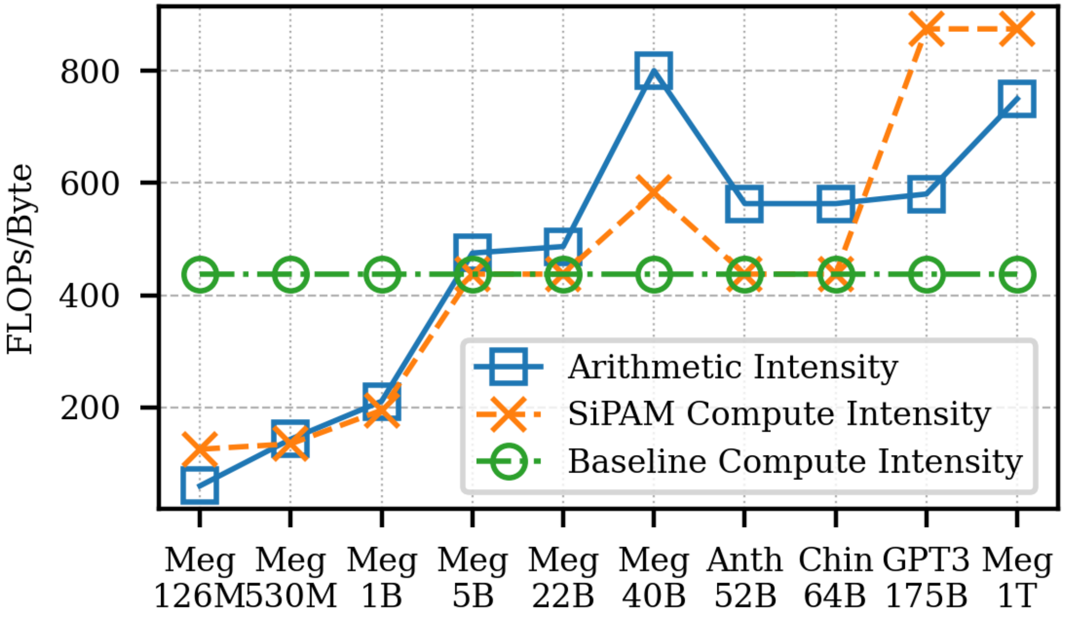
- ❖ NVL Domain: up to 72 GPUs

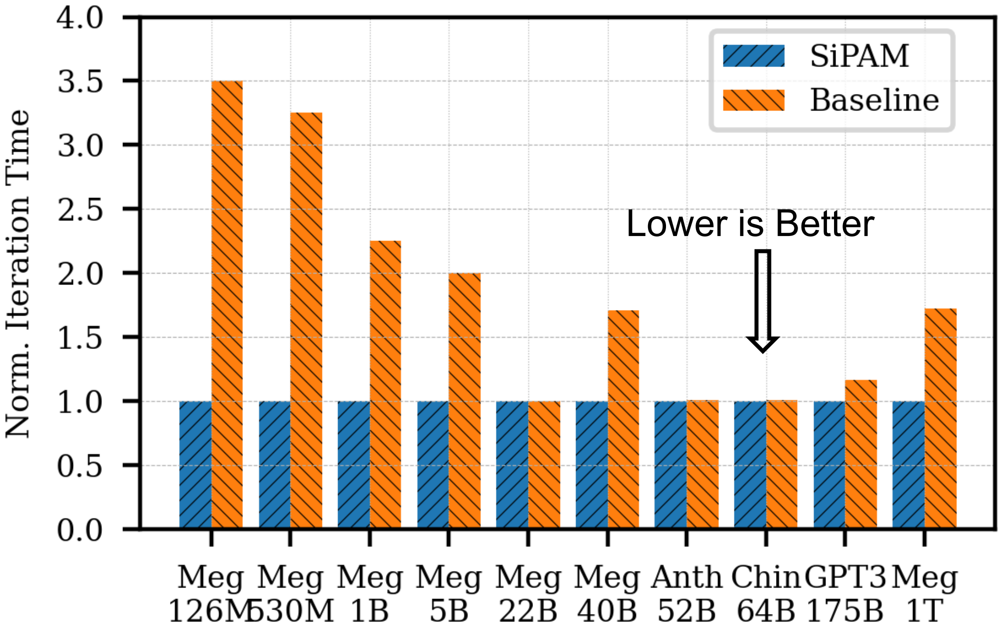
SiPAM Configuration:



Simulation Results - Training

- ❖ **Workloads:** Megatron-126M/5B/22B/40B/1T, Anthropic 52B, Chichilla-64B, GPT3-175B (**Training**)
- ❖ **Baseline:** Up to 256 B100 GPUs each with fixed 192 GB HBM memory @ 8 TBps total memory bandwidth
- ❖ **SiPAM:** Up to 256 GPUs, with compute, memory bandwidth, and capacity optimized based on each workload



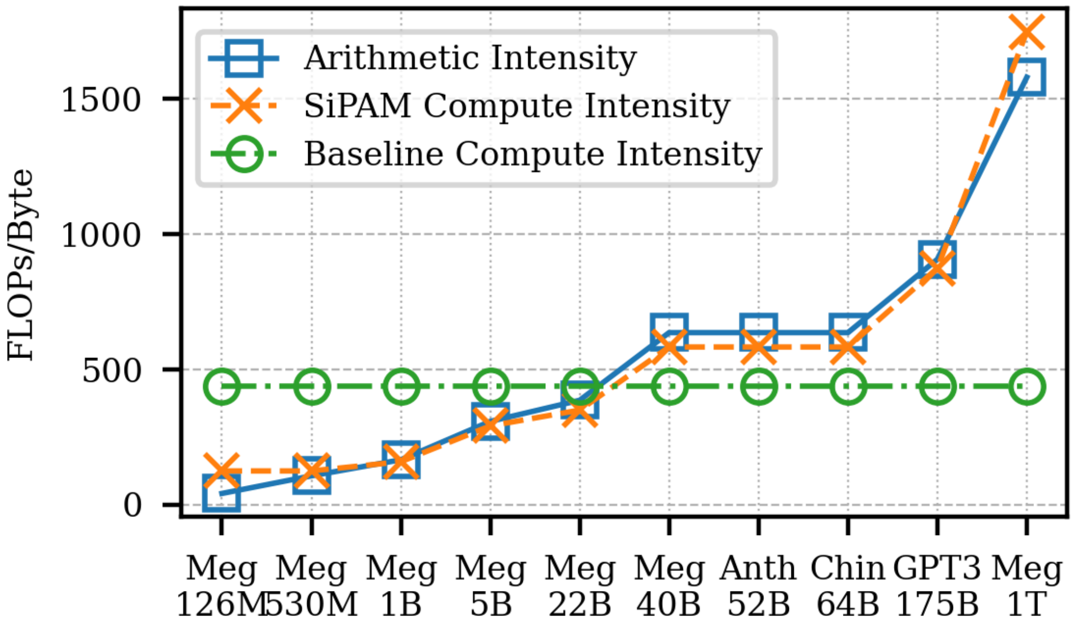


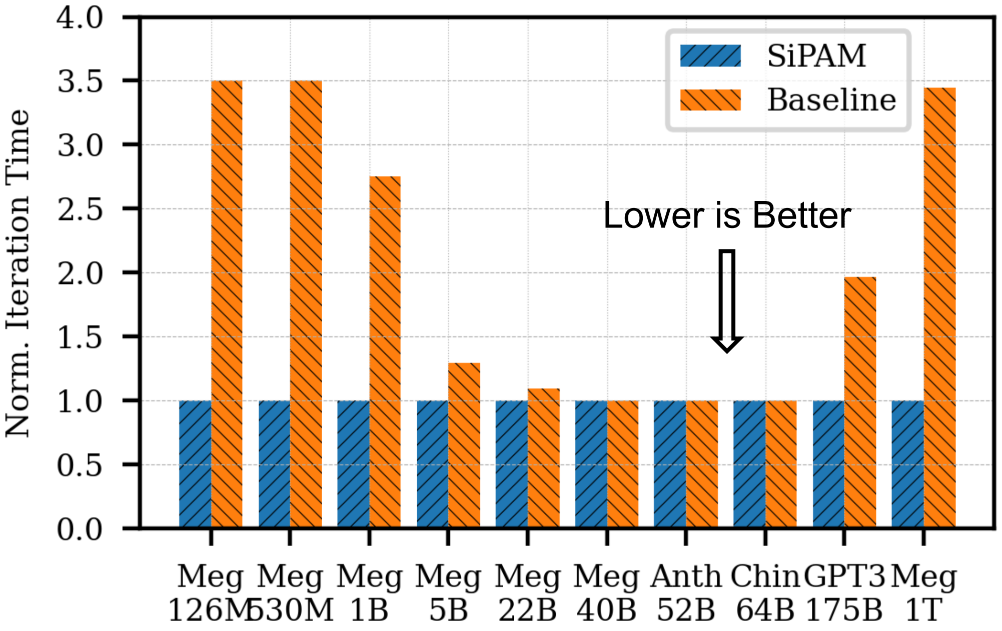
a) SiPAM tracks arithmetic intensity closely, while the baseline remains constant

b) SiPAM improves training time by up to **3.5x**

Simulation Results - Inference

- **Workloads:** Megatron-126M/5B/22B/40B/1T, Anthropic 52B, Chichilla-64B, GPT3-175B (**Inference**)
- **Baseline:** Up to 64 B100 GPUs each with fixed 192 GB HBM memory @ 8 TBps total memory bandwidth
- **SiPAM:** Up to 64 GPUs, with compute, memory bandwidth, and capacity optimized based on each workload

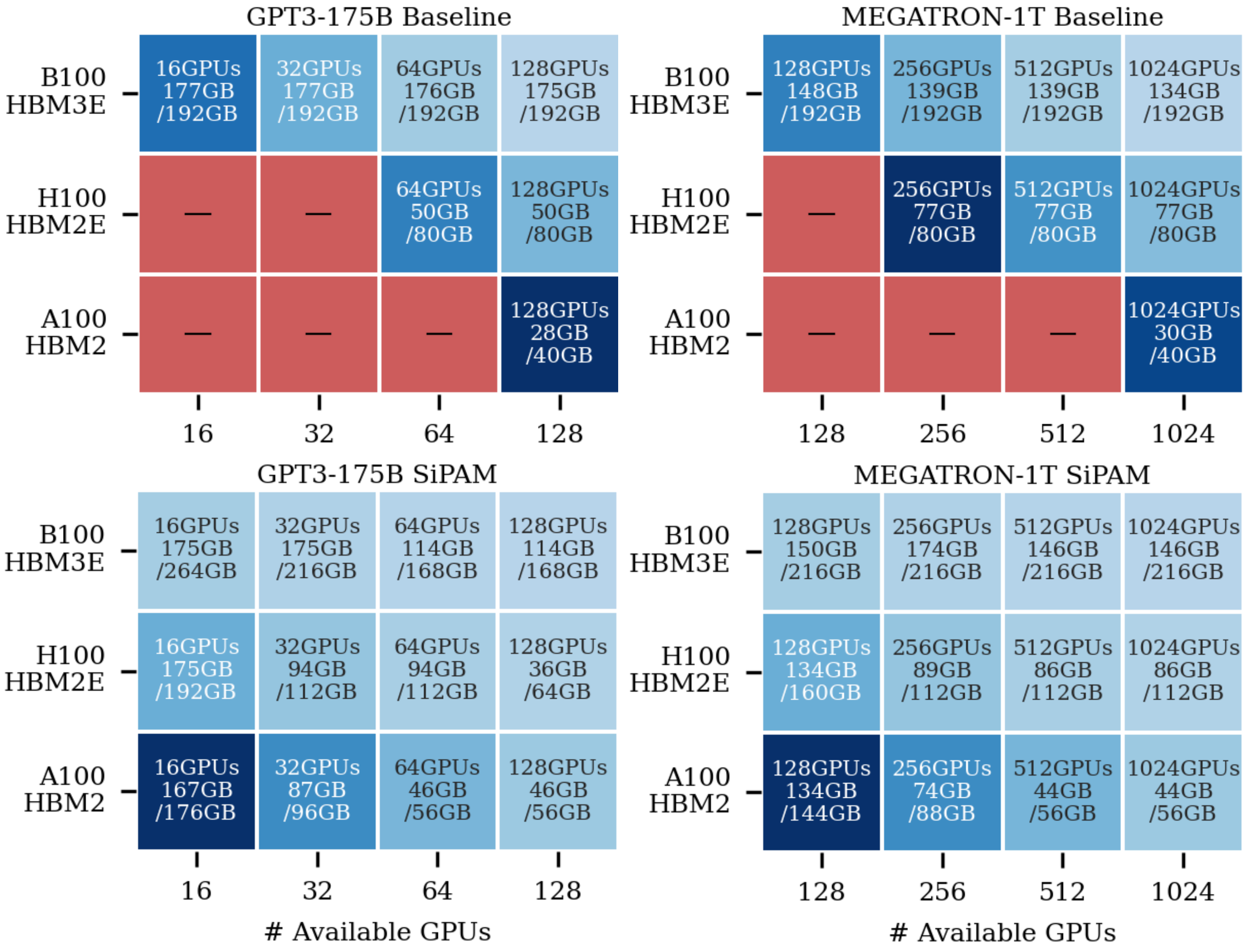




a) SiPAM tracks arithmetic intensity closely, while the baseline remains constant

b) SiPAM improves inference time by up to **3.5x**

Performance Under Limited Compute Resource



GPT3-175B Baseline

	16	32	64	128
B100 HBM3E	16GPUs 177GB /192GB	32GPUs 177GB /192GB	64GPUs 176GB /192GB	128GPUs 175GB /192GB
H100 HBM2E	—	—	64GPUs 50GB /80GB	128GPUs 50GB /80GB
A100 HBM2	—	—	—	128GPUs 28GB /40GB

MEGATRON-1T Baseline

	128	256	512	1024
B100 HBM3E	128GPUs 148GB /192GB	256GPUs 139GB /192GB	512GPUs 139GB /192GB	1024GPUs 134GB /192GB
H100 HBM2E	—	256GPUs 77GB /80GB	512GPUs 77GB /80GB	1024GPUs 77GB /80GB
A100 HBM2	—	—	—	1024GPUs 30GB /40GB

GPT3-175B SiPAM

	16	32	64	128
B100 HBM3E	16GPUs 175GB /264GB	32GPUs 175GB /216GB	64GPUs 114GB /168GB	128GPUs 114GB /168GB
H100 HBM2E	16GPUs 175GB /192GB	32GPUs 94GB /112GB	64GPUs 94GB /112GB	128GPUs 36GB /64GB
A100 HBM2	16GPUs 167GB /176GB	32GPUs 87GB /96GB	64GPUs 46GB /56GB	128GPUs 46GB /56GB

Available GPUs

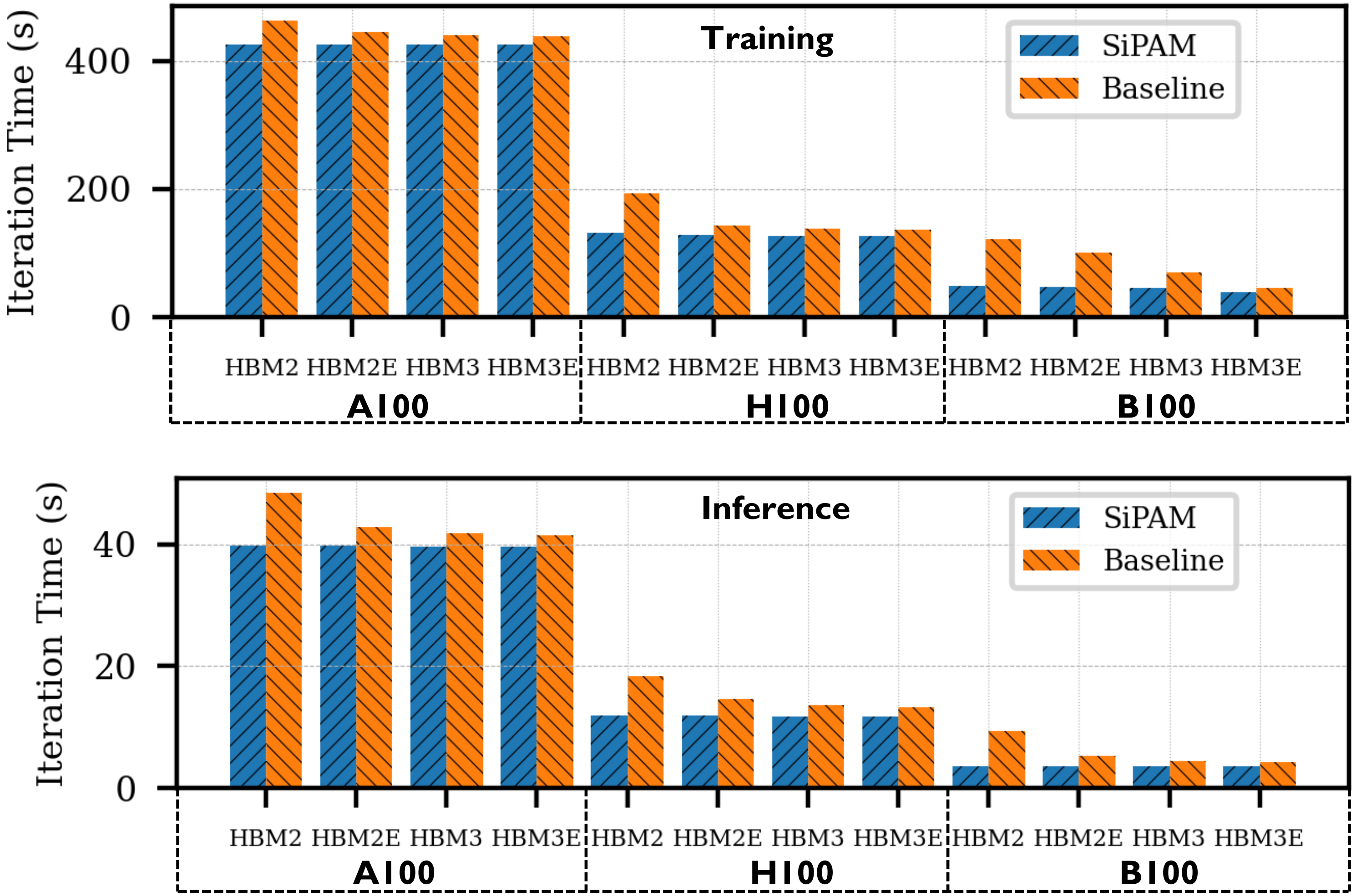
MEGATRON-1T SiPAM

	128	256	512	1024
B100 HBM3E	128GPUs 150GB /216GB	256GPUs 174GB /216GB	512GPUs 146GB /216GB	1024GPUs 146GB /216GB
H100 HBM2E	128GPUs 134GB /160GB	256GPUs 89GB /112GB	512GPUs 86GB /112GB	1024GPUs 86GB /112GB
A100 HBM2	128GPUs 134GB /144GB	256GPUs 74GB /88GB	512GPUs 44GB /56GB	1024GPUs 44GB /56GB

Available GPUs

- **Configurations**:
  - Workload: GPT3-175B and Megatron-1T
  - Red cells: no feasible parallelization strategy
  - Darker color: higher iteration time
- **Takeaway**: SiPAM consistently enables feasible deployment of larger models under constrained GPU resources
  - SiPAM flexibly allocates memory capacity and bandwidth.
  - For a fixed GPU-HBM combination, iteration time decreases as the number of available GPUs increases.
  - For a fixed number of GPU, newer GPU generations yield lower iteration time.

Compute & Memory Technology Scaling

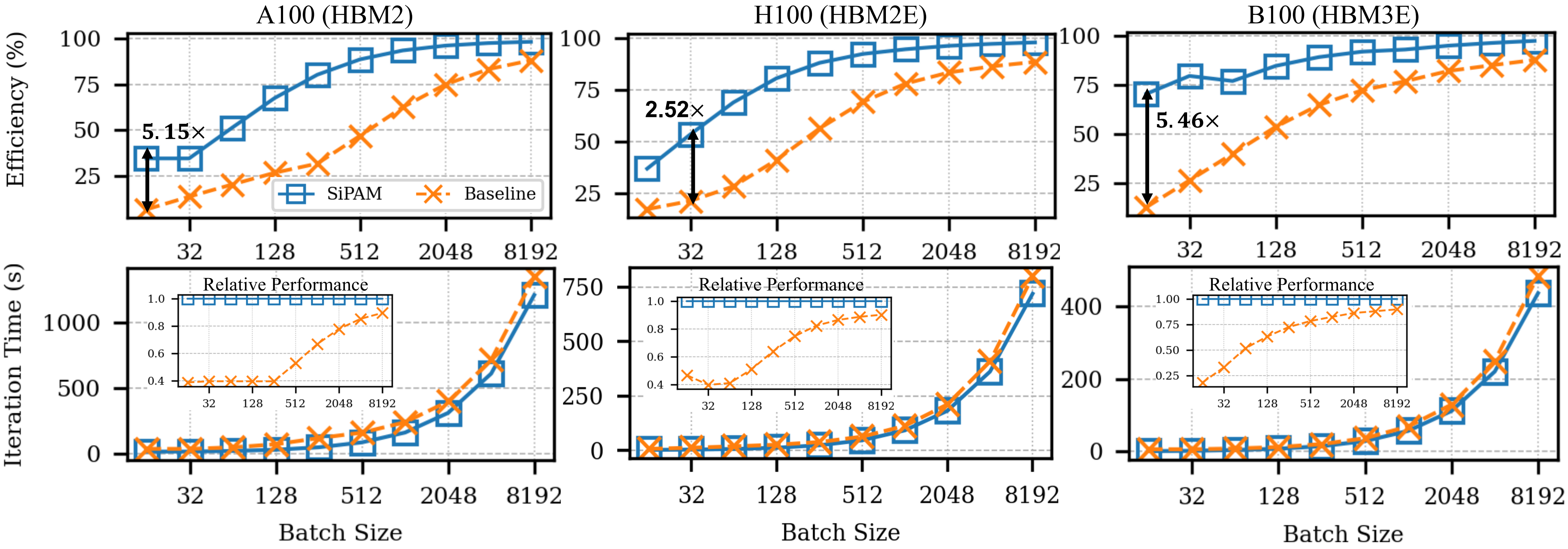


- **Configurations**:
  - Workload: GPT3-175B
  - Network size: 128 GPUs
  - Cross GPU-HBM pairing
- **Takeaway**: SiPAM consistently outperforms the baseline by allocating the needed compute and memory resources
  - Newer GPU generations outperform earlier ones.
  - For each GPU generation, performance improves as memory generation advances.

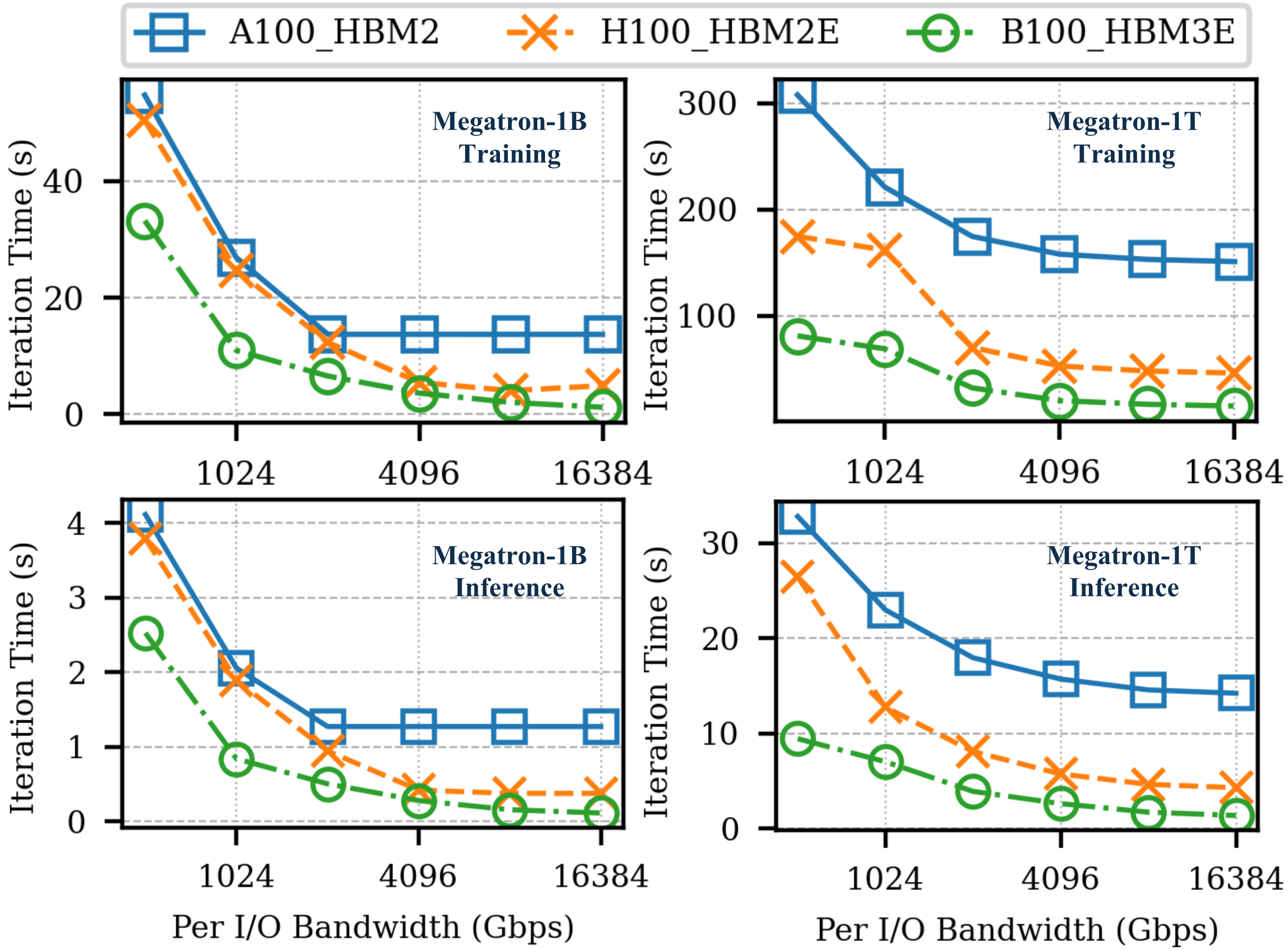
System Efficiency Analysis

- ❖ **Configurations**:
  - ❖ Workload: Megatron-1T
  - ❖ Metrics: System Efficiency & Iteration Time (inset shows relative performance)
- ❖ SiPAM consistently outperforms the baseline in both efficiency and iteration time.

System Efficiency = $\frac{T_{compute}}{T_{total}}$



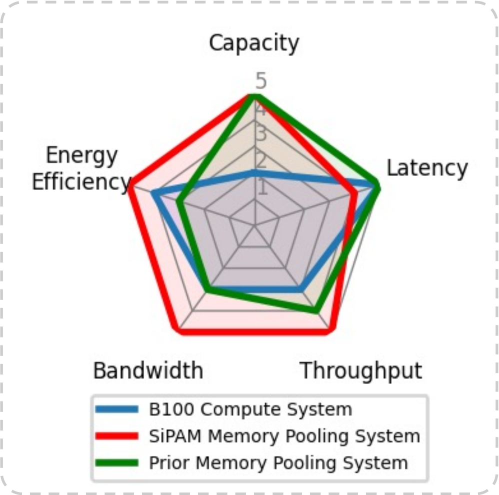
SiP I/O Bandwidth Scaling



- ❖ **Configurations**:
  - ❖ Workload: Megatron-1T & Megatron-1T
  - ❖ Per I/O Bandwidth: 512 Gbps to 16 Tbps
  - ❖ Total injection bandwidth / GPU:
    - ❖ A100/H100: 6 Tbps to 192 Tbps
    - ❖ B100: 7.6 Tbps to 240 Tbps
- ❖ **Takeaway**: Newer GPU generations with higher compute capability require greater memory bandwidth to achieve continued performance scaling.
- ❖ A100 on Megatron-1B: performance plateaus
  - ❖ Compute throughput becomes saturated

Conclusion

- **Problems addressed:** memory capacity & bandwidth bottlenecks in AI/ML
- **Design:**
  - Direct photonic integration along the perimeter of the compute die
  - Unified high-bandwidth communication domain
- **Optimization:** co-designed roofline-model based allocation algorithm
- **Results:**
  - Showed up to 3.5x faster iteration time
  - Highlights the critical need for bandwidth scaling in next-generation compute.



- **Future Works:**
  - Cost and power modeling
  - Capture network demand in addition to memory demand

Acknowledgement

Semiconductor Research Corporation

CUbiC
Center for Ubiquitous Connectivity

arpa·e

PINE

LPS LABORATORY FOR PHYSICAL SCIENCES

arm

IBM

ANALOG DEVICES
AHEAD OF WHAT'S POSSIBLE™

SK hynix

Micron

BOEING

EMD Electronics

HRL LABORATORIES

tsmc

qorvo

NSF

intel

DARPA

GlobalFoundries™

Raytheon Technologies

SAMSUNG

MEDIATEK