Exploring High-Bandwidth Flash for Modern LLM Inference: Opportunities and Challenges

SAFARI Live Seminar

Jisung Park

POSTECH CAOS
COMPUTER ARCHITECTURE & OPERATING SYSTEMS LABORATORY

2026.07.24

Brief Self-Introduction

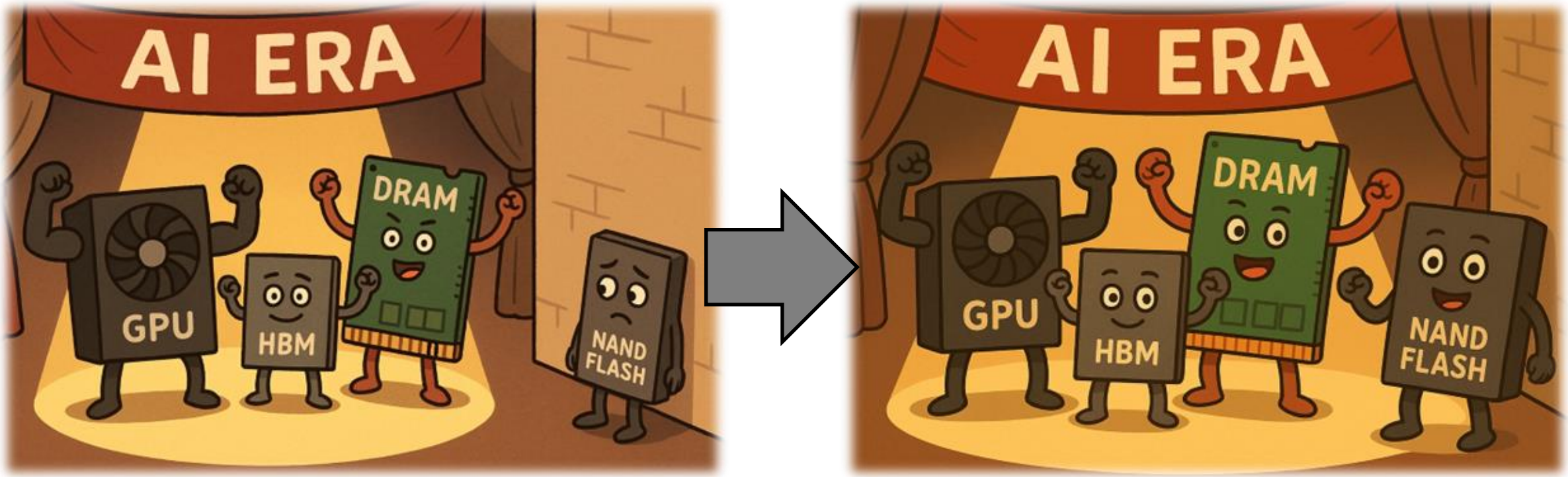
- Dr. Jisung Park
  - Assistant Professor @ POSTECH CSE, 2022.09 – Present
  - Postdoctoral Research Associate and Lecturer, ETH Zurich, 2019 – 2022
  - Ph.D. from SNU EECS, 2019
  - Homepages
    - Research Group – CAOS (Computer Architecture and Operating Systems) Laboratory: http://caos.postech.ac.kr
    - Personal: https://jisung-park.github.io/
  - Email: jisung.park@postech.ac.kr

- Research Interests
  - Computer Architecture and System Software
  - Memory and Storage Systems
  - System Security
  - HW/SW Cooperation
  - Application-Specific System Design (e.g., Bioinformatics and AI)

Research on NAND Flash Memory (Selected)

- Performance Optimizations
  - RPS: Relaxed Program Sequence – DAC 2016
  - P&AR²: Pipelined & Adaptive Read Retry – ASPLOS 2021
  - Flash-Cosmos: In-Flash Bulk Bitwise Operation – MICRO 2022
  - Experimental Study on Performance Impact of Read Disturbance – SIGMETRICS 2026
- Cost-effectiveness Optimizations
  - DAC: Dedup-Assisted Compression – DATE 2017
  - DeepSketch: ML-Assisted Delta Compression – USENIX FAST 2022
  - AERO: Adaptive Erase Operation – ASPLOS 2024
  - Norns: Learning-Based Post-Fabrication Optimization – ICCAD 2024
- Security & Privacy
  - RansomBlocker: Ransomware-Proof SSD – DAC 2019
  - Evanesco: Low-Cost Data Sanitization – ASPLOS 2020
- +14 publications in ISCA, ASPLOS, MICRO, HPCA, DAC, and DATE

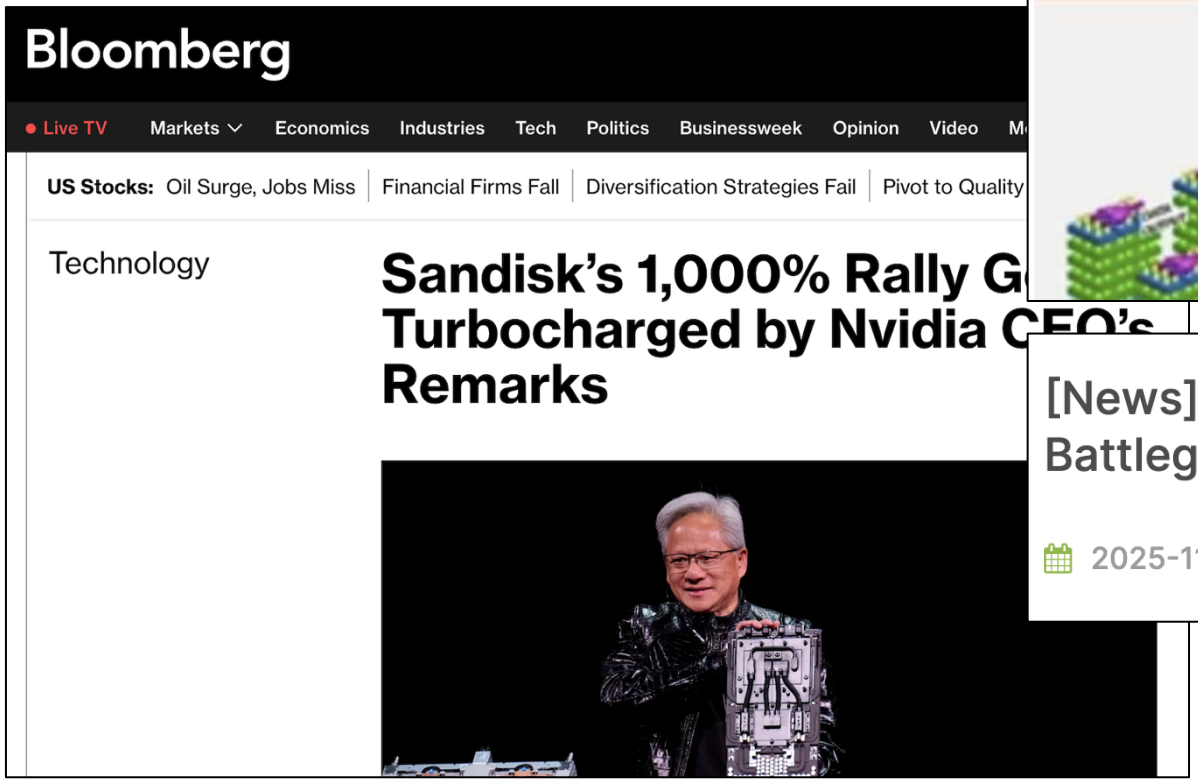
NAND Flash in AI Era

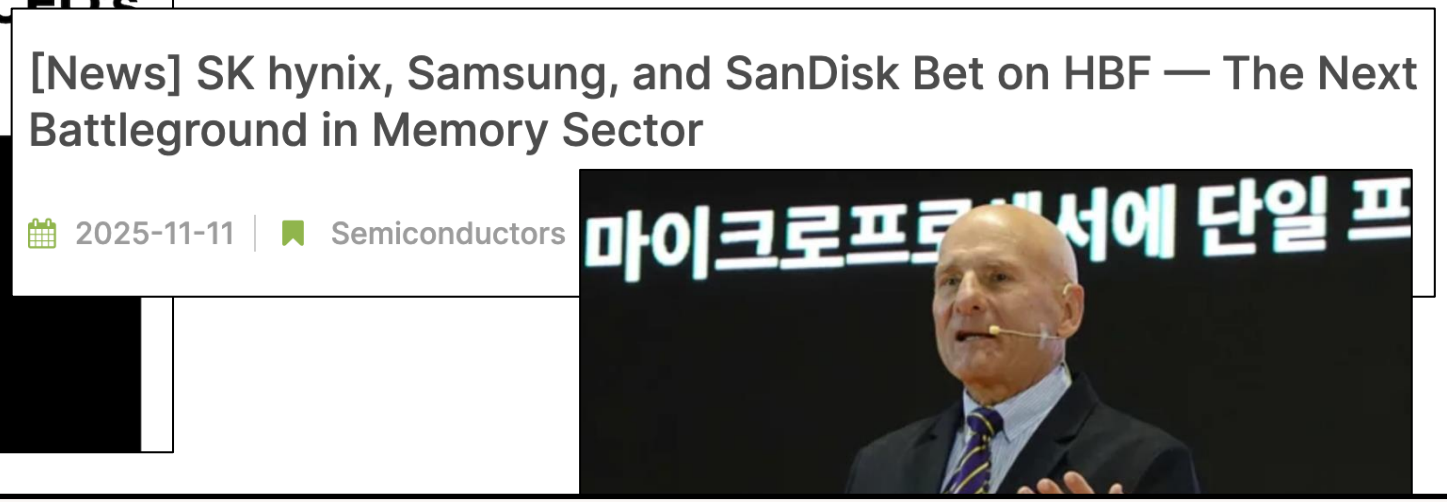


- NAND flash has received far less attention in (AI) systems research
- This is rapidly changing as AI systems face explosive data growth

Growing Interest in NAND Flash

- Industry is paying increasing attention to importance and new possibilities





Now is the time for academia to explore potential and address challenges

Exploring HBF for LLM-Serving Systems

- D. Son et al., "Exploring High-Bandwidth Flash for Modern LLM Inference: Opportunities and Challenges," IEEE Computer Architecture Letters (IEEE CAL), 2026 (early access).

Exploring High-Bandwidth Flash for Modern LLM Inference: Opportunities and Challenges

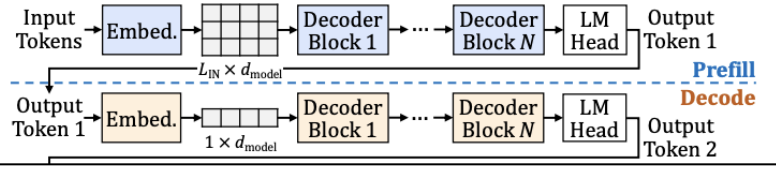
Dowon Son, Yonggon Park, Hyunuk Cho, Hyungkyu Ham, Onur Mutlu, Sungjin Lee, Gwangsun Kim, and Jisung Park

***Abstract*—This work investigates the potential benefits and technical challenges of using high-bandwidth flash (HBF) for large language model (LLM) inference. HBF has gained increasing attention as a promising solution to mitigate memory-capacity bottlenecks in modern LLM-serving systems, but its benefits and challenges remain largely uninvestigated. To address this gap, we thoroughly analyze HBF-based LLM-serving systems under diverse system configurations and operating scenarios in which HBF serves as a main GPU-memory component to handle both reads and writes. Our analysis shows that, despite its limited write performance, HBF can significantly improve the batch size, throughput, and flexibility of LLM-serving systems while reducing the minimum GPU requirements, but realizing these benefits critically depends on sustaining HBM-comparable read bandwidth and requires significant endurance improvements.**

***Index Terms*—High-bandwidth flash, LLM-serving systems**

II. Background and Motivation

LLM Inference. Fig. 1(a) depicts the overall process of LLM inference, which consists of two phases: (*i*) *prefill* and (*ii*) *decode*. Both phases pass through the same token-embedding layer, decoder blocks, and language-model (LM) head. Given a query with L_{IN} tokens, the prefill phase processes all tokens in parallel to generate the first output token. The decode phase then generates L_{OUT} tokens *auto-regressively*, using each output token as a new input token for the next iteration.



Executive Summary

- Motivation: High-bandwidth flash (HBF) is gaining significant attention as a promising alternative or complement to high-bandwidth memory (HBM)
- Goal: Understand HBF's potential benefits and technical challenges
  - In modern LLM-serving systems
  - When used as the main GPU-memory component
- Methodology: Simulation of HBF-enabled LLM-serving systems
  - Using recent large-scale LLMs
  - While varying GPU count, SLO constraint, and context length
- Ten key observations and five takeaways, including
  - HBF can significantly improve the batch size, throughput, and system flexibility
  - Providing larger capacity alone is insufficient; requires (1) HBM-comparable read bandwidth and (2) significant endurance enhancement (+ write performance)

Outline

- Background & Motivation
- Methodology
- Evaluation Results
- Conclusion

Large Language Models (LLMs)

- Predicts the next token from the given context
  - Statistically based on patterns learned from training datasets
  - Autoregressively by using a new output token as an input for the next token

- Requires access to and computation on massive amounts of data
  - For every token
  - Increasing amount of model weights
  - KV cache also increasing w/ context length

    → e.g., Generating the 100,000th token in Llama-3 405B model requires reading ~800-GB data (756-GB weights + 50-GB KV cache)

LLM inference: Significant memory requirements in both capacity and bandwidth

Memory Requirements of LLMs

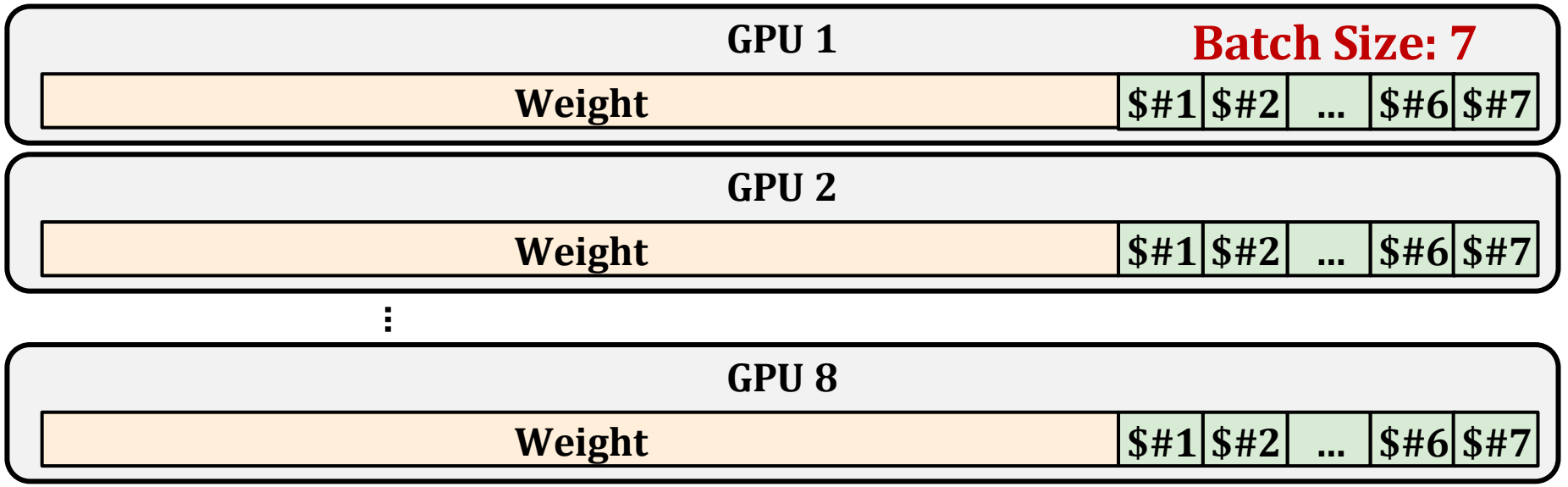
- Llama-3 405B case
  - Model weights ≈ 750 GB
  - Max. KV cache per query ≈ 50 GB (128K context length)

NVIDIA H200
- 141-GB capacity
- 4.8-TB/s bandwidth (HBM3e)

- Capacity Requirement: > 800 GB
  - Including memory space for activations and compute buffers
- Bandwidth Requirement: > 8 TB/s
  - Service-Level Objective (SLO) = 10 Tokens Per Second (TPS)
    = 0.1-second Time Per Output Token (TPOT)
  - Max. amount of data to read per token ≈ 800 GB

Today's Solution: Scale Up GPUs

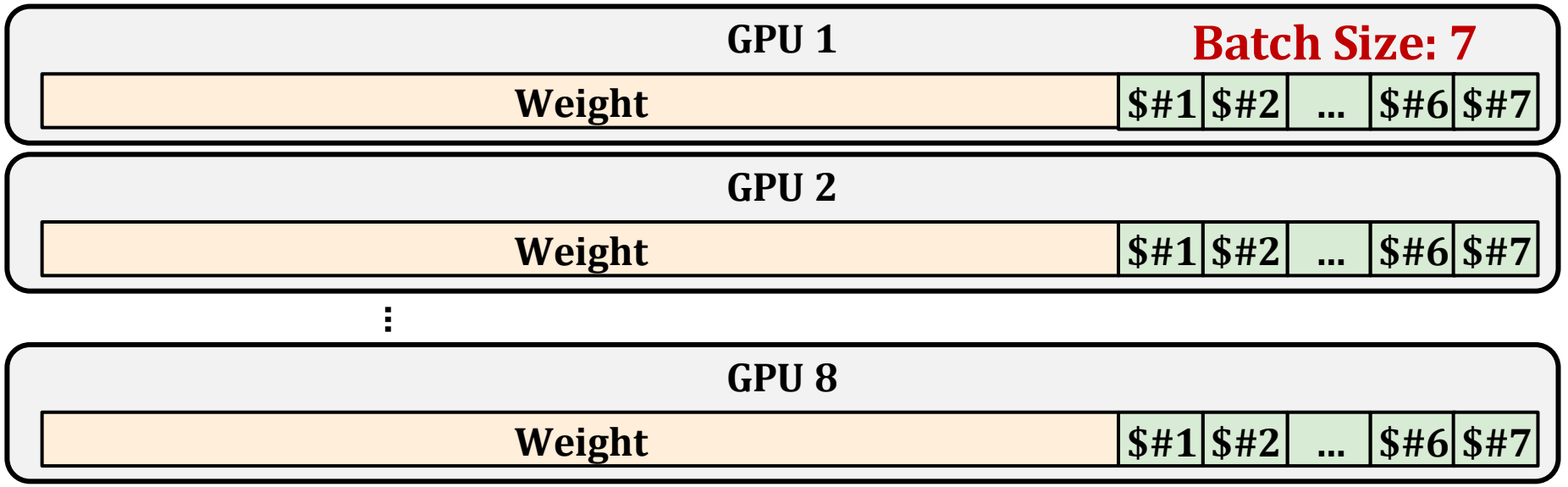
- Aggregate HBM capacity and bandwidth
  - HGX H200 (8 GPUs)
  - Capacity: 141×8 = 1,128 GB > 800 GB
  - Bandwidth: 4.8×8 = 38.4 TB/s > 8 TB/s
- Partition a single large model across GPUs
  - Relying on high-bandwidth interconnects (e.g., NVLink)



Today's Solution: Scale Up GPUs

- Aggregate HBM capacity and bandwidth
  - HGX H200 (8 GPUs)
  - Capacity: 141×8 = 1,128 GB > ~~800 GB~~ 1,100 GB
  - Bandwidth: 4.8×8 = 38.4 TB/s > ~~8 TB/s~~ 11 TB/s
- Partition a single large model across GPUs
  - Relying on high-bandwidth interconnects (e.g., NVLink)

HBM bandwidth is largely underutilized



Today's Solution: Scale Up GPUs

- Aggregate HBM capacity and bandwidth
  - HGX H200 (8 GPUs)
  - Relying on high-bandwidth interconnects (e.g., NVLink)

LLM inference: Significant memory requirements on both capacity and bandwidth

Current solution: Scaling memory capacity w/ multiple GPUs - costly and wasteful

NAND Flash-Based High-Bandwidth Memory

- SanDisk's High-Bandwidth Flash (HBF)

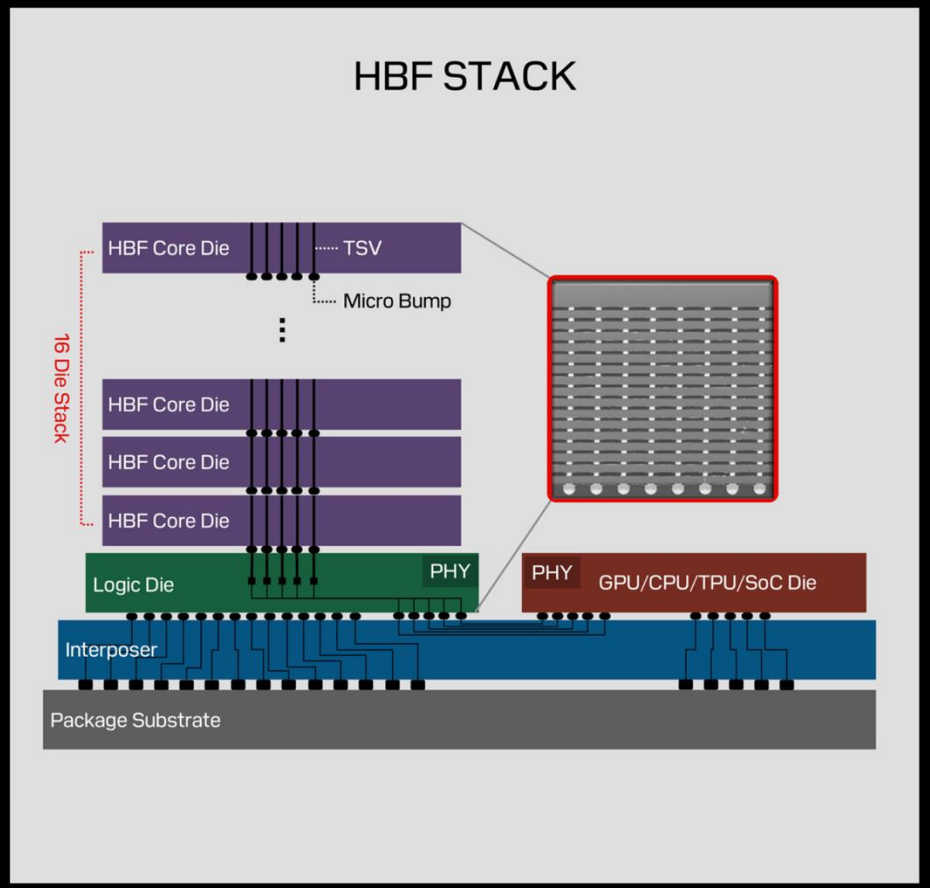
High Bandwidth Flash (HBF™)
Augmenting HBM Memory with NAND Flash for AI Inference Workloads

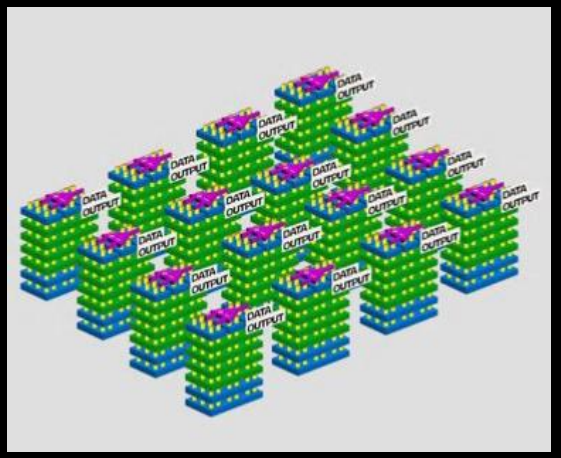
(→) Match HBM Bandwidth
Deliver 8-16X Capacity at Similar Cost

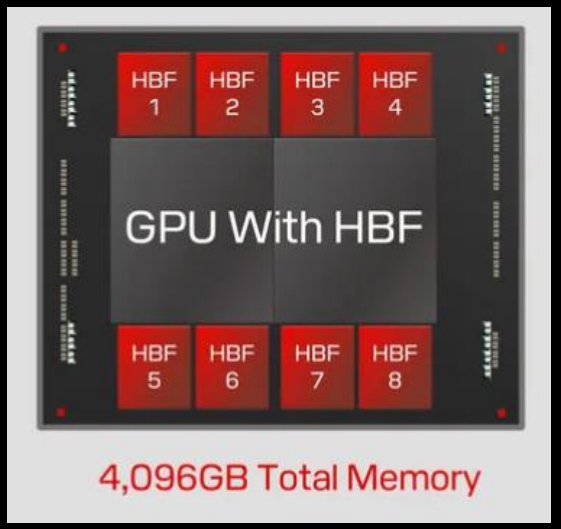
(→) Enabled by BiCS Technology
With CBA Wafer Bonding

(→) Proprietary Stacking Technology
Ultra-Low Die Warpage for 16H Stacking

(→) Architecture Developed Over the Past Year
With Inputs From Major AI Players

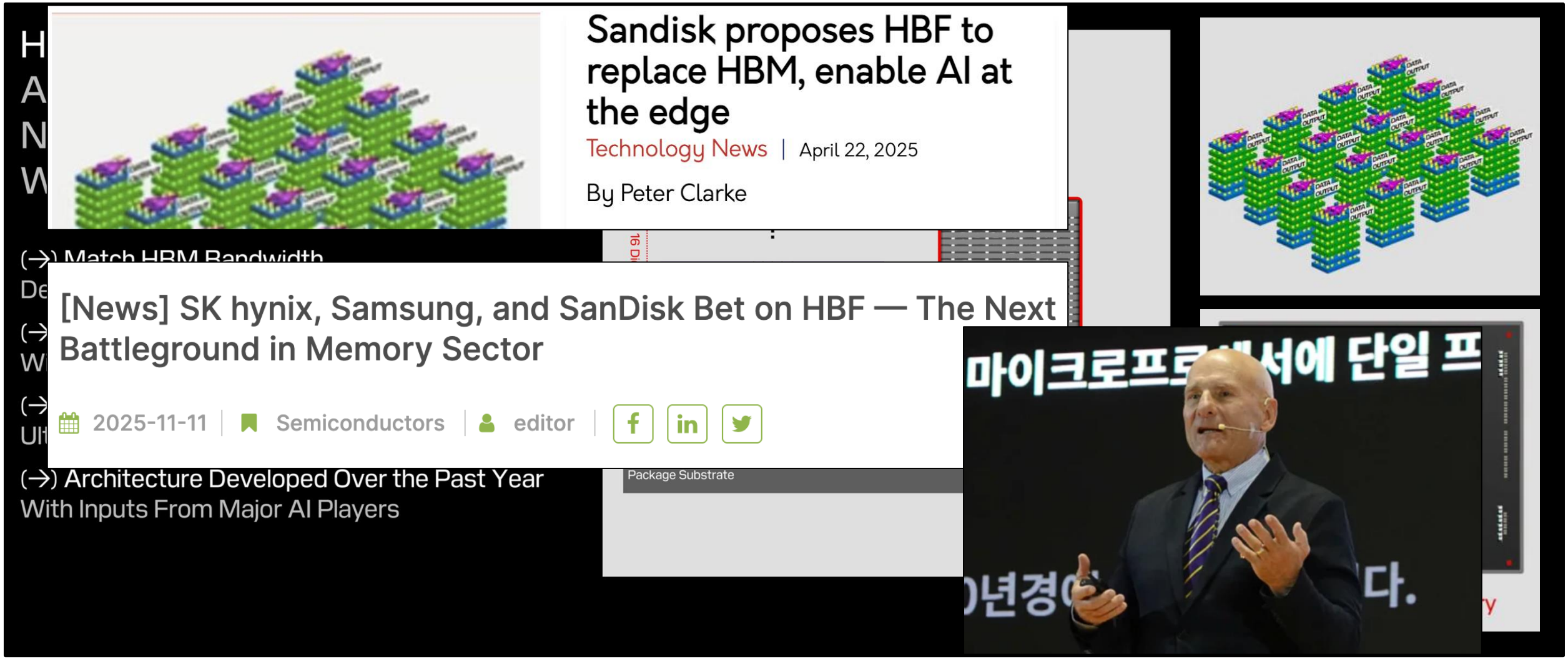






NAND Flash-Based High-Bandwidth Memory

- SanDisk's High-Bandwidth Flash (HBF)



NAND Flash-Based High-Bandwidth Memory

- SanDisk's High-Bandwidth Flash (HBF)

Sandisk's promise:
>10x capacity than & comparable bandwidth to HBM4

10% of GPUs or 10x throughput?:
We need a systematic analysis of the potential of HBF

Our Goal

Provide new insights on HBF's potential benefits and challenges through a systematic analysis of HBF-based LLM-serving systems under realistic, challenging workloads

Outline

- Background & Motivation
- Methodology
- Evaluation Results
- Conclusion

Methodology: Key Assumption

All HW components (GPU cores, memory stacks, and interconnects) operate at their specified peak performance

\+ Rapid exploration of HBF's potential benefits

\- How to achieve such performance? (future work)

Methodology: LLM-Serving Systems

- Focusing on decode nodes in disaggregated prefill-decode systems
  - Assuming high-performance prefill nodes for zero-stall batching
  - KV-cache transfer from prefill nodes to decode nodes

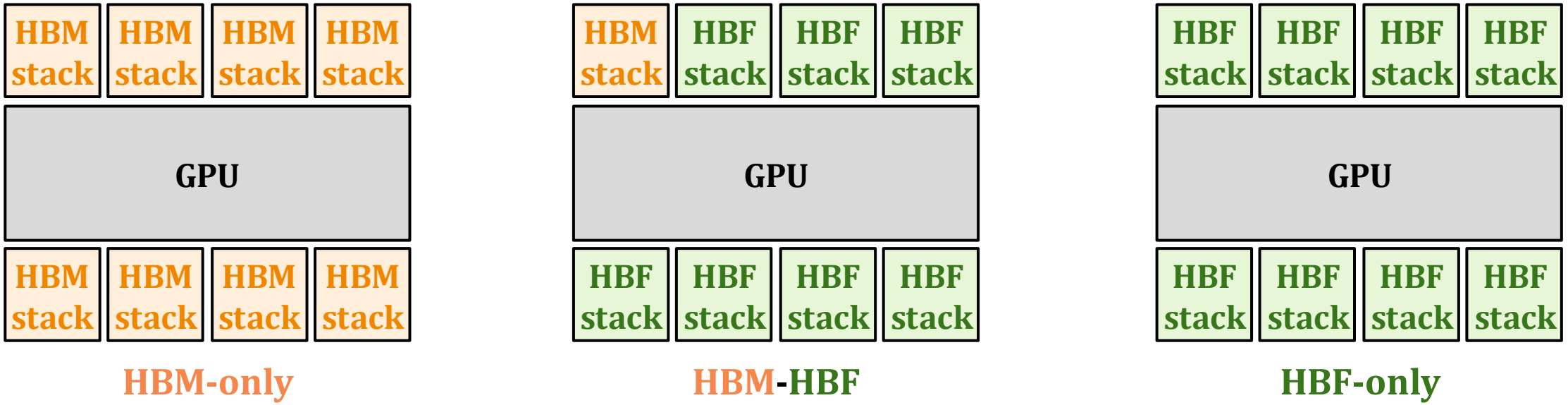
- Continuous batching, constrained by
  - GPU-memory capacity (no KV-cache spill)
  - SLO requirement (e.g., TPOT $\leq$ 0.1 s)
  - On-die SRAM capacity for HBF-only systems

- Using best parallelism scheme to maximize TPS (tokens per second)

- Modeled using an extended version of LLMSimulator [S. Yun+, MICRO'24]

GPU Configurations

- Three GPU-memory architectures
  - HBM-only: eight HBM stacks per GPU
  - HBM-HBF: one HBM stack and seven HBF stacks per GPU
    - HBM stack dedicated to intermediate data (not KV)
  - HBF-only: eight HBF stacks per GPU
    - Equipped with 40-MB SRAM on logic die to store intermediate data



Methodology: GPU Configurations (Cont'd)

- R100 GPU core (4 PFLOPS for BF16)
- Interconnects
  - Intra-node (~8 GPUs): 1,800 GB/s NVLink
  - Inter-node: 100 GB/s InfiniBand
- HBM stacks (HBM4)
  - Per-stack read bandwidth: 1.6 TB/s
  - Per-stack capacity: 36 GB
- HBF stacks (based on Sandisk's projection)
  - Per-stack read bandwidth: 1.6 TB/s
  - Per-stack capacity: 512 GB

Methodology: GPU Configurations (Cont'd)

- HBF model based on conventional low-latency NAND flash
  - It may not be easy to achieve the 1.6-TB/s per-stack bandwidth with NAND flash
  - e.g., Samsung Z-NAND and Kioxia XL Flash
    - Page size: 2 KiB and 4 KiB
    - Planes per die: 8 and 16
    - Page-read latency: 3 µs and 4 µs
    - Die bandwidth: ~5 GB/s and ~16 GB/s
    - Stack bandwidth: ~80 GB/s and ~256 GB/s (16 dies per stack)
  - Achieving target bandwidth requires 27 planes per die even with 1-µs latency
  - Optimistic modeling of conventional NAND-based HBF
    - 4-KiB page, 16 planes per die, 3-µs latency
      → ~350-GB/s per-stack bandwidth

Methodology: GPU Configurations (Cont'd)

- Five LLM-serving systems w/ different GPU-memory configurations

- HBM4 (HBM-only), HBF and CONV (HBM-HBF), HBF$^+$ and CONV$^+$ (HBF-only)
  - Assume 100-µs program latency (from Z-NAND) for all HBF configurations

GPU	# Stacks (HBM/HBF)	Capacity [GB]	RD BW [TB/s] (HBM/HBF)	WR BW [TB/s] (HBM/HBF)
HBM4	8/0	288	12.8/-	12.8/-
HBF	1/7	3,620	1.6/11.2	1.6/0.112
HBF$^+$	0/8	4,096	-/12.8	-/0.128
CONV	1/7	3,620	1.6/2.45	1.6/0.073
CONV$^+$	0/8	4,096	-/2.80	-/0.084

Methodology: Workloads

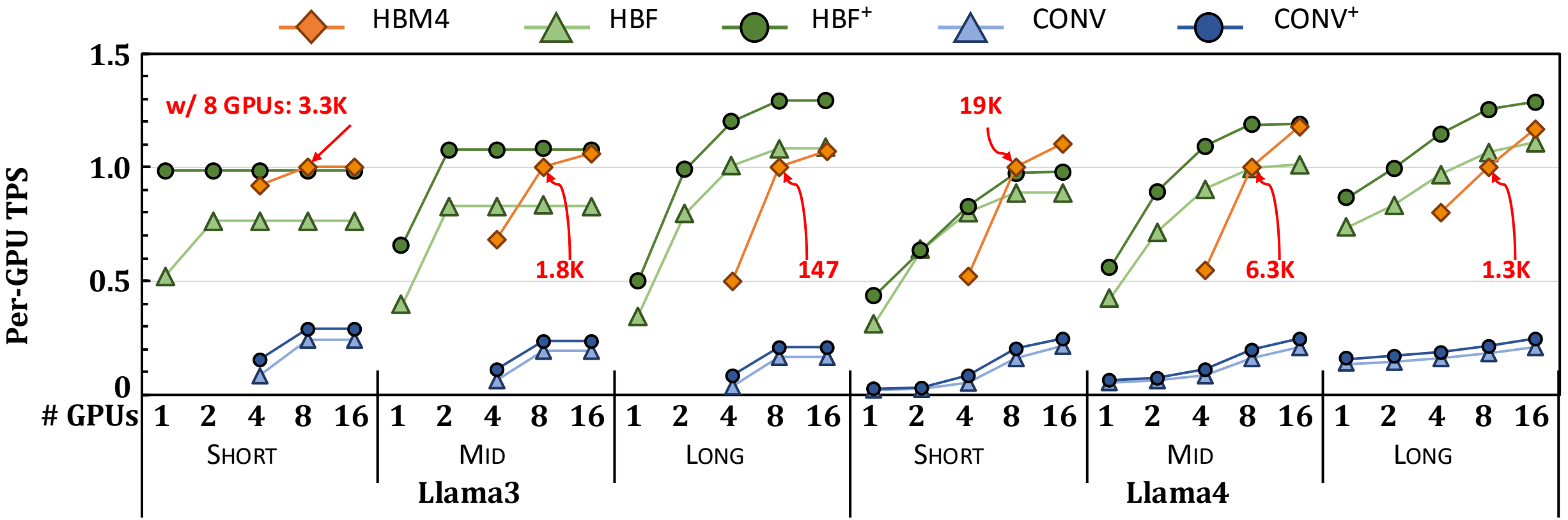
- Target models: Llama 3 405B and Llama 4 Maverick
  - Llama 3 405B (dense): 756-GB weights
  - Llama 4 Maverick (MoE): 746-GB weights

- Three workloads w/ different $\langle L_{IN}, L_{out} \rangle$ based on real-world benchmarks
  - SHORT (ShareGPT): $\langle 1660, 373 \rangle$
  - MID (LongBench): $\langle 5.9K, 499 \rangle$
  - LONG (English summarization): $\langle 103.5K, 1.1K \rangle$

Outline

- Background & Motivation
- Methodology
- **Evaluation Results**
- Conclusion

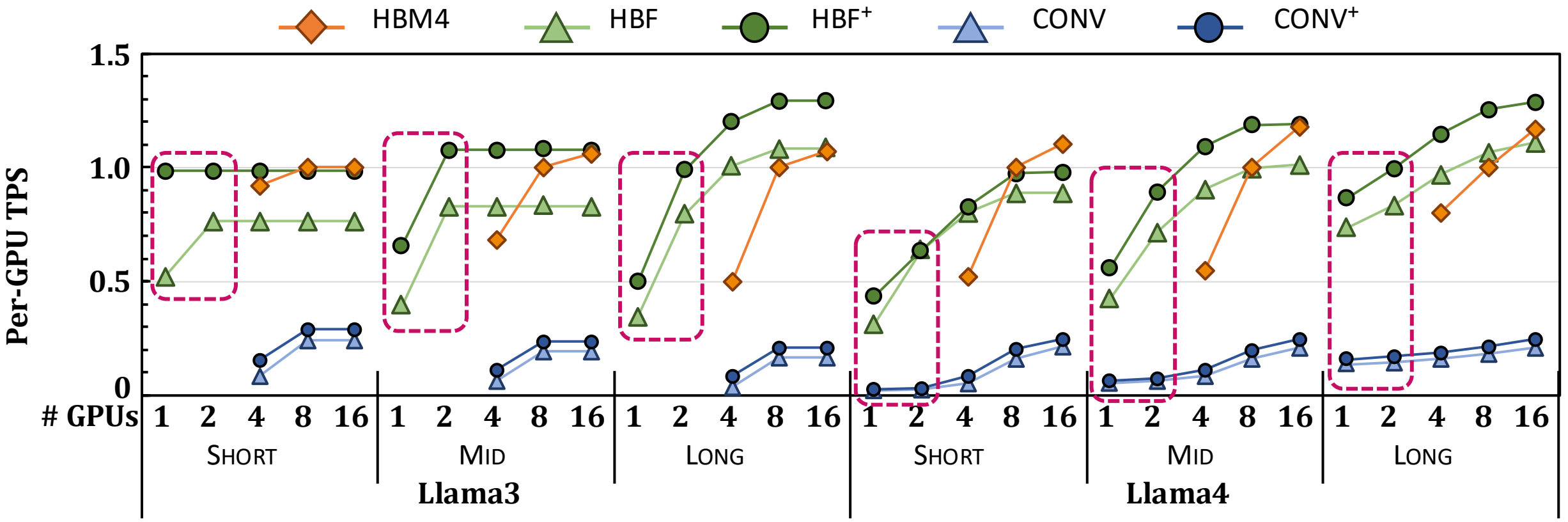
Evaluation Results: System Throughput



Per-GPU TPS under 0.1-second TPOT SLO

Normalized to HBM4 w/ 8 GPUs: Significantly higher TPS in MoE models

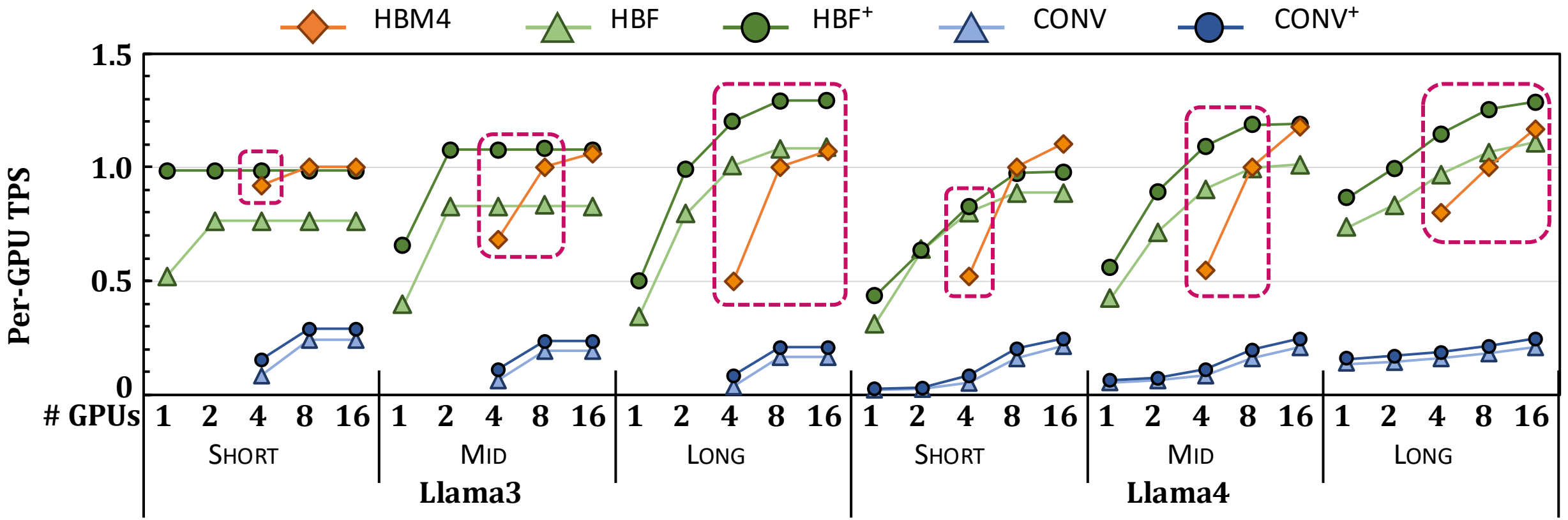
Evaluation Results: System Throughput



Per-GPU TPS under 0.1-second TPOT SLO

Benefit: HBF reduces the minimum GPU count to support the given models

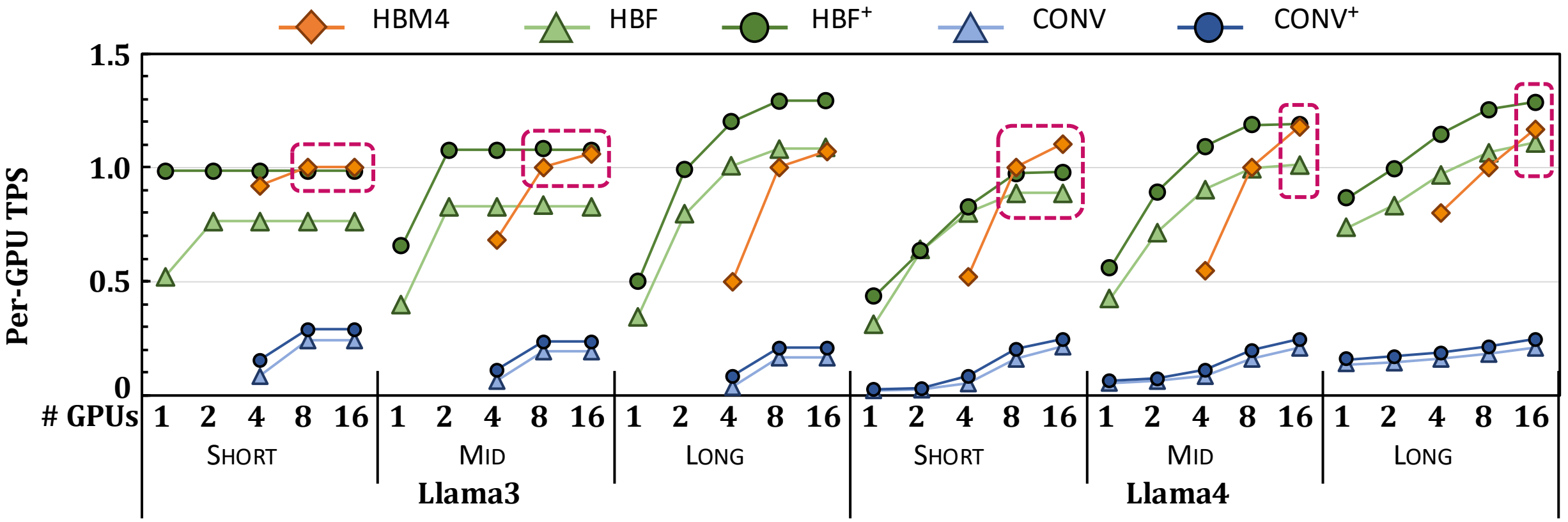
Evaluation Results: System Throughput



Per-GPU TPS under 0.1-second TPOT SLO

Benefit: HBF can improve system throughput at the same GPU count: larger improvements for longer-context workloads

Evaluation Results: System Throughput

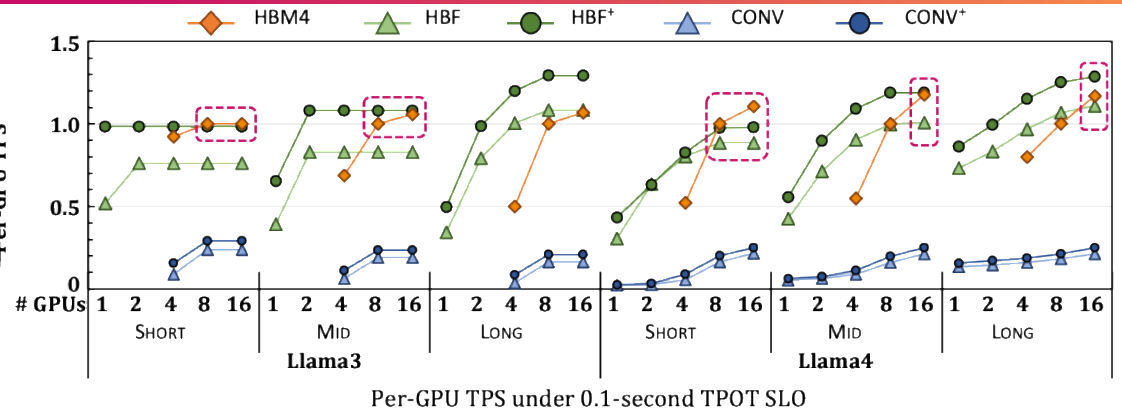


Per-GPU TPS under 0.1-second TPOT SLO

Diminishing improvements as GPU count increases
→ Is HBF effective only for small-scale systems?

Evaluation Results: Diminishing Gains w/ more GPUs

- GPU scaling
  → Larger aggregate capacity & bandwidth
  → Overcome the memory wall (at high cost)

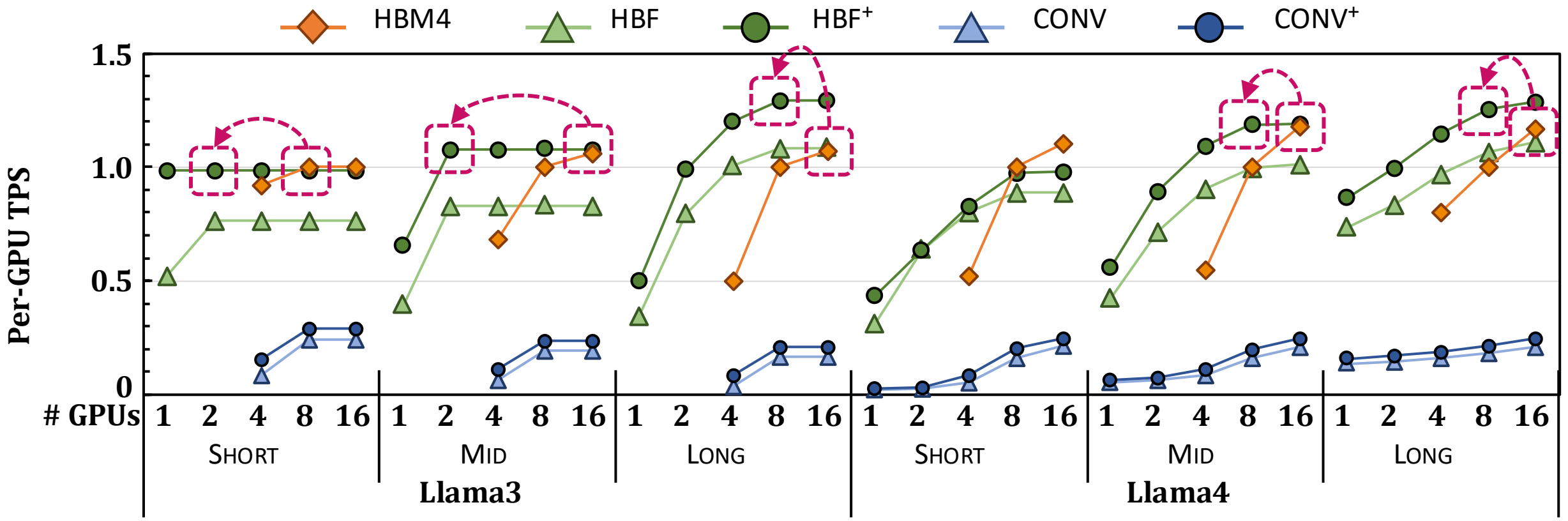


Per-GPU TPS under 0.1-second TPOT SLO

- The benefits of GPU scaling diminish under all evaluated systems
  - Regardless of the memory technology (i.e., also in HBM-based systems)
  - Larger batch → Amortizing more weight accesses (diminishing after a certain point)

- HBF enables achieving such batch sizes w/ fewer GPUs, which can potentially
  - Provide larger throughput gains for larger models (e.g., w/ trillion-parameter LLMs)
  - Allow developing even larger LLMs or leveraging multiple LLMs

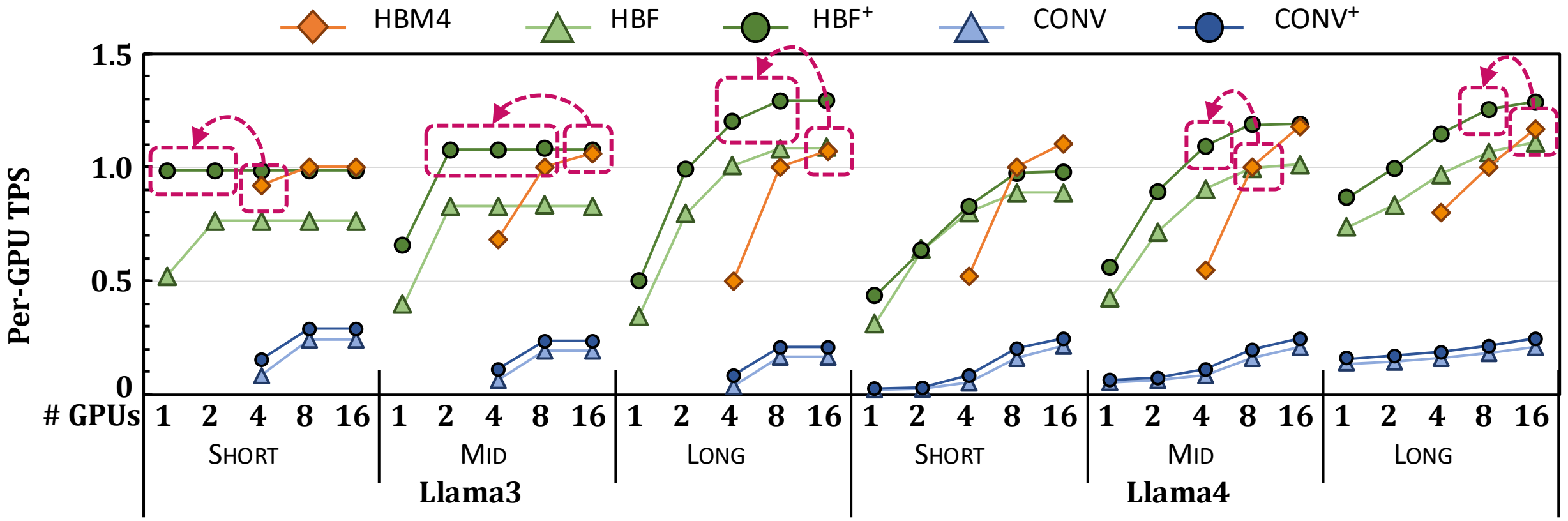
Evaluation Results: System Throughput



Per-GPU TPS under 0.1-second TPOT SLO

Diminishing improvements as GPU count increases?
→ Benefit: HBF enables achieving sufficient batch sizes w/ fewer GPUs

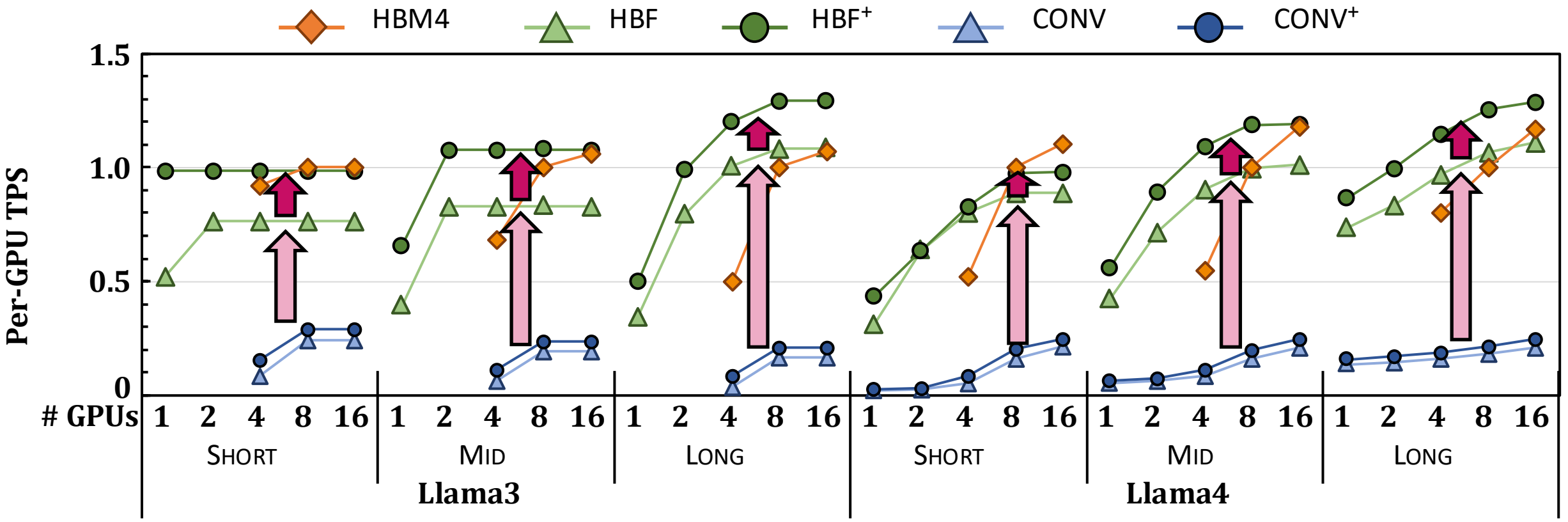
Evaluation Results: System Throughput



Per-GPU TPS under 0.1-second TPOT SLO

HBF can enable high per-GPU TPS even w/ fewer GPUs
→ Benefit: HBF can simultaneously improve performance and flexibility

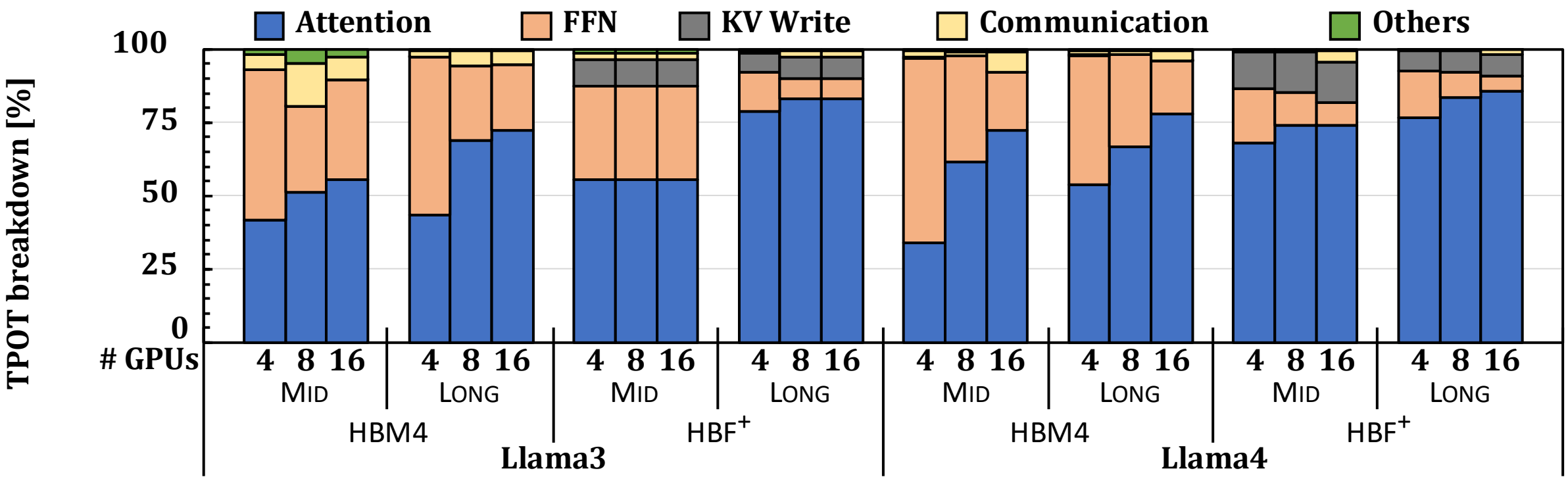
Evaluation Results: System Throughput



Per-GPU TPS under 0.1-second TPOT SLO

HBF+ significantly outperforms other HBF-based systems
→ Requirement: Achieving high read bandwidth is the key to benefits over HBM

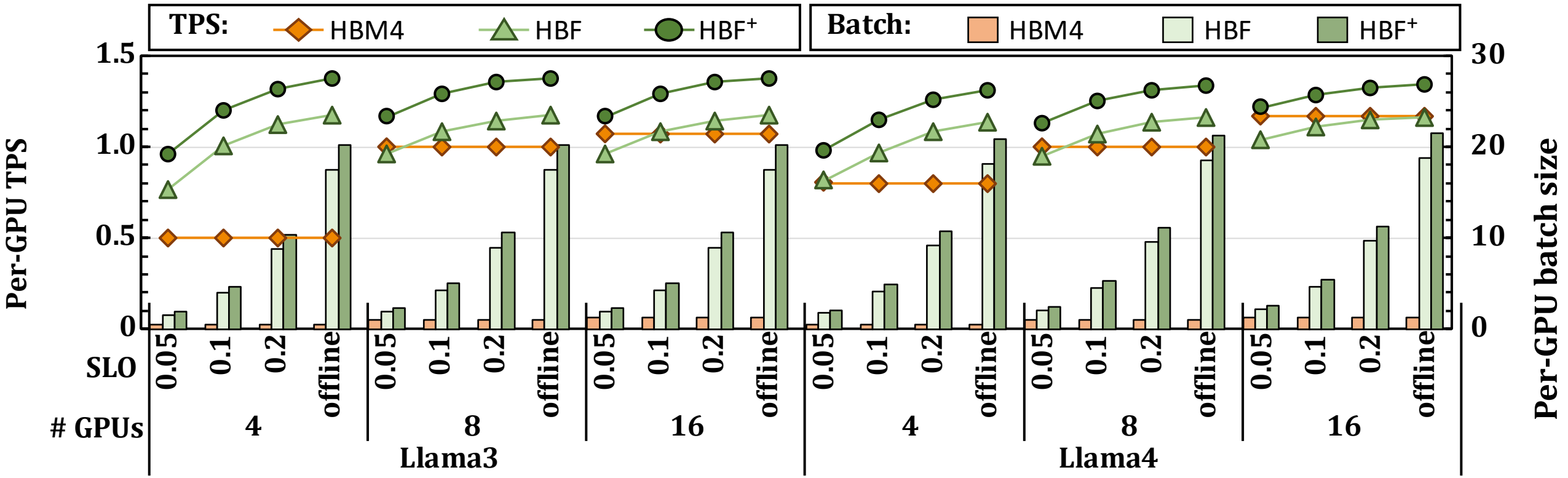
Evaluation: Performance Breakdown



Decoding time breakdown under 0.1-second TPOT SLO

KV-cache writes (from prefill nodes) introduce non-trivial performance overhead
→ Requirement: Reducing tPROG would further improve performance

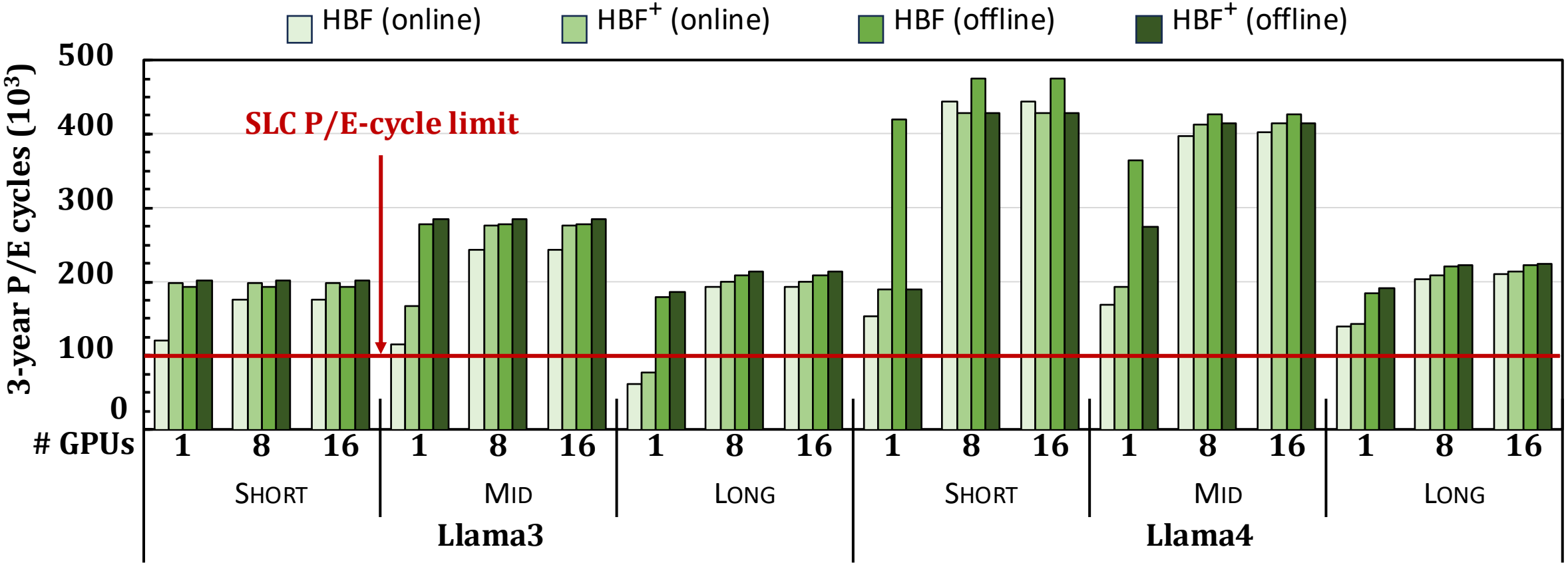
Evaluation: Effect of TPOT SLO Constraint



Performance and batch size impact of TPOT SLO

Benefit: HBF provides larger benefits under relaxed SLO constraints

Evaluation: Write-Traffic Analysis



Required PEC under different operating scenarios

Requirement: HBF requires significant endurance enhancement (2~5x)

Summary: Exploring HBF for LLM-Serving Systems

- Various potential benefits
  - Reducing the minimum GPU count
  - Improving system throughput at the same GPU count, especially
    - Before GPUs are sufficiently scaled for a given model
    - For long-context workloads
    - Under relaxed SLO requirements
  - Improving the performance and flexibility of LLM-serving systems
- Requirements
  - Maximizing read bandwidth
  - Significantly enhancing endurance
  - Improving program speed
  - w/ extensive architecture- and system-level optimizations

Exploring High-Bandwidth Flash for Modern LLM Inference: Opportunities and Challenges

SAFARI Live Seminar

Jisung Park

POSTECH CAOS
COMPUTER ARCHITECTURE & OPERATING SYSTEMS LABORATORY

2026.07.24