Is High-Bandwidth Flash All You Need?

Vinicius Petrucci, Felippe Zacarias, Vishal Tanna
Micron Technology
San Jose, CA, USA

Abstract

Large language model (LLM) inference accelerators are converging on multi-tier memory to hold multi-terabyte mixture-of-experts (MoE) weights and long-context key-value (KV) state. We use a roofline-based analytical model to compare three augmentations to high-bandwidth memory (HBM): on-package high-bandwidth flash (HBF), on-package LPDDR, and LPDDR over a chip-to-chip (C2C) link to a host CPU. We find that on-package LPDDR matches or beats every alternative at every operating point we evaluate, by up to 5.6× over HBM-only, via three properties. (1) *Capacity*: on-package augmentation admits compact layouts that HBM-only cannot fit; at peak throughput under the latency SLO, HBF and LPDDR both shrink the model-instance footprint 4× (16 to 4 GPUs). (2) *Writability*: LPDDR absorbs KV spill that HBF cannot serve, given NAND's write-rate and endurance limits, yielding up to 2.1× over HBF on KV-heavy 1 M-context workloads. (3) *On-package bandwidth*: where C2C carries critical-path traffic, LPDDR's higher effective decode bandwidth yields 1.3×–2.0× over C2C. The larger LPDDR-over-C2C gaps are replica-shrink (property (1)): on-package LPDDR can host the cold-expert pool whereas off-package C2C cannot, positioning C2C-attached LPDDR as a KV-overflow rather than expert-offload tier.

Disclaimer. *The views expressed in this paper are those of the authors and do not necessarily reflect the official position of Micron Technology, Inc. This work represents ongoing research and early findings based on an analytical model and publicly available information.*

1 Introduction

Mixture-of-experts (MoE) weights now reach multiple terabytes, and reasoning-era inference holds long-context key-value (KV) state that grows with batch and sequence length, outrunning the capacity of an on-package high-bandwidth memory (HBM) stack. Prior work argues that HBM alone is insufficient for future large language model (LLM) inference accelerators and calls for high-bandwidth flash (HBF) capacity expansion [5, 8, 11].

The 3rd Workshop on Hot Topics in System Infrastructure (HotInfra'26), co-located with ISCA'26, June 28, 2026, Raleigh, NC, USA

We refer to the LPDDR, HBF, and chip-to-chip (C2C) augmentations collectively as *tier-2* memory (capacity tiers beneath HBM). The recently proposed H^3 architecture [5] pairs HBM with on-package HBF on a daisy-chained substrate. We adopt the same packaging but swap HBF for on-package LPDDR. The operational payoff of such tier-2 augmentation is replica-shrink: packing an instance into fewer GPUs frees the scarce accelerators that frontier-GPU supply constrains [2]. The figure of merit is tok/s/kW under a per-user time-per-output-token (TPOT) service-level objective (SLO), capturing the two binding constraints in production serving that operators cannot let degrade at tail latency [2]: rack power and interactive latency.

Two limits make the choice non-obvious. HBF buys capacity, but NAND is *write-limited* on the on-package decode timescale: its program rate and write endurance are orders of magnitude below DRAM, so it cannot absorb KV that spills during decode as batch and context grow, nor reshuffle the expert pool that serving churn demands (LoRA swaps, model versioning, batch-driven expert-set drift [1]). LPDDR does both at nanosecond rewrite latency. The off-package alternative, LPDDR over a C2C link, is equally writable but lacks the on-package bandwidth to overlap cold-expert fetches with decode.

We adopt a MoE expert-offload [4, 7, 16] allocator for on-package tier-2 (LPDDR, HBF): the frequently-routed ("hot") experts are pinned in HBM and the rarely-routed ("cold") complement offloaded to tier-2, with KV kept in HBM and spilled to a writable tier (LPDDR) only under capacity pressure. This is feasible for on-package tier-2 because the H^3 daisy-chain [5] provides a dedicated die-to-die (D2D) interconnect between HBM and the on-package tier-2 that runs in parallel with HBM compute, so cold-expert fetches overlap decode rather than stalling it. Off-package C2C lacks this parallel path, so it instead uses the *weights-first* allocator that ships as the default in major open-source inference stacks: weights stay in HBM and C2C absorbs only KV spill (Section 4 quantifies why expert-offload is worse for C2C). We *spill* cold KV to a writable tier rather than *evict* it: spilling reads tokens back at full precision, so every attention term is bit-identical to the HBM-resident computation and reasoning accuracy is intact [17], making placement purely a performance and power question.

Table 1. Per-GPU tier specs on a 72-GPU rack. BW is nominal peak read-rate; effective BW under scattered access in Section 2. LPDDR design grid swept in Section 4.

Tier	Cap. (GiB)	BW (TB/s, rd)	Latency (ns)	Energy (pJ/bit)
HBM	512	32.0	100	3.0
On-pkg LPDDR	2,048	2.0	100	2.4
HBF	4,096	8.0	20,000	8.5
Off-pkg C2C	3,072	1.8	500	5.0

Results and contributions. We contribute (i) the first comparison of HBF, on-package LPDDR, and C2C-attached LPDDR on a single figure of merit, tok/s/kW under a TPOT SLO, across four production-scale inference workloads at 1 M-token context; and (ii) a decomposition of the on-package LPDDR advantage into three separable properties, *capacity* (replica-shrink), *writability* (KV-spill absorption that write-limited HBF lacks at decode rate), and *on-package bandwidth* (which C2C lacks), each mapped to the workload regime it governs (Section 3).

2 Analytical Model

Platform. The analytically modeled platform (no silicon measurement) is a 72-GPU rack with 1.8 kW per-GPU TDP and 1.8 TB/s per-GPU unidirectional fabric bandwidth. Per-GPU tier specs include eight HBM stacks (512 GiB), bounding the largest MoE models at moderate tensor-parallel (TP) degree and motivating tier-2 augmentation.

Workloads. We evaluate four MoE models spanning three attention families. The DeepSeek-V4 pair (DSV4-Pro, DSV4-Flash) combines compressed-sparse and heavily-compressed attention (CSA/HCA) with shared-KV multi-query attention (MQA) for a light ∼3–5 KB/tok KV footprint (BF16/FP8). Kimi-K2.6 uses KV-heavy multi-head latent attention (MLA) at ∼39 KB/tok. Nemotron-3-Ultra (549 B total, 55 B active) is a Mamba-2 + grouped-query attention (GQA) hybrid with LatentMoE on ∼44% of layers; its ∼12 KB/tok BF16 KV lives only on the 12 GQA layers of 108. All are sparse MoE (6–22 of 256–512 experts routed, plus 1 shared), with MoE in every block except Nemotron-3-Ultra's interleaved layout. At 512 GiB FP8, Nemotron-3-Ultra's weights nearly fill a single GPU's HBM, so it cannot co-locate weights and long-context KV on one GPU without tier-2.

Memory tiers. Per-bit read energy is media at-pin [6, 10] plus 0.5 pJ/bit/hop in the H^3 daisy-chain [5]; HBF latency matches H^3's SLC-NAND assumption. The on-package LPDDR tier reuses the H^3 daisy-chained substrate (same D2D link and hop as HBF). The default LPDDR design point is 2 TiB at 2 TB/s (Table 1).

Metric. Throughput is the per-step decode rate (batch/TPOT) at fixed input context, evaluated at a single decode step (fixed KV state) rather than accumulated over the full output sequence, so output length does not affect the metric. Rack throughput is the request-rate proxy (requests/s = rack-tok/s ÷ mean output length). We omit the time-to-first-token (TTFT) prefill term: prefill is commonly disaggregated onto separate hardware, making the decode tier a self-contained choice. We do not claim it for prefill, whose sustained-read profile may favor a NAND tier (Section 4).

Solver and SLOs. At its core a roofline computes per-op time as the maximum of its compute-bound and memory-bound limits from published hardware peaks (Williams et al. [14]); the step model extends this by running the HBM, on-package D2D, and C2C paths in parallel (step time = slowest path sequential collectives), with D2D contention clipped to its base-die share. The solver enumerates parallelism strategies per operating point and selects the one with maximum tok/s/kW meeting the TPOT SLO. We evaluate at 10 ms and 25 ms TPOT (tighter than typical production [13]) to reflect agentic workloads where per-step TPOT compounds across multi-turn tool-use chains. We write an operating point as *batch @ SLO* (e.g., b=128 @ 25 ms is batch 128 under the 25 ms TPOT SLO).

KV partition and placement. The KV hot-access pattern follows H2O's heavy-hitter calibration [18]: a small fraction of tokens absorbs the majority of attention reads, defining which KV bytes stay in HBM and which spill to tier-2. Because HBF is write-limited (programmed once with the static cold-expert pool at load, read many times during inference), the solver hard-rejects any placement that routes KV to it, confining KV spill to the writable DRAM tiers (on-package LPDDR or off-package C2C); no NAND program/erase traffic enters the decode model. The 2 TiB LPDDR design point is the smallest capacity admitting a 1-GPU layout on the only LPDDR-binding configuration (b=128 @ 25 ms on DSV4-Pro; Section 3); BW above ∼1 TB/s is unused on the other three workloads (Section 4).

Effective tier-2 BW. Effective BW = nominal peak $\times \eta$, with η derived from each medium's access model under two patterns: *scattered* cold-expert reads (random-position KB–MB fetches) and *paged* KV reads. LPDDR delivers $\eta \approx 0.80$ scattered (1.6 TB/s, from DRAM row-buffer hit rates) and $\eta \approx 0.85$ paged (1.7 TB/s). HBF is heavily clipped on scattered access by NAND's ∼20 µs read latency on small fetches ($\eta \approx 0.36$, 2.9 TB/s). C2C

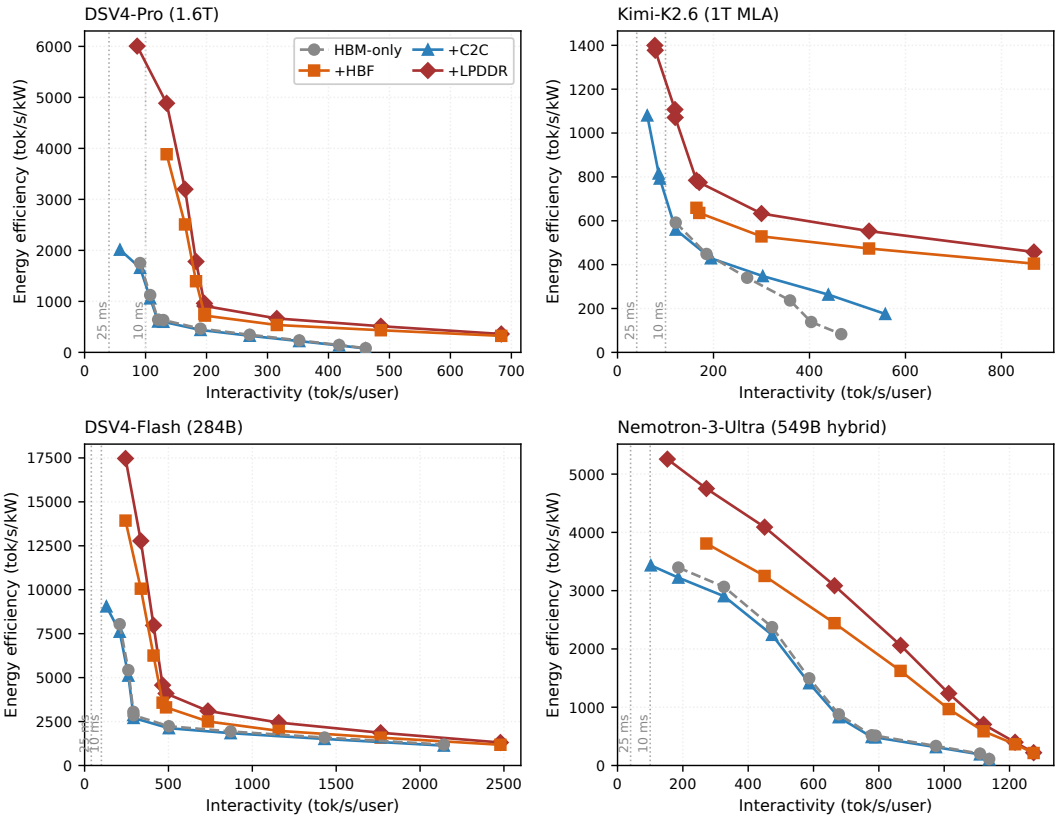

Figure 1. Pareto envelopes per workload at 1 M-token context on a 72-GPU rack, LPDDR at the 2 TiB / 2 TB/s design point. Each curve is the upper-right (interactivity, energy efficiency) frontier across every feasible layout (TP/EP/DP degrees; attention-mode; batch $\in \{4, 8, 16, 32, 64, 128, 256\}$) under the per-tier allocators of Section 1 (expert-offload on-package, weights-first for C2C). Dotted guides mark the 25 ms and 10 ms TPOT SLOs (markers to their right meet that SLO). Each curve's rightmost extent is its interactivity ceiling; its y-plateau, its energy-efficiency ceiling.

paged-KV streaming pays a protocol penalty calibrated to Strata [15] ($\eta \approx 0.55$, 1.0 TB/s). KV-bound comparisons (e.g., Kimi-K2.6) use the paged figures (LPDDR 1.7 vs. C2C 1.0 TB/s); expert-bound comparisons use the scattered figures.

The scattered $\eta \approx 0.36$ is a worst case for HBF, but the comparison is insensitive to it: a modern HBF stack overlaps reads, and because an expert is large (∼66 MB) bandwidth, not the ∼20 µs latency, sets the cost [3, 6, 9, 12]. Raising HBF scattered η from 0.36 to 0.50 leaves every LPDDR-over-HBF ratio byte-identical (the winning expert reads overlap HBM compute, so HBF η never reaches the critical path), and even pairing an optimistic HBF (0.50) with a pessimistic LPDDR (0.65) flips no cell.

3 Findings

Two ceilings organize Figure 1. The *interactivity ceiling* (rightmost extent of each curve) is set by the smallest TP the placement admits: HBM-only is forced to TP≥4 on both Kimi-K2.6 and DSV4-Pro, for different reasons. DSV4-Pro is weight-bound (1.43 TiB FP8 weights need 4 GPUs = 2 TiB HBM; TP=2 at 1 TiB does not fit); Kimi-K2.6 is KV-bound (0.93 TiB weights fit at TP=2, but the KV-heavy MLA state at 1 M context does not). Either way, the per-layer AllReduce floors TPOT regardless of batch. The *energy-efficiency ceiling* (highest y per curve) is set by how much batch packs into a replica before some tier overflows. Tier-2 augmentations relieve both ceilings via *replica-shrink*: holding the cold-expert pool on tier-2 admits compact layouts (peak 5.6× on DSV4-Pro b=256 @ 10 ms, where HBF and LPDDR each shrink the replica from 16 to 4 GPUs; the gain drops to 3.8× at 25 ms as HBM-only packs denser).

Writable tier (vs. HBF). The LPDDR-over-HBF gap is widest on Kimi-K2.6 (orange-vs-red separation in the second panel of Figure 1), where KV-heavy MLA grows linearly with batch×context. Write-limited NAND cannot absorb the KV tail, so HBF keeps KV in HBM and

TP stays pinned at 8 across batches, capping its energy efficiency at ∼660 tok/s/kW. LPDDR's writable on-package tier absorbs ∼5.5 TiB of KV spill at nanosecond rewrite latency, freeing HBM to grow per-replica batch from 64 to 256 and lifting efficiency to ∼1,400 tok/s/kW (2.1× over HBF, the abstract figure, at b=128 @ 25 ms). The gap narrows or closes on the other three models, where HBF's compact layout keeps KV in HBM and the write limit is never exercised. On Nemotron-3-Ultra the tier-2 win is capacity-driven, not writability-driven: both tiers hold the cold-expert pool that lets the 512 GiB-weight model fit on one GPU (8 → 1 replica), so they tie until LPDDR's higher BW pulls ahead at ≥ 2 TB/s.

Effective BW (vs. C2C). The 1.3×–5.6× LPDDR-over-C2C gap has two regimes. At the peak (the same 5.6× DSV4-Pro point) the gap is set by allocator semantics: weights-first cannot shrink the C2C layout while expert-offload puts the cold pool on LPDDR and admits a 4-GPU replica. The narrow end (1.3×–2.0×) is the BW-limited regime, set by either the hybrid layer mix bounding MoE traffic (Nemotron-3-Ultra) or C2C's 1.7×-lower effective paged-KV bandwidth on KV-bound Kimi-K2.6 (Section 2).

Per-step LPDDR traffic. Only DSV4-Pro b=128 @ 25 ms is LPDDR-binding (1-GPU layout, ∼7 ms SLO slack); elsewhere LPDDR traffic runs in parallel with the binding HBM path, and at the 5.6× point cold-expert reads fully overlap HBM-MoE compute via the H^3 daisy-chain [5].

4 Discussion

Regimes favoring HBF. HBF's differentiator is sustained on-package read BW at multi-TiB capacity, which RDMA-attached NAND (∼100 GB/s) cannot match; capacity-only workloads with loose BW budgets, by contrast, are well served by that RDMA path. Two regimes that we do not evaluate rely on HBF's bandwidth advantage: prefix-cache reuse under tight TTFT constraints [5], and decode-phase expert streaming when weights exceed LPDDR's on-package capacity.

HBF under weights-first. Under the production-default weights-first allocator, HBM+HBF reduces to HBM-only within 0.4% on every operating point: KV is the spillover target, write-limited HBF cannot accept it, and the TP sweep picks a layout where per-GPU experts already fit in HBM. Only expert-offload activates HBF.

LPDDR design points. Sweeping LPDDR BW 0.5–4 TB/s at the 2 TiB design point, LPDDR matches or beats HBF at 5.8× lower tier peak power (95 W at 2 TB/s vs HBF's 549 W per GPU). Throughput is BW-flat on the two models whose tier-2 traffic stays below the BW knee (Kimi-K2.6, DSV4-Flash); only the weight-heavy two scale with LPDDR BW (DSV4-Pro b=128 @ 25 ms 459 → 542 K tok/s; Nemotron-3-Ultra 565 → 618 as BW rises 1 → 4 TB/s), the only ones placing enough weight on LPDDR to reach the critical path. The BW floor is set by hotness, not capacity: the hot-expert pin (bias-8 [16], ∼73% of reads in HBM) and the H2O 20/30/50 split [18] fire tier-2 traffic only once HBM is exhausted, across a 0.10–0.40 hot-ratio sweep. Predictor drift adds a further consideration: re-pinning a drifting predictor is a write HBF's program rate cannot serve at decode rate, so HBF must keep the whole pool resident. Even with a flat predictor, the LPDDR-over-HBF gap holds at ≥1.24× across the expert-bound points. Capacity is the binding knob: 2 TiB is the smallest pool admitting the compact layouts (DSV4-Pro's 1-GPU layout at b=128 @ 25 ms, 4.2× over HBM-only; Nemotron-3-Ultra's 8→1 replica), while 512 GiB is too tight on DSV4-Pro and Kimi-K2.6.

Why not C2C for experts? Each tier's allocator is its own throughput-maximizing choice, so the LPDDR-over-C2C comparison is unconfounded: HBM+C2C under expert-offload (EF) underperforms its weights-first (WF) mode by 6.9× geomean (4.7–9.7×), whereas the on-package tiers are the mirror image (EF beats WF by 1.8–1.9× on LPDDR and HBF). The asymmetry is physical: dropping C2C latency to 100 ns has no effect, while raising effective BW to 4 TB/s recovers ∼2.2×, and the off-package paged read pays an NVLink-C2C protocol and CPU-DMA penalty (η=0.55) the on-package H^3 path avoids. C2C-attached LPDDR thus serves best as a higher-capacity KV-overflow tier.

Smaller fault domain. Replica-shrink also shrinks the unit that a single GPU fault disables, a first-order reliability concern for operators [2]: because every decode step synchronizes across the model instance, one failed GPU stalls all of it, so shrinking from 16 GPUs to 4 (a single compute tray) cuts that blast radius to one quarter. We report only this direction, leaving recovery-cost quantification to future work.

5 Conclusion

High-bandwidth flash augmenting HBM is not the whole answer: it adds capacity but its write-limited NAND cannot absorb the KV that the decode phase spills, so on-package LPDDR is the more general tier-2 choice. Across all workloads and operating points evaluated, HBM+LPDDR matches or outperforms every alternative (up to 5.6× over HBM-only): capacity drives most of the gain (HBF and LPDDR tie when KV fits in HBM), writability separates LPDDR from HBF on KV-heavy workloads, and on-package BW separates it from C2C on critical-path traffic. On-package LPDDR (expert-offload) and C2C-attached LPDDR (KV-overflow) cover both tier-2 roles.

References

[1] Abhimanyu Bambhaniya, Geonhwa Jeong, Jason Park, Jiecao Yu, Jaewon Lee, Pengchao Wang, Changkyu Kim, Chunqiang Tang, and Tushar Krishna. 2026. Scaling Multi-Node Mixture-of-Experts Inference Using Expert Activation Patterns. *arXiv preprint arXiv:2604.23150* (2026).

[2] Ying Chen, Wendy Hu, Ankit Mathur, Mike Eastham, Pei-Lun Liao, Wai Wu, and Arjun DCunha. 2026. Reliable LLM Inference at Scale: Lessons from Building Reliable LLM Inference Infrastructure. Databricks Engineering Blog. Non-peer-reviewed industrial engineering blog post; cited as production-operator evidence for serving-layer motivation only.

[3] W. Cheong et al. 2018. A Flash Memory Controller for 15 μs Ultra-Low-Latency SSD Using High-Speed 3D NAND Flash with 3 μs Read Time. In *IEEE International Solid-State Circuits Conference (ISSCC)*. Samsung Z-NAND SLC-fast: 3 μs tR at reduced page size.

[4] Zehao Fan, Zhenyu Liu, Yunzhen Liu, Yayue Hou, Hadjer Benmeziane, Kaoutar El Maghraoui, and Liu Liu. 2025. Context-Aware Mixture-of-Experts Inference on CXL-Enabled GPU-NDP Systems. arXiv:2512.04476 [cs.LG] https://arxiv.org/abs/2512.04476

[5] Minho Ha, Euiseok Kim, and Hoshik Kim. 2026. H3: Hybrid Architecture Using High Bandwidth Memory and High Bandwidth Flash for Cost-Efficient LLM Inference. *IEEE Computer Architecture Letters* 25, 1 (2026), 49–52. https://doi.org/10.1109/LCA.2026.3660969

[6] Po-Kai Hsu, Weihong Xu, Qunyou Liu, Tajana Rosing, and Shimeng Yu. 2026. HAVEN: High-Bandwidth Flash Augmented Vector ENgine for Large-Scale Approximate Nearest-Neighbor Search Acceleration. *arXiv preprint arXiv:2603.01175* (2026). Section 4.1 calibrates read energy using industry-grade 3D NAND data, scaling from the $\sim$30 pJ/bit read energy reported for contemporary devices; Figure 8(b) shows re-architected distributed-subarray HBF at $\sim$2–16 pJ/bit across page-size/block-count trade-offs.

[7] Taehyun Kim, Kwanseok Choi, Youngmock Cho, Jaehoon Cho, Hyuk-Jae Lee, and Jaewoong Sim. 2024. MoNDE: Mixture of Near-Data Experts for Large-Scale Sparse Models. arXiv:2405.18832 [cs.LG] https://arxiv.org/abs/2405.18832

[8] Xiaoming Ma and David Patterson. 2026. Challenges and Research Directions for Large Language Model Inference Hardware. *arXiv preprint arXiv:2601.05047* (2026).

[9] H. Maejima et al. 2018. A 512Gb 3b/Cell 3D Flash Memory on a 96-Word-Line-Layer Technology. In *IEEE International Solid-State Circuits Conference (ISSCC)*. Commodity 3D TLC tR $\approx$ 50 μs at 96 WL layers.

[10] Samsung Electronics. 2025. LPW DRAM Keynote. In *IEEE International Solid-State Circuits Conference (ISSCC)*. San Francisco, CA. Keynote by Song Jae-hyuk, Samsung DS Division CTO; reports 1.9 pJ/bit end-to-end at 200+ GB/s per channel for LPW (Samsung's Low-Power Wide I/O DRAM).

[11] Anton Shilov. 2025. SanDisk's New High Bandwidth Flash Memory. Tom's Hardware.

[12] C. Siau et al. 2019. A 512Gb 3b/Cell 3D Flash Memory at 1.066Gbps Peak I/O Bandwidth. In *IEEE International Solid-State Circuits Conference (ISSCC)*. Commodity 3D TLC tR $\approx$ 20–50 μs at 128 WL layers.

[13] Qipeng Wang and Zhendong Yang. 2026. Taming Request Imbalance: SLO-Aware Scheduling for Disaggregated LLM Inference. arXiv:2605.02329 [cs.DC] https://arxiv.org/abs/2605.02329

[14] Samuel Williams, Andrew Waterman, and David Patterson. 2009. Roofline: An Insightful Visual Performance Model for Multicore Architectures. *Commun. ACM* 52, 4 (2009), 65–76. https://doi.org/10.1145/1498765.1498785

[15] Zhongwei Xie et al. 2025. Strata: Hierarchical Context Caching. *arXiv preprint arXiv:2508.18572* (2025).

[16] Leyang Xue, Yao Fu, Zhan Lu, Luo Mai, and Mahesh K. Marina. 2024. MoE-Infinity: Offloading-Efficient MoE Model Serving. *arXiv preprint arXiv:2401.14361* (2024).

[17] Aojie Yuan, Tianqi Shen, and Dajun Zhang. 2026. Not All Thoughts Need HBM: Semantics-Aware Memory Hierarchy for LLM Reasoning. *arXiv preprint arXiv:2605.09490* (2026). Proves full-precision KV offload to DDR is zero-approximation-error (accuracy depends on the eviction ratio, not the HBM ratio); the accuracy-side complement to our performance/power analysis.

[18] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. 2023. H2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.