LPDDR for General-Purpose and AI Workloads in Large-Scale Data Center Deployments

Authors [Micron]: Khayam Anjam, Naveen Krishna Yarlagadda, Sujit Somandepalli, Sudharshan Vazhkudai, Praveen Vaidyanathan

Authors [Meta]: Kiran Malwankar, Prakash Chauhan, Anil Agrawal, Craig Ross

Introduction

As data centers face increasingly daunting power and thermal constraints, the shift to LPDDR5X is a critical evolution in server architecture. Historically used in mobile and edge devices, LPDDR5X is uniquely positioned for modern data center workloads because it delivers improved memory bandwidth while operating at a fraction of the power of conventional DDR5 implementations. As AI and general-purpose computing (GPC) workloads continue to scale, this energy efficiency is paramount for managing total cost of ownership (TCO) at the data center level. New modular form factors like SOCAMM2 further enable this shift—delivering the efficiency of mobile memory alongside the serviceability that data centers demand and establishing LPDDR5X as a practical and compelling option for hyperscale deployments.

This work is the result of a collaborative effort between Meta and Micron to evaluate the suitability of LPDDR5X for data center environments, characterizing memory bandwidth, capacity, latency, and power efficiency under representative workloads. We adopted DCPerf, an open-source benchmark suite from Meta that reflects real-world data center workloads.

Meta developed DCPerf to represent the dominant workload categories in its hyperscale fleet and to guide platform selection using both application-level metrics and system-level signals. It is used internally by Meta for product evaluation, platform configuration, and procurement decisions in data centers. The suite is designed to model major workload categories based on production fleet profiling, and its benchmarks are used to guide hardware/software co-design and early-stage optimizations.

Higher LPDDR5X capacity and bandwidth enable measurable gains across both AI and GPC workloads in the DCPerf suite. This approach enables better resource utilization and reduces energy consumption, making it an attractive option for enterprises and data centers seeking to balance performance with sustainable growth. We present a comprehensive characterization of LPDDR5X across these dimensions.

Key Takeaways

<7%
Of system power consumption

LPDDR delivers improved power efficiency: LPDDR5X consumes approximately one-third the power of DDR5, accounting for **<7%** of total system power.[1]

3x
Throughput improvement

Memory capacity Is a performance multiplier: Doubling LPDDR5X capacity eliminates disk spill in Spark SQL workloads, delivering **2x to 3x** throughput improvement.

13%
More usable bandwidth

Bandwidth drives performance: Higher LPDDR5X speeds translate to workload gains—a system with faster LPDDR memory (8533 MT/s) delivered **13% more** DRAM bandwidth than the 6400 MT/s configuration, resulting in 11% higher Tensor Rebatching throughput.

~9%
Lower latency

Latency advantage for GPC workloads: Faster LPDDR5X reduces memory latency (**~9% lower** P99 tail-latency)—a measurable gain for services operating under strict SLA requirements.

[1] See our analysis on page 3 for power efficiency comparisons

Performance Evaluation and Scaling Analysis

Our analysis is structured across four key dimensions: **bandwidth**, **latency**, **capacity**, and **power efficiency**. These axes capture the primary characteristics that determine memory subsystem suitability for modern data center workloads. Workloads that are bandwidth-sensitive benefit directly from LPDDR5X's higher signaling rates, while latency-sensitive workloads stress memory access response times under realistic concurrency. Capacity sensitivity captures scenarios where working set size is the binding constraint. Power efficiency underpins all these dimensions, and LPDDR5X's low operating voltage enables it to optimize TCO across an entire fleet by delivering competitive performance at a significantly reduced energy footprint. Collectively, these four dimensions provide a comprehensive framework for assessing where and how LPDDR5X delivers value in production data center environments.

Workloads

From DCPerf, we selected a suite of AI and GPC benchmarks for our study. These included Tensor Rebatching, Deserialization, MediaWiki, and SparkBench to evaluate LPDDR5X's bandwidth, power, capacity, and latency characteristics under realistic data center conditions.

Memory bandwidth-sensitive workloads: Tensor Rebatching performs sustained multi-threaded memory copy (memcpy) operations from a large memory pool to contiguous output buffers, creating a memory bandwidth-sensitive workload. Deserialization complements this by exercising memory access patterns typical of data ingestion pipelines, where serialized tensor data must be parsed and reconstructed in memory before inference. Both operations represent critical stages in AI data pipelines where data movement dominates performance and energy consumption.

Memory latency-sensitive workloads: To characterize how memory performance affects tail latencies on GPC workloads, we ran the mediawiki_mem workload across a sweep of connections-per-thread values (-c), measuring throughput (requests per second, or RPS) and request latency at increasing concurrency. The MediaWiki benchmark models Facebook web-serving workloads in Meta data centers.

Memory capacity-sensitive workloads: To demonstrate LPDDR5X's capacity advantages, we evaluated SparkBench under default and higher-load configurations. Spark SQL workloads are memory-capacity sensitive as they rely on keeping large datasets, intermediate shuffle data, and OS page cache resident in memory to avoid expensive spill-to-disk behavior. When memory capacity is insufficient, shuffle, sort, and scan stages incur repeated disk I/O, significantly inflating end-to-end execution time. SparkBench models data analytics workloads by running a data-warehouse-style Spark SQL workload.

While LPDDR was traditionally constrained by its soldered-down design, the SOCAMM form factor removes that barrier through a modular, high-density interface. This allows data centers to expand memory footprints significantly, enabling these workloads to handle massive datasets resident in memory without sacrificing performance.[2]

Hardware Platforms

We evaluated two LPDDR memory configurations. SKU1 uses a 4-rank setup, providing 512GB per socket at 6400 MT/s. SKU2 uses a 2-rank configuration, providing 256GB at 8533 MT/s. The higher rank count of SKU1 increases capacity by stacking additional DRAM die, but the added electrical load restricts the memory signaling rate—which explains why SKU2 achieves a higher data rate despite lower capacity.

[2] Micron's High-Capacity 256GB LPDRAM SOCAMM2 for Data Center Infrastructure

	SKU1	SKU2
CPU	ARM 72c / socket	ARM 72c / socket
Memory Capacity	512GB/socket LPDDR5X (1TB total)	256GB/socket LPDDR5X (512GB total)
Memory Speed	6400 MT/s; 384 GB/s	8533 MT/s; 512 GB/s
Memory Rank	4	2
Micron Part #	62F8	62F4

Table 1: Hardware Platforms

Analysis

Bandwidth Sensitivity: LPDDR's Higher Bandwidth Drives Better Performance

Figure 1 shows that SKU2's faster LPDDR5X memory (8533 MT/s) delivers higher sustained DRAM bandwidth than SKU1 (6400 MT/s) on the Tensor Rebatching workload. This observed 13% bandwidth advantage translates directly to improved workload performance, achieving 11% higher rebatch throughput and 10% lower time-per-batch.

In AI data pipelines, Tensor Rebatching sits at a critical junction between data ingestion and compute. Any reduction in time-per-batch directly translates to less idle time at the accelerator, improving overall pipeline throughput. SKU2's bandwidth advantage demonstrates that higher LPDDR5X signaling rates can meaningfully accelerate this stage, reducing the memory bottleneck that would otherwise limit how quickly data reaches the compute layer.

Deserialization also exhibits bandwidth sensitivity. This workload parses serialized tensor data and reconstructs it in memory, combining memory reads with lightweight compute operations. SKU2 achieved 6% higher memory bandwidth and 18% higher IPC (instructions per cycle), confirming that the faster memory interface reduces backend stalls and enables the CPU to spend more cycles doing useful work rather than waiting for memory.

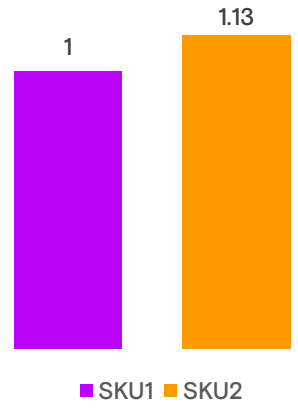


Figure 1: Normalized DRAM Bandwidth; Tensor Rebatching

Power Efficiency: LPDDR's Power Advantage

The power breakdown analysis (Figure 2) reveals LPDDR5X's lower power consumption under memory-intensive workloads. During the Tensor Rebatching benchmark, DRAM power consumption registers as just 6.8% of total system power. This contrasts sharply with traditional DDR5 server configurations, where memory power can consume a substantially larger share of total system power. Previous Micron analysis has highlighted this advantage, demonstrating up to 75% lower LPDDR5X power usage versus DDR5 (Figure 3).[3]

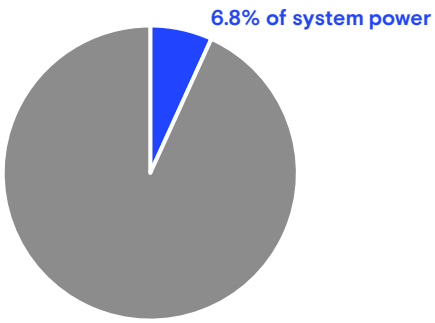


Figure 2: LPDDR5X Share of System Power (6.8%); Tensor Rebatching 18

[3] The role of low-power (LP) memory in data center workloads

LPDDR5X achieves superior power efficiency while delivering higher memory bandwidth. Its reduced operating voltage and advanced signaling allow bandwidth to scale without a corresponding rise in power consumption, resulting in better bandwidth-per-watt performance than conventional server memory.

At data center scale, this efficiency advantage can materially reduce power-related operating costs—delivering the memory bandwidth modern AI workloads demand without the proportional power and cooling overhead of conventional DDR5 solutions.

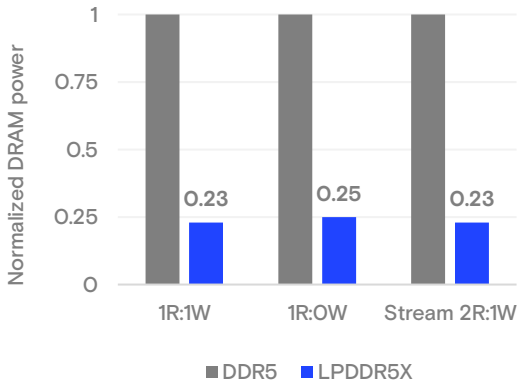


Figure 3: Multichase Memory Power Comparison

GPC Latency: mediawiki_mem

The oss_performance_mediawiki_mem benchmark in DCPerf is a Tuned + MLP (memory-level parallelism) variant of mediawiki_mem, augmented with LambdaChase to inject additional memory pressure. We swept connections-per-thread (c=36–288) to characterize throughput requests per second (RPS) and tail-latency scaling under increasing concurrency.

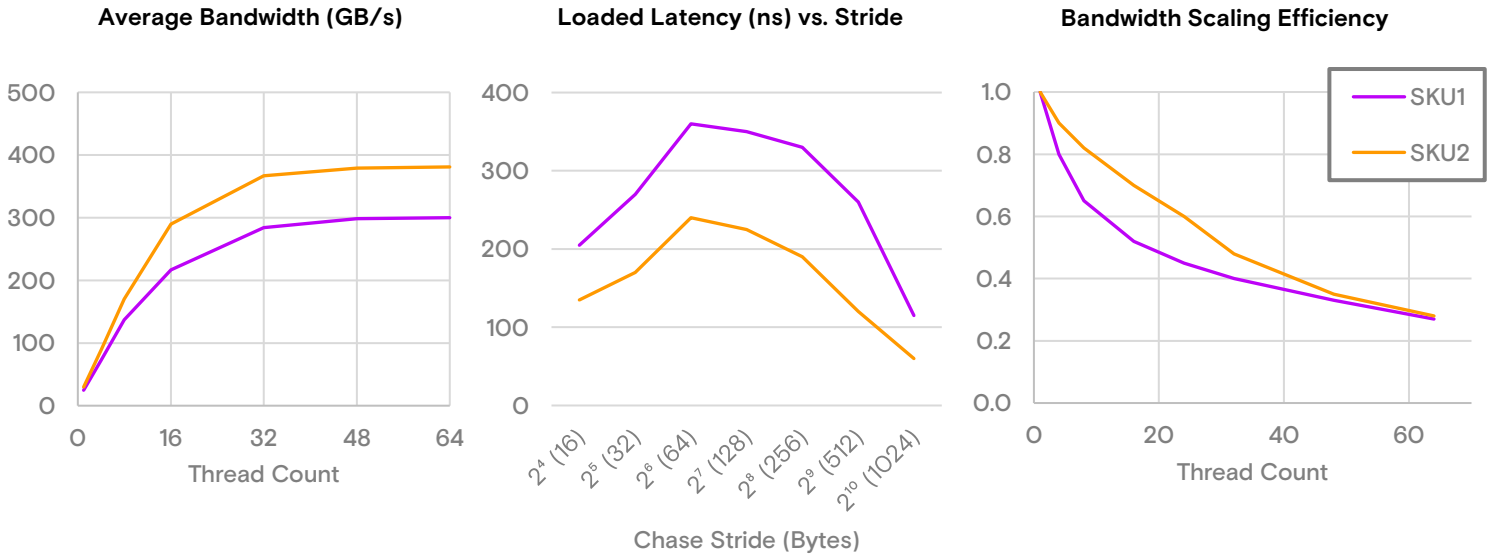
Across the mediawiki_mem concurrency sweep, SKU2 achieves a quality-of-service advantage through higher throughput while reducing P99 memory read latency by 12–20 ms (~9%) (Figure 4). This tail-latency compression enables the load generator to sustain higher effective concurrency, directly translating into higher RPS.

Figure 4: MediaWiki P99 Tail Latencies

The primary differentiator is SKU2's faster LPDDR5X data rate (8533 MT/s versus 6400 MT/s), which provides higher bandwidth and lower latency. This reduces backend stalls and improves core efficiency (IPC)—tightening the tail of request latency rather than merely shifting the mean.

Microbenchmark cross-validation using Multichase shows the same root cause: SKU2 sustains higher bandwidth and lower loaded latency across stride and thread sweeps (Figures 5(a) – 5(c)), confirming the application-level QoS improvement is fundamentally memory-subsystem-driven.

Finally, none of the throughput/latency curves saturate within the evaluated range; with HHVM auto-scaling capped at two instances, the achieved concurrency maps to the lower thread regime of Multichase, indicating additional scaling headroom—where SKU2's advantage is expected to widen as SKU1's loaded latency inflects at ≥16 threads.



Figures 5(a)-5(c): Bandwidth, Loaded Latency, and Scaling Efficiency – Multichase Benchmark

Capacity Sensitivity: LPDDR's Larger Capacity Advantage

To evaluate capacity sensitivity, we ran the SparkBench from the DCPerf suite under two configurations—standard query load (spark_standalone_remote) and a 3x heavier variant (spark_standalone_remote_3x)—on both platforms. The systems are otherwise identical in CPU architecture, core count, cache hierarchy, and Spark worker memory allocation (276GB), isolating **memory capacity as the primary performance variable**.

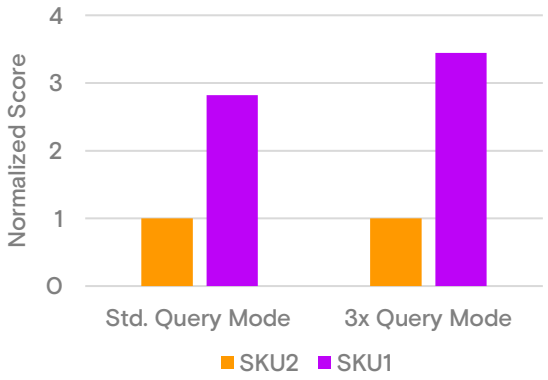


Figure 6: Spark Standard and 3x Load Performance Gain

Figure 7 clearly illustrates the root cause of SKU2's performance gap. Under the standard workload, the shuffle-heavy stage becomes the dominant bottleneck, showing a 38x slowdown once shuffle data exceeds available DRAM and spills to SSD. Under the 3x workload, the bottleneck shifts to a scan stage dominated by re-reads where SKU2 experiences a 5.96x slowdown because its smaller OS page cache cannot retain the dataset, forcing repeated SSD re-reads on each scan.

By contrast, SKU1's larger memory capacity serves a dual function: It retains the entire shuffle working set in DRAM, eliminating disk spills, while also sustaining an OS page cache large enough to keep frequently accessed data resident across stages. Together, these effects demonstrate that memory capacity is the decisive performance factor for Spark, and that provisioning sufficient DRAM to hold both shuffle partitions and page cache is essential to avoiding order-of-magnitude slowdowns caused by disk I/O.

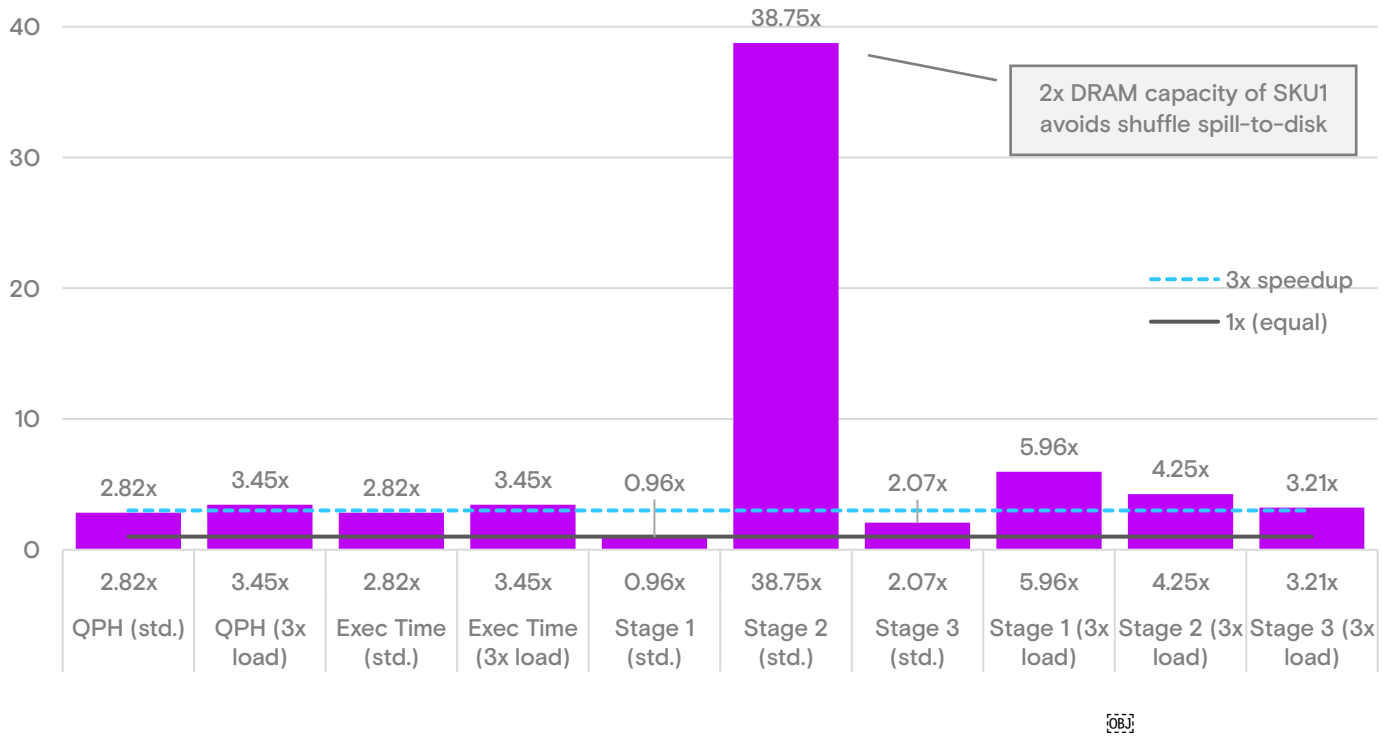


Figure 7: SparkBench Stage-Level Breakdown, SKU1 Speedup vs. SKU2 (higher is better)

Conclusion

Our evaluation using Meta's DCPerf suite confirms that LPDDR5X delivers across all four dimensions of the analytical framework presented here:

- On **bandwidth sensitivity**, SKU2's higher memory speed translated to improved throughput and reduced time-per-batch in Tensor Rebatching, with Deserialization further confirming that faster memory reduces backend stalls and improves CPU utilization.
- On **latency sensitivity**, the faster LPDDR5X interface delivered better tail latency in the MediaWiki workload at higher concurrency, demonstrating that memory speed improvements benefit GPC workloads as well.
- On **capacity sensitivity**, expanding the LPDDR5X memory footprint eliminated disk spilling in SparkBench, delivering a dramatic speedup.
- On **power efficiency**, LPDDR5X sustains these performance levels while consuming a smaller fraction of total system power.

As LPDDR5X signaling rates advance toward 9600 MT/s and beyond, the bandwidth advantages observed in workloads like Tensor Rebatching are expected to increase—further improving throughput and reducing latency. On the capacity front, Micron's 256GB SOCAMM2 module enables up to 2TB of LPDDR5X per CPU socket, a configuration that would eliminate disk spilling across an even broader class of large-scale workloads.

Ultimately, LPDDR5X provides a highly compelling combination of data center-class bandwidth, meaningful capacity scaling, and improved energy efficiency, proving it is no longer just an “edge” technology, but a vital and viable standard for hyperscale GPC and AI platforms.

References

[The role of low-power (LP) memory in data center workloads]

[LPDDR at Scale: Enabling Efficient LLM Inference Through High-Capacity Memory]

[Micron’s High-Capacity 256GB LPDRAM SOCAMM2 for Data Center Infrastructure]

https://github.com/facebookresearch/DCPerf

https://github.com/google/multichase

Author Bios

Khayam Anjam is a Staff Engineer at Micron Technology. As part of Micron’s Data Center Workload Engineering team, he works on characterizing and optimizing modern AI workloads across GPU-accelerated and data center platforms. He also contributes to technical papers, internal reports, and customer-facing collateral. Prior to Micron, Khayam designed AI systems for computer vision and anomaly detection. At Dell Technologies, he authored six patents while contributing to ML, analytics, and reliability initiatives. He brings more than 10 years of industry experience spanning performance engineering, generative AI, and distributed systems. Khayam holds an M.S. in Computer Engineering from Clemson University.

Naveen Krishna is a Principal Engineer in Systems Performance at Micron Technology. As part of the Data Center Workload Engineering team, he focuses on performance characterization and optimization of AI and data center workloads, with an emphasis on memory subsystem behavior across emerging architectures. His work spans benchmarking large-scale AI pipelines, analyzing memory bandwidth and capacity trade-offs, and driving insights that inform next-generation memory and system design. He brings over 15 years of industry experience, including prior work in system software design. Currently, his work centers on performance analysis of distributed computing systems, with a strong interest in understanding workload behavior and emerging system architectures to drive performance at scale.

Sujit Somandepalli is a Senior Engineering Manager in the Data Center Workload Engineering group at Micron Technology, Inc., where he leads performance characterization and workload-driven optimization of next-generation memory and storage solutions. His work focuses on bridging application behavior with system-level architecture to drive differentiated value in data center deployments. Sujit brings prior experience from Dell Inc. and Qualcomm Inc., with a strong background in systems design, performance modeling, and workload analysis, and is particularly interested in emerging memory technologies, memory hierarchy tuning, and application-aware system design.

Dr. Sudharshan S. Vazhkudai is a Fellow of systems design engineering at Micron Technology. At Micron, he established the Data Center & Client Workload Engineering team, which brings an end-to-end systems perspective in understanding how deep-memory hierarchy is used to create modern system architectures optimized for workloads. Prior to this, for over two decades, he worked at the US Department of Energy's Oak Ridge National Lab as a Distinguished Scientist and a Program Director, building HW/SW codesign solutions for generations of the world's fastest production supercomputers, serving a large user base in AI and HPC. Dr. Vazhkudai holds a Ph.D. in computer science from the University of Mississippi and has also served as a joint faculty at the University of Tennessee. He has published over a hundred peer-reviewed papers on the fundamental underpinnings of data center systems design, system architecture, deep-memory/storage hierarchy, heterogeneous CPU/GPU architectures, data management, and distributed systems.

Praveen Vaidyanathan is Vice President and General Manager of the Cloud Memory Business Unit at Micron Technology. In this role, Praveen leads Micron's cloud memory strategy and execution, working closely with global hyperscale and enterprise customers. He oversees product line management, applications engineering, and customer qualification teams to deliver next-generation memory solutions optimized for cloud, AI, and data-centric workloads.

Anil Agrawal is a Hardware Systems Engineer at Meta, where he integrates advanced Reliability, Availability, and Serviceability (RAS) technologies into hyperscale compute and AI/ML platforms. With over 40 years of professional experience—including 20+ years in the semiconductor industry—he brings deep expertise in hardware fault management across x86, ARM, and accelerator-based systems. Prior to Meta, Anil spent a decade at Intel developing RAS solutions that significantly reduced server crash rates and delivered over $100M in cost savings. He is a patent holder, a published researcher, and a regular presenter at OCP and JEDEC forums.

Prakash Chauhan is a Hardware Systems Engineer at Meta. He is responsible for evaluation, pathfinding and adoption of critical server technologies into Meta's next-generation compute roadmap. Prakash works closely with industry partners and consortia to produce standards and products that are deployable at scale with high performance and reliability.

Kiran Malwankar is a Hardware Systems Engineer at Meta on the Hardware Design & Release-To-Production (HDRTP) team, specializing in hyperscale infrastructure, advanced memory architectures, and system-level optimization. With decades of experience in high-speed digital and analog hardware design, he has driven multiple products from concept to multi-billion-dollar revenue. He has architected next-generation server and storage systems, and designed industry-first NVMeOF platforms. He holds 25+ patents and has co-authored several papers. His expertise spans DDR, LPDDR, MRDIMM, and ARM/x86 performance tuning. Outside of work, he enjoys travel, culinary exploration, and cricket history, always seeking new ways to learn, refine, and contribute.

micron.com