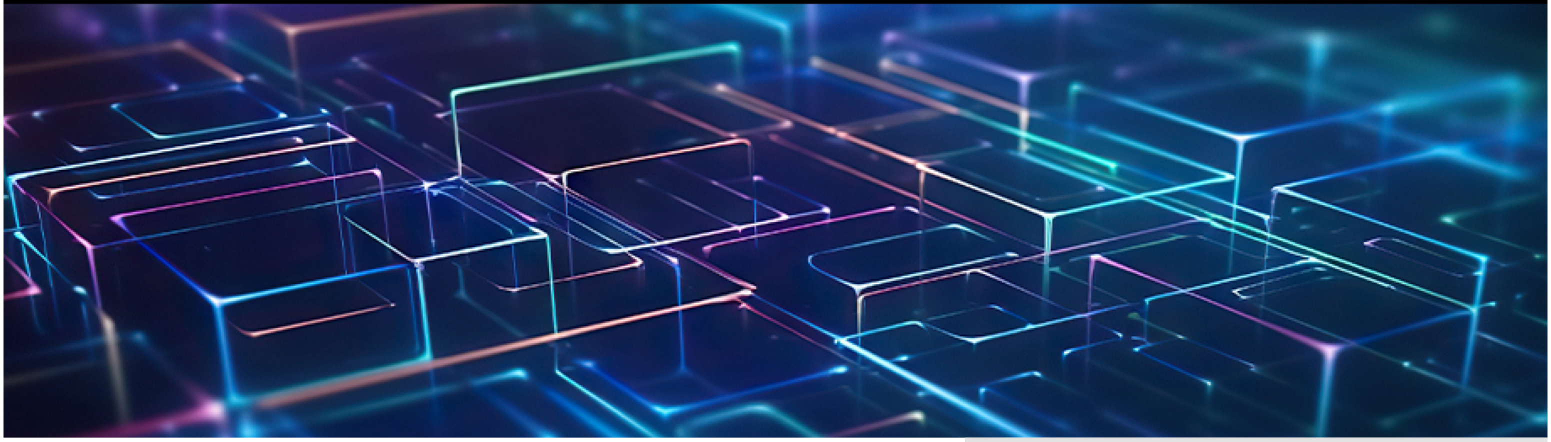

The role of low-power (LP) memory in data center workloads

LP memory, achieving energy efficiency and performance

Micron LPDDR5X is a low-power memory technology, specifically engineered to address the escalating energy demands of modern data centers by delivering high performance while consuming less power than DDR5 memory. By implementing advanced power management features—including lower operating voltages, deep sleep modes, and intelligent power optimization algorithms—LPDDR5X achieves impressive power savings. In data center environments, where every watt matters, this technology helps to reduce operational costs and environmental impact.

It is well-known that AI and high-performance computing (HPC) workloads are increasingly constrained by bandwidth and power. To evaluate the potential value of LPDDR5X against DDR5 for next-generation data center architectures, we comprehensively tested the performance of LPDDR5X across critical metrics: memory bandwidth, power, application runtime, power efficiency (performance per watt), and task energy. Using a diverse test suite—including a microbenchmark (multichase) to assess fundamental memory characteristics, a computationally intensive HPC simulation (POT3D), and an AI model (Llama 3) at varying parameter sizes (8B and 70B)—we provide data-driven insights to highlight how LPDDR5X improves both efficiency and performance in the data center.

For IT leaders facing increasing and ongoing pressure to optimize efficiency while managing increasing energy costs, this report provides a focused analysis of how advanced memory technologies like LPDDR5X help to improve performance and power efficiency for HPC and AI workloads in the data center. Technology investors and decision-makers will find our results particularly valuable in understanding the potential impact of LP memory technologies in addressing the challenges of scaling, power, and performance.

Key takeaways

>75% **lower memory power**

For Multichase and POT3D, LPDDR5X memory consumed up to 77% less power, compared to DDR5 memory.

>35% **higher memory bandwidth**

The Multichase benchmark shows up to 36% higher memory bandwidth for LPDDR5X. Both AI and HPC applications benefit from this higher bandwidth.

10% **better overall system power efficiency**

For Llama 3 (8B), our LPDDR5X system delivers 10% better power efficiency (perf/watt) due to lower power.

>5x **higher throughput**

For Llama 3 (70B), the GH200 system with LPDDR5X delivers more than 5 times higher inference throughput and about 80% lower inference latency. The improvement in performance is due to the combination of Grace CPU, LP memory, and NVLink.

LPDDR in a data center system

To investigate the potential of low-power memory in data center architectures, we selected the NVIDIA Grace Hopper GH200 system—the first commercial implementation featuring LPDDR5X technology—as our primary test platform. This innovative system, combining an ARM CPU with an H100 GPU, represents a cutting-edge approach to high-performance computing infrastructure.

Our comparative analysis benchmarked the LPDDR5X-based Grace Hopper against a contemporaneous DDR5 server configuration from the 2022-2023 product cycle, as detailed in Table 1. By systematically comparing these systems, we aim to quantify the performance and power implications of advanced memory technologies across diverse computational workloads, including microbenchmarks, AI, and HPC applications.

The fundamental architectural difference emerges in memory packaging: LPDDR5X memory is directly soldered onto the Grace Hopper board, in contrast to DDR5 modules with a 64-bit width to the CPU. The Grace Hopper's architecture leverages 32 memory controllers, with each controller managing a 16-bit channel from each LPDDR5X package. This configuration provides more parallelism and efficiency in data handling as each channel can operate independently.

In comparison, the DDR5 system employs a more traditional approach: 4 memory controllers, each with 4 channels of 32 bits (utilizing 2x 32-bit subchannels), totaling 16 channels of 32-bit width. The LPDDR5X configuration offers 4 ranks versus DDR5's 2 ranks, further enhancing access parallelism since each rank operates independently. Performance metrics highlight LPDDR5X's advantages, with a peak theoretical bandwidth of 384 GB/s—slightly higher that DDR5's 358 GB/s. This combination of higher data rate, enhanced parallelism, and greater bandwidth positions LPDDR5X as a superior technology for HPC applications and mixed memory access patterns.

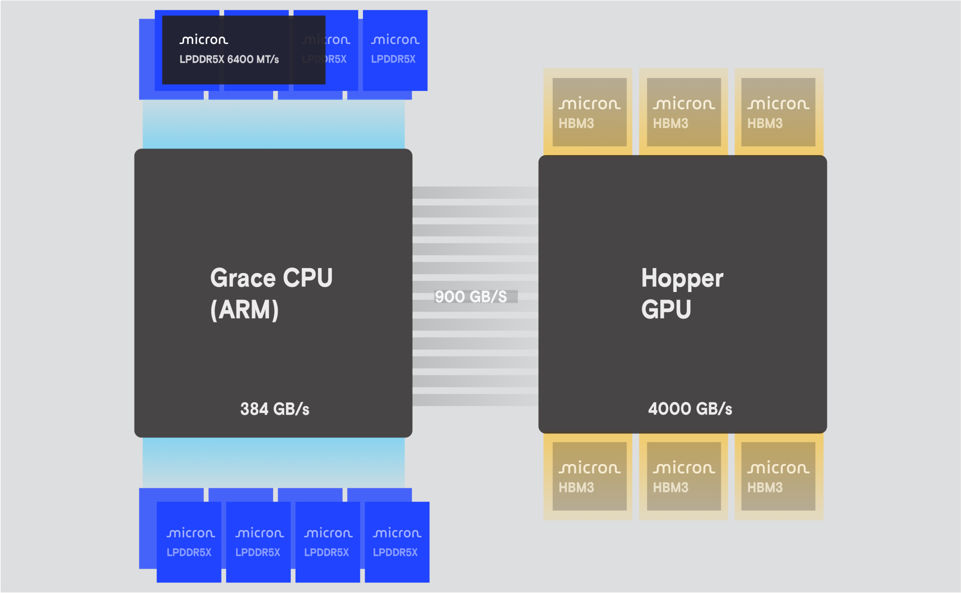

Figure 1: LPDDR5X system – LPDDR5X and ARM architecture

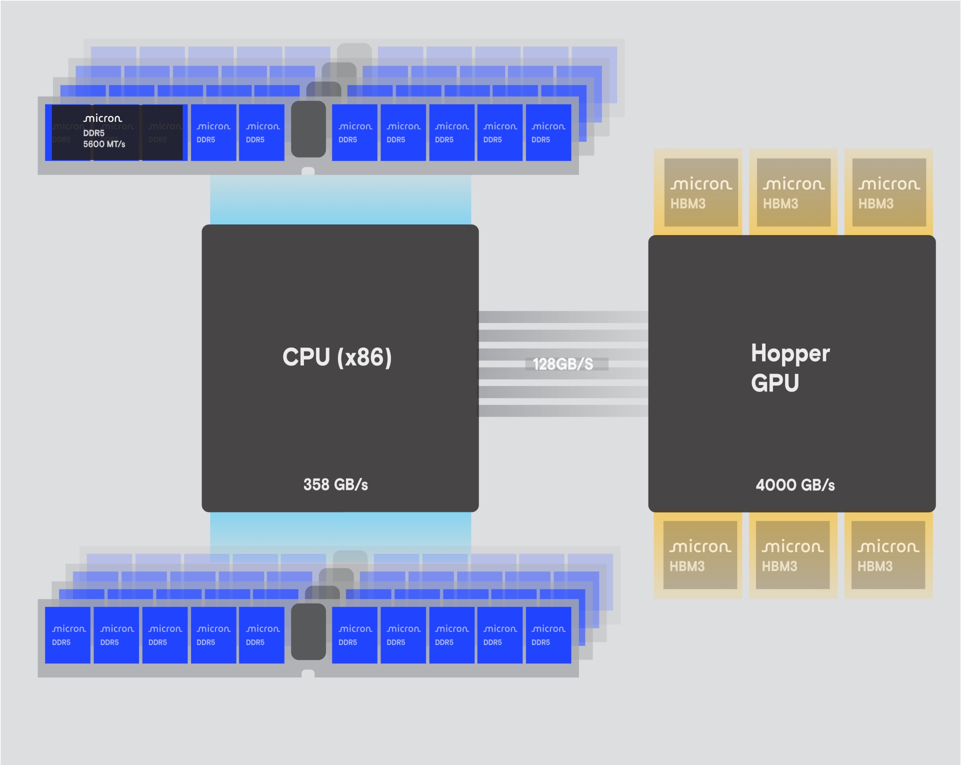

Figure 2: DDR5 system – DDR5 and x86 architecture

System[1]	LPDDR5X System	DDR5 System
Platform	NVIDIA GH200 Grace Hopper Superchip [2]	x86 [2]
Frequency	3.1 GHz	3.9 GHz
Cores	72	64
L3 Cache	114MB	320MB
Memory	LPDDR5X 6400 MT/s Rank = 4 Memory controllers (MC): 32 Channels per MC: 1 Width per channel: 16 bits Total width = 512 bits	DDR5 5600 MT/s Rank = 2 Memory controllers (MC): 4 Channels per MC: 2 Width per channel: 64 bits Total width = 512 bits
GPU	H100 96GB HBM3	H100 96GB HBM3
CPU-GPU Interconnect	NVLink: C2C 900 GB/s bidirectional	PCIe: 128 GB/s bidirectional

Table 1: System configurations

1. CPU for LPDDR5X system: ARM Neoverse® V2; CPU for DDR5 system: Intel® Xeon® Platinum 8592+ Processor

LPDDR5X vs. DDR5 bandwidth and power analysis using microbenchmark

We use the multichase benchmark—a performance testing tool that evaluates memory bandwidth and latency. Multichase performs a series of pointer-chasing operations through an array to simulate different memory access patterns. This benchmark is particularly effective for evaluating how well a memory system handles non-sequential data access (random), which is common in real-world applications. The benchmark's methodology involves multiple threads executing simultaneous memory read and write operations to measure bandwidth. The benchmark records the time taken to complete these operations, allowing it to calculate memory bandwidth in GB/s.

We compiled multichase from the source with default optimization flags for both platforms. Multichase can assess how each memory type performs under various conditions, providing valuable insights into their efficiency and suitability for different applications. Since only a fraction of the compute cores are needed to saturate the multichase memory bandwidth on both the LPDDR5X and the DDR5-based systems, this is a good way to isolate and compare the memory bandwidth and memory power on two different systems that have different instruction set architectures (ISAs).

Results

Our comparative analysis reveals LPDDR5X's improved performance across diverse memory access patterns. In the 1R:1W scenario, LPDDR5X delivers a bandwidth of 293 GB/s—a 36% improvement over DDR5's 215 GB/s. Similar gains are observed in 1R:0W and Stream 2R:1W scenarios, with 11% and 32% performance increases, respectively. These improvements stem from LPDDR5X's innovative architecture: 4R ranks and finer channel granularity that optimize memory access efficiency and reduce latency.

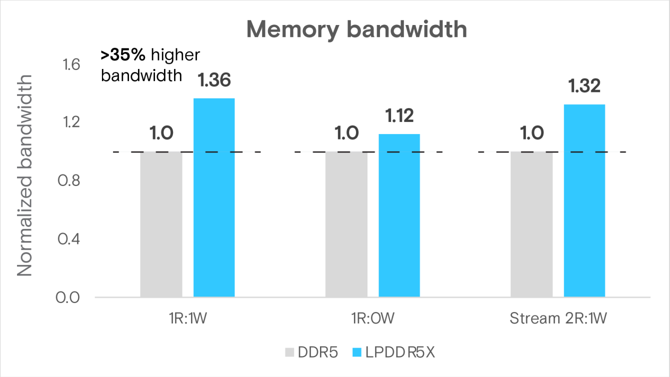

Figure 3: Maximum bandwidth (gigabytes per seconds[GB/s]) for Multichase benchmark

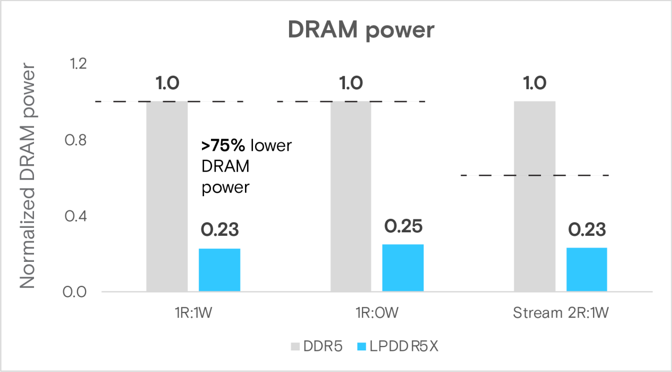

Figure 4: DRAM power (Watts) for multichase benchmark

To measure DRAM power on Grace, we used the NVIDIA DRAM power model for LPDDR5X. For DDR5, power is measured using performance counters. In the 1R:1W scenario, LPDDR5X consumes just 19.9 watts—a significant 77% reduction from DDR5's 86.5 watts. Comparable power savings of 75-77% are evident across different access patterns. LPDDR5X achieves these power savings through advanced features like lower operating voltages, deep sleep modes, and optimized power management algorithms.

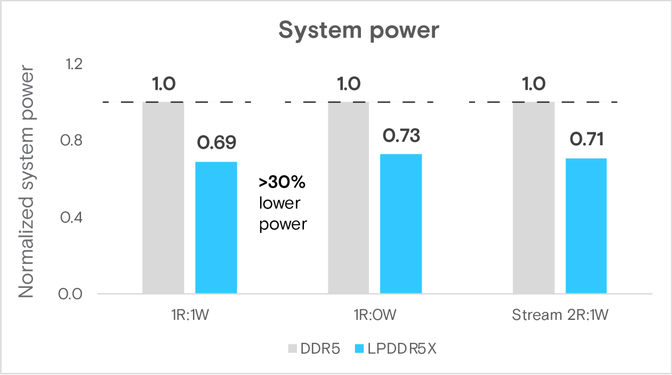

Figure 5: System Power (Watts) for multichase benchmark

System-level power measurements, collected via ipmitool, consistently show LPDDR5X system's lower power consumption, ranging 29-34% lower than the DDR5 system. Although the Grace solution is designed to be power efficient, 40% of the system-level power reduction corresponds to LPDDR5X's lower power draw.

The multichase microbenchmark analysis demonstrates LPDDR5X's bandwidth and power advantages, establishing a promising framework for evaluating its potential in data center applications. The sections that follow will the explore performance and power gains across real-world applications in HPC and AI.

LPDDR5X for HPC in Datacenter

We evaluate LPDDR5X's performance in high-performance computing (HPC) applications using the Solar Physics (POT3D) — a workload that simulates coronal magnetic field dynamics. Our analysis reveals significant performance and efficiency advantages for LPDDR5X-based systems.

The POT3D experiment demonstrated a **10%** runtime improvement on the LPDDR5X system compared to DDR5, driven by enhanced memory bandwidth and architectural differences. LPDDR5X also achieved **20%** better memory bandwidth utilization. Note that HPC applications are bound by memory bandwidth, and typically also follow the trend of memory bandwidth gain. Recall for the multichase microbenchmark that LPDDR5X realized a ~36% memory bandwidth gain (for 1R:1W) compared to DDR5, and herein a real-world application, with its complex data access pattern can achieve more than 55% of that potential. This highlights the opportunity for LPDDR5X to alleviate the bandwidth boundedness of memory-intensive applications.

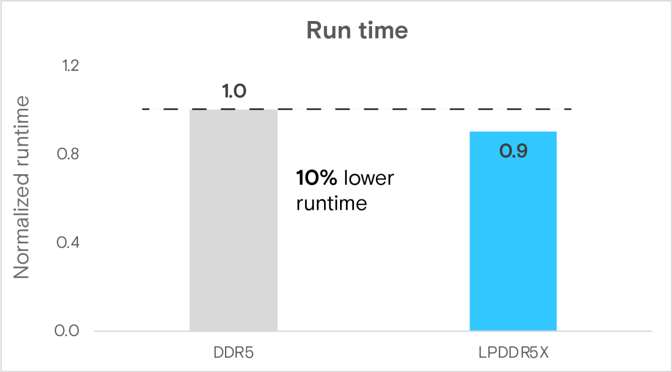

Figure 6: Run time for POT3D (solar physics)

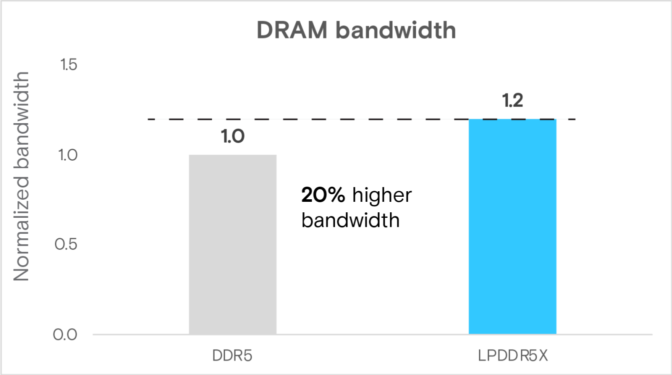

Figure 7: DRAM bandwidth for POT3D (solar physics)

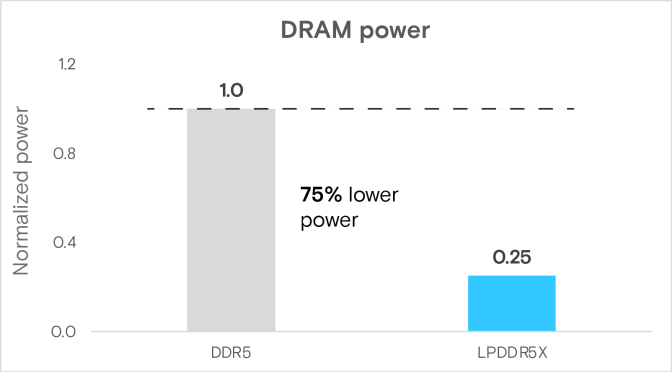

Figure 7: DRAM power for POT3D (solar physics)

Power consumption results were equally compelling: the LPDDR5X memory consumed 75% less power during POT3D execution, consistent with microbenchmark observations. This dramatic reduction not only decreases energy consumption but also contributes to more sustainable data center infrastructure.

In summary, the POT3D analysis highlights LPDDR5X's transformative potential: a 75% DRAM power benefit, 20% memory bandwidth gain, and 10% runtime reduction position it as a promising technology for memory-intensive HPC applications.

LLM inference with LPDDR5X in CPU/GPU Systems

We evaluated LPDDR5X performance for LLM inference across two scenarios: CPU-only and CPU+GPU configurations.

CPU-only | Llama 3 8B

We ran Llama 3 8B model on both LPDDR5X and DDR5 systems. Models in the 8-20 billion parameter range are often considered acceptable for CPU-only execution. The DDR5 system, featuring a high-performance x86 CPU with a 3.9 GHz clock and large last-level cache (L3), demonstrated better raw performance: generating tokens **1.7x** faster with approximately **1.1x** better first-token latency.

However, when evaluating performance per watt—a critical measurement of power efficiency—the LPDDR5X system excelled. Leveraging both LPDDR5X memory and a low-power ARM-based Grace CPU, it achieved a **1.1x** power efficiency improvement, which has the potential to significantly reduce inference deployment costs.

Note: the power results of the multichase benchmark, shown earlier in this report, showed that LPDDR5X contributed to approximately 40% of the overall power reduction.

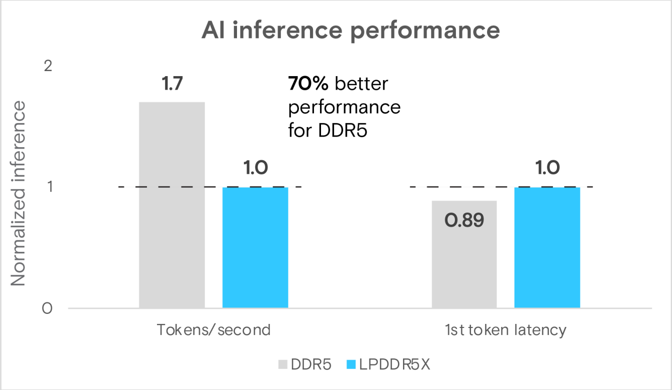

Figure 8: AI inference for Llama 3 8B

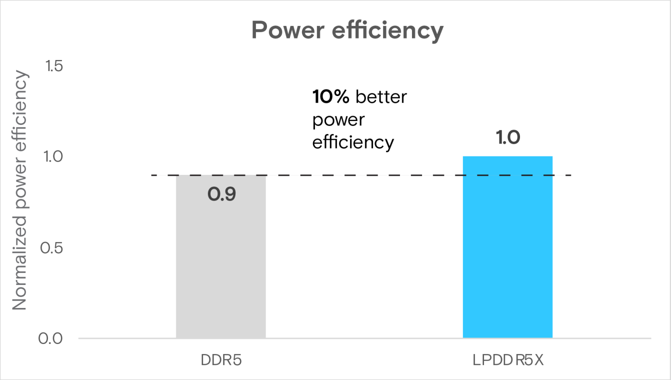

Figure 9: Power efficiency (performance/watt) for Llama 3 8B

CPU and GPU | Llama 3 70B

To better understand the role of LPDDR5X in an inference run on a CPU+GPU scenario, we examined a Llama 3 model with 70 billion parameters. Models of this scale need GPU and HBM resources due to their higher bandwidth and computational requirements. We employed the H100/HBM3 GPU in two configurations:

- LPDDR5X system with H100/HBM3 GPU integrated (NVIDIA Grace Hopper Superchip)
- Standard DDR5 system, where we installed the same H100/HBM3 to be consistent in our comparison

The key differentiator was interconnect performance. The Grace Hopper Superchip features an integrated NVIDIA NVLink with 900 GB/s bidirectional bandwidth, compared to the standard DDR5 system's PCIe Gen5 link, which offers only 128 GB/s bidirectional bandwidth. The LPDDR5X system greatly outperformed the DDR5 system:

- **7x** higher interconnect speed (CPU-GPU)
- **346 GB/s** device-to-host and **334 GB/s** host-to-device transfer speeds, compared to 55 GB/s (unidirectional) for DDR5
- **5x** higher inference throughput
- **80%** lower inference latency

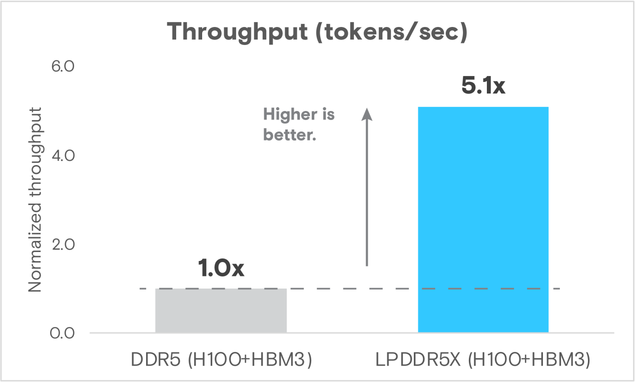

Figure 10: Normalized throughput (tokens per second) Llama 3 70B

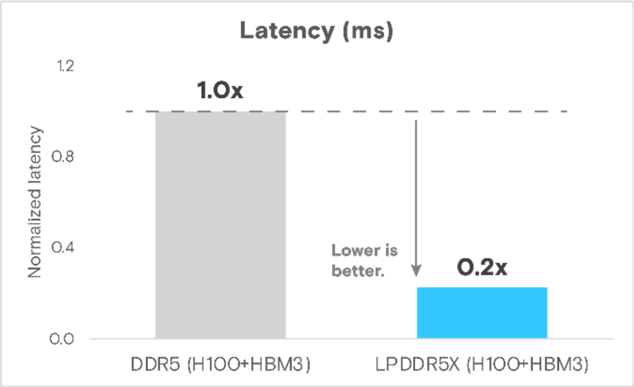

Figure 11: Normalized latency (milliseconds [ms]) Llama 3 70B

When it comes to power and energy-efficiency, we observed that the LPDDR5X system not only completes inference tasks faster but also uses less power and energy. The LPDDR5X DRAM power was 60% less than the DDR5 DRAM only power. Recall that overall, the LPDDR5X DRAM's power savings from both the multichase microbenchmark and HPC is also around 77% and 75%, respectively. Therefore, overall, LPDDR5X offers significant power advantages over DDR5 modules across all workloads.

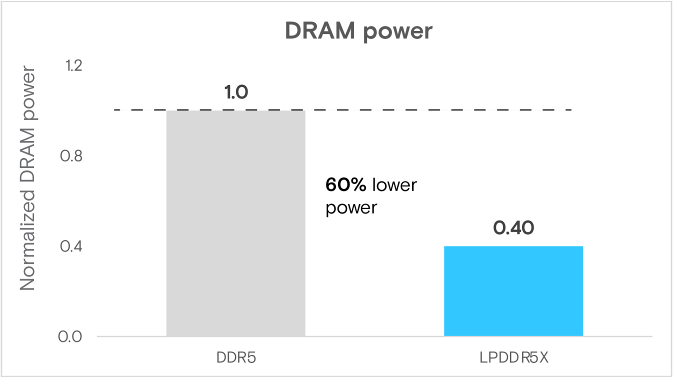

Figure 12: DRAM power for Llama 3 70B

Translating power metrics to task energy, the LPDDR5X system consumed **73%** less energy per task, demonstrating substantial efficiency advantages for large AI workloads. While CPU-only performance favored the DDR5 system, the LPDDR5X system's power efficiency and CPU-GPU integration via NVLink provide compelling benefits for next-generation AI infrastructure. See figure below. Task energy is particularly relevant to data center operators and managers, who are looking to optimize operational expenditure of their infrastructure deployments to support emerging workloads in AI.

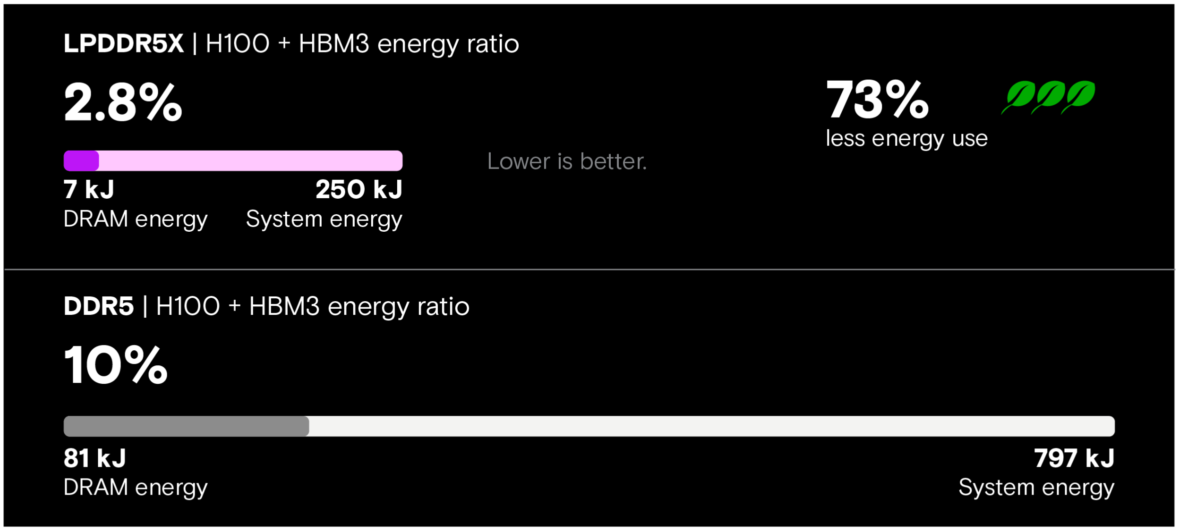

Figure 13: LLM inference energy efficiency

The future of data centers

Low-power memory represents an opportunity for reshaping data center infrastructure, offering an energy-efficient solution to the escalating computational and energy challenges of AI and high-performance computing. Our comprehensive analysis of LPDDR5X reveals transformative potential across critical performance dimensions:

Memory bandwidth: Up to **36%** performance improvement

Application runtime: Enhanced through optimized memory utilization

Power efficiency: Significant **77%** DRAM power reduction by enabling more sustainable, power-efficient computing architectures

Low-power memory technologies can help data centers simultaneously address three main challenges: increasing computational demands, rising energy costs, and environmental sustainability.

- To learn more about how LP memory technologies are fundamentally reshaping the way data centers approach energy efficiency for AI tasks, check out our blog: Every watt matters: How low-power memory is transforming data centers
- Explore our LPDDR5X product page.

References

[1]NVIDIA. (2024). Grace Hopper Superchip datasheet. https://resources.nvidia.com/en-us-data-center-overview-mc/en-us-data-center-overview/grace-hopper-superchip-datasheet-partner

[2] Intel. (2024). 5th Generation Intel® Xeon® Scalable Processors product documentation.

https://www.intel.com/content/www/us/en/products/docs/processors/xeon/5th-gen-xeon-scalable-processors.htm

Authors

Sudharshan Vazhkudai, Henrique Pötter, Khayam Anjam, Moiz Arif

micron.com/ddr5