DRAFT

Low-Power (LP) DDR memory in Datacenters

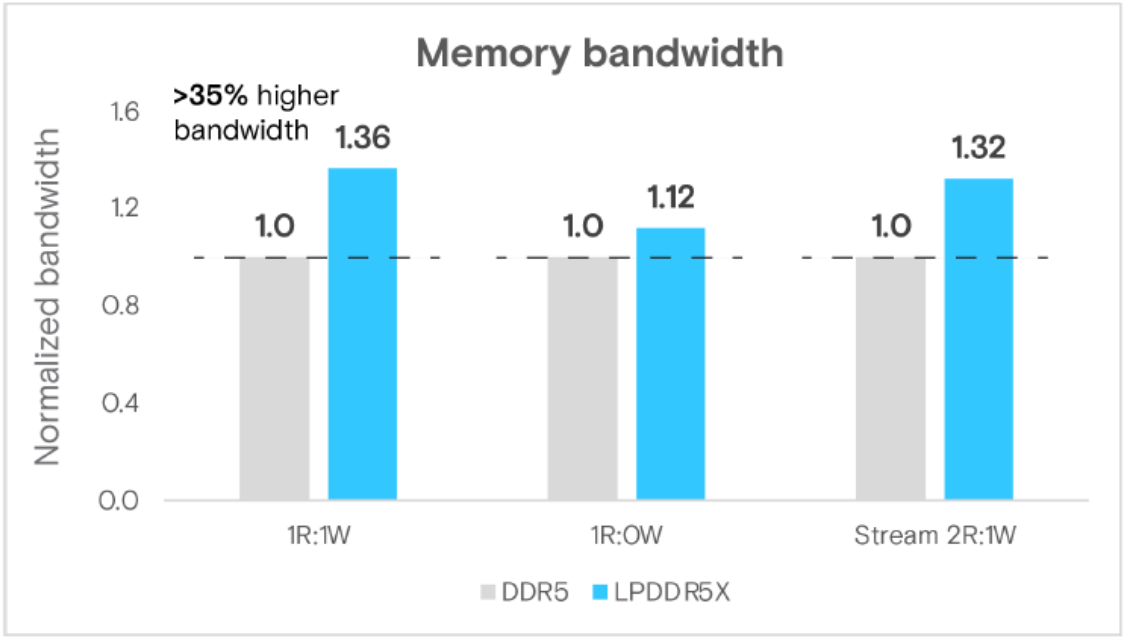
Khayam Anjam

micron

Role of LP memory in Datacenters

- AI and high-performance computing (HPC) workloads are constrained by bandwidth and power.
- To evaluate the performance of LP memory, we comprehensively tested across: memory bandwidth, power, application runtime, power efficiency and task energy.
- We used NVIDIA Grace Hopper GH200 system. This system combine an ARM CPU with an H100 GPU.

Microbenchmarks



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

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.

Microbenchmarks

In the 1R:1W scenario, LPDDR5X consumes just 19.9 watts—a significant 77% reduction from DDR5's 86.5 watts.

LPDDR5X system uses lower power, ranging 29-34% lower than the DDR5 system.

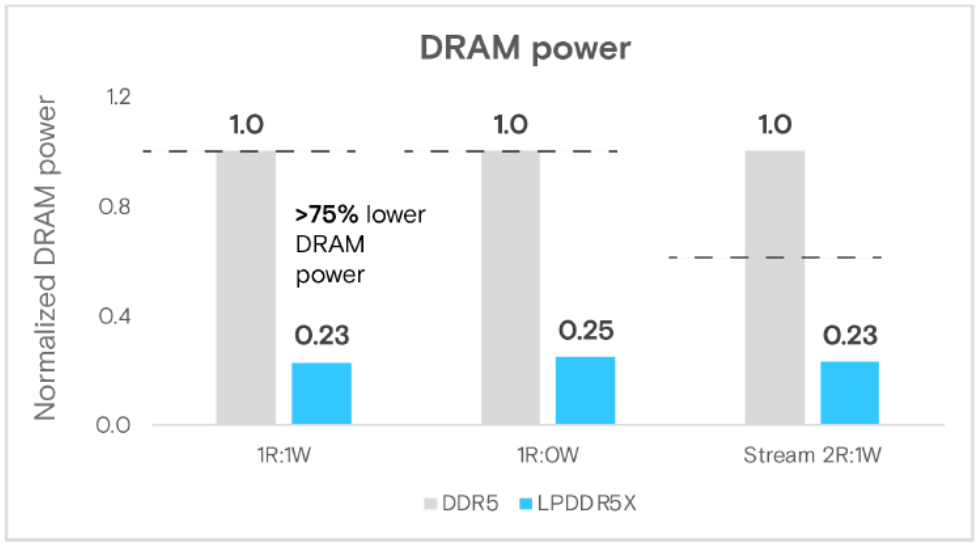


Figure 4: DRAM power (Watts) for multichase benchmark

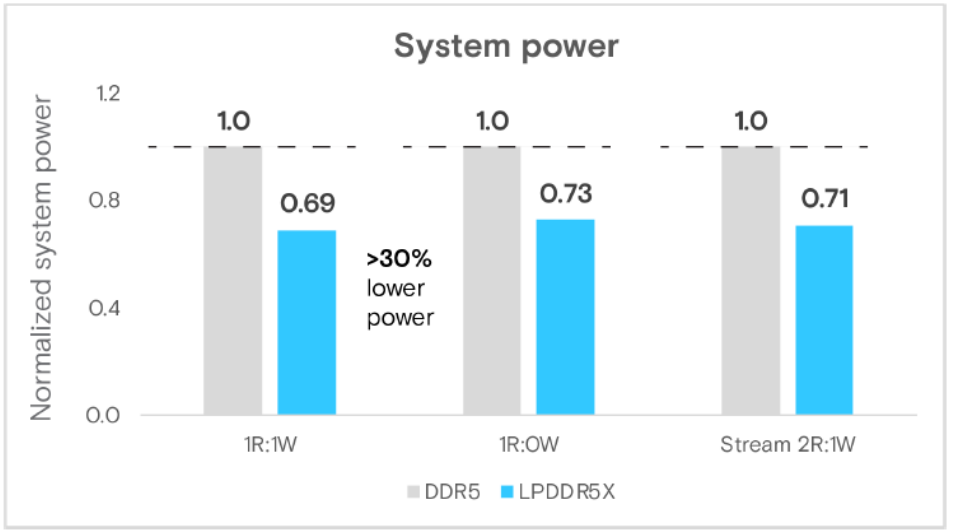


Figure 5: System Power (Watts) for multichase benchmark

LP for HPC in Datacenter

We evaluated LPDDR5X's performance in high-performance computing (HPC) applications using a Solar Physics (POT3D) workload

- **10%** runtime improvement on the LPDDR5X system compared to DDR5.
- LPDDR5X also achieved **20%** better memory bandwidth utilization.
- LPDDR5X memory consumed **75%** less power during POT3D execution

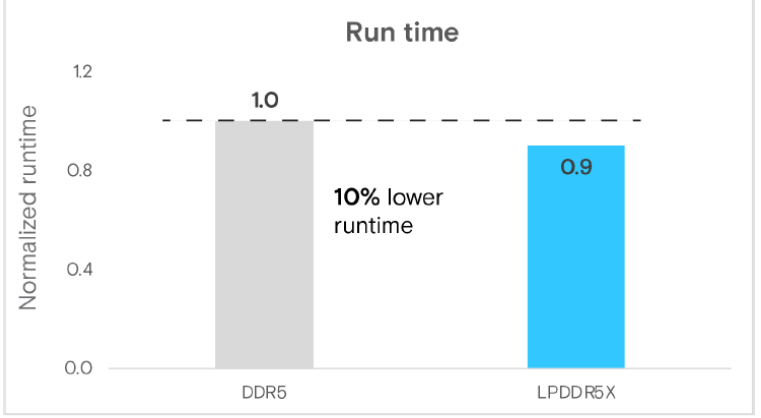


Figure 6: Run time for POT3D (solar physics)

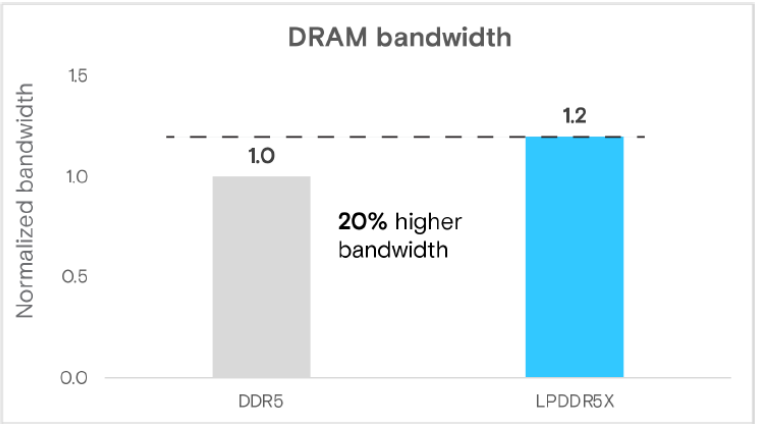


Figure 7: DRAM bandwidth for POT3D (solar physics)

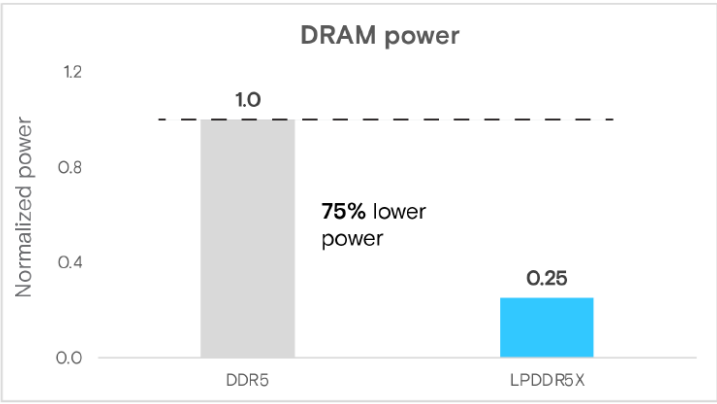


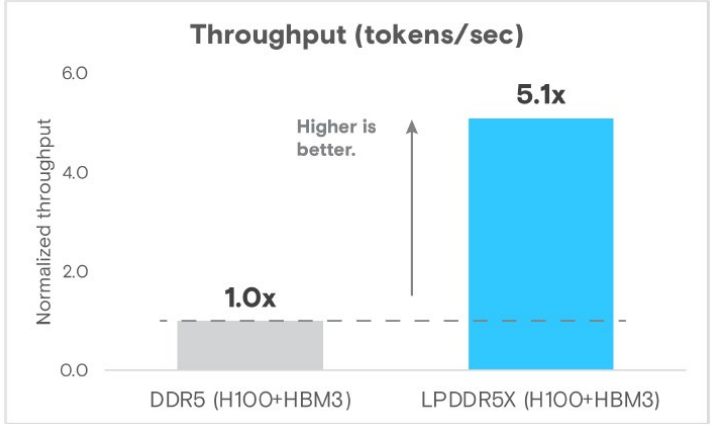
Figure 7: DRAM power for POT3D (solar physics)

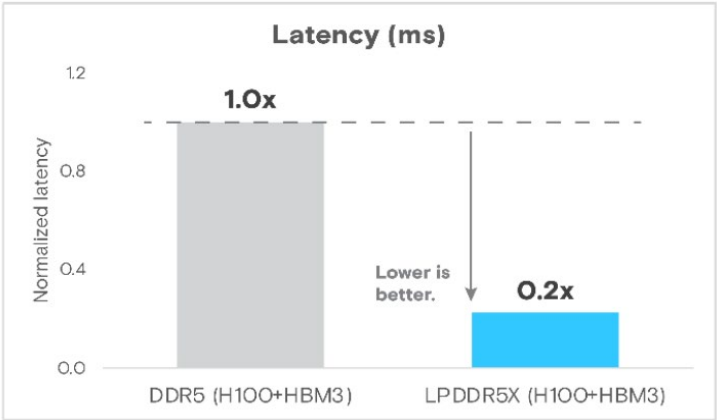
LP for AI in Datacenter

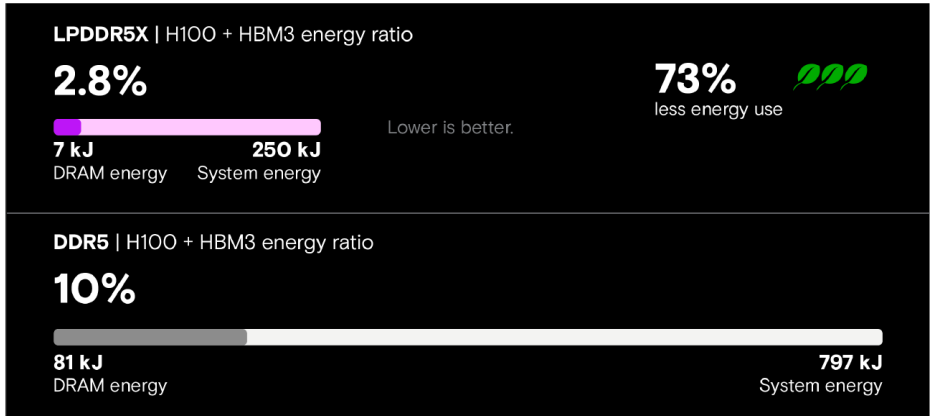
We ran CPU+GPU inference with LLaMA-3 70B.

- 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
- The LPDDR5X DRAM power was 60% less than the DDR5 DRAM only power.
- 73% less energy use by LPDDR5X







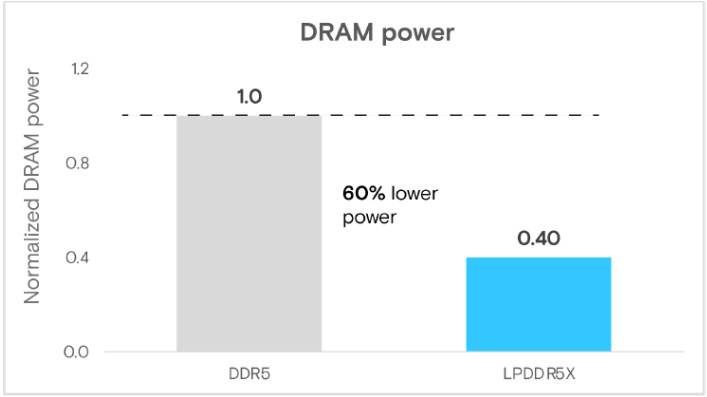
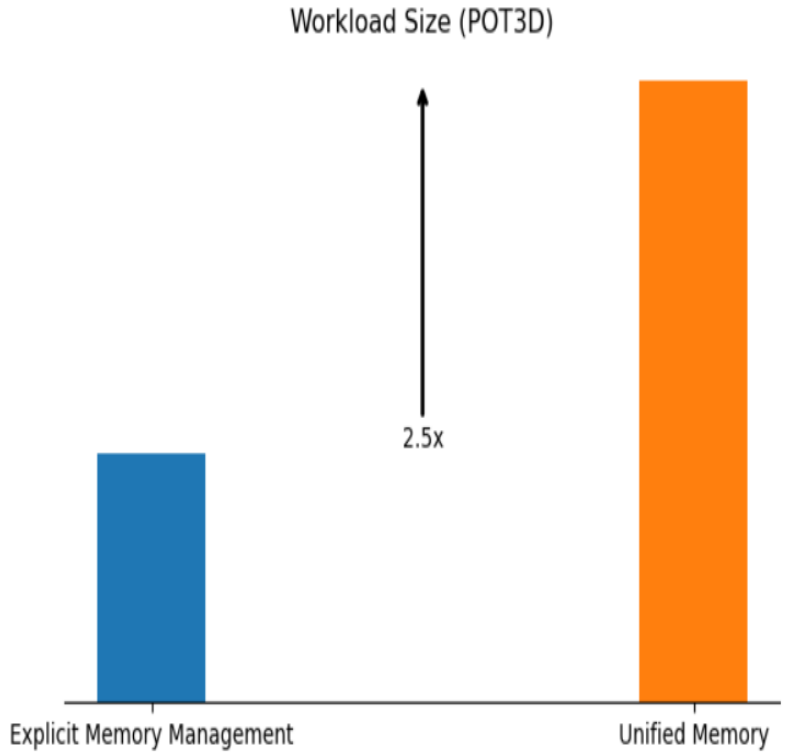


Figure 12: DRAM power for Llama 3 70B

Unified Memory on GH200

- 7x fasters GPU-CPU interconnect, cache coherency and central ATS-TBU architecture makes GH200 an ideal choice to run UM applications.
- Why Use unified memory?
- Results:
  - With unified memory we can run a workload that is 2.5x larger compared to managed memory.



The future of data centers

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.

Checkout our tech brief: Every watt matters: How low-power memory is transforming data centers | Micron Technology Inc.