Micron Sets New Benchmark With the World's First High-Capacity 256GB LPDRAM SOCAMM2 for Data Center Infrastructure

March 3, 2026 at 9:00 AM EST

News highlights:

- **1/3 the power consumption and 1/3 smaller footprint versus standard RDIMMs** — enabled by the industry's first monolithic 32Gb LPDDR5X die
- **2.3 times faster time to first token** for long-context LLM inference, and **3 times better performance per watt** in stand-alone CPU applications
- **1.33 times more capacity per module** — enabling 2TB LPDRAM per 8-channel server CPU for both AI and high-performance compute (HPC)

A Media Snippet accompanying this announcement is available by clicking on this link.

BOISE, Idaho, March 03, 2026 (GLOBE NEWSWIRE) -- Micron Technology, Inc. (Nasdaq: MU) today extended its leadership in low-power server memory by shipping customer samples of the industry's highest-capacity LPDRAM module — 256GB SOCAMM2. Enabled by the industry's first monolithic 32Gb LPDDR5X design, this milestone represents a transformational step forward for AI data centers, delivering low-power memory capacity that can unlock new system architectures.

The convergence of AI training, inference, agentic AI and general-purpose compute are driving more demanding memory requirements and reshaping data center system architectures. Modern AI workloads drive large model parameters, expansive context windows and persistent key value (KV) caches, while core compute continues to scale in data intensity, concurrency and memory footprint.

Across these workloads, memory capacity, bandwidth efficiency, latency and power efficiency have become primary system level constraints, directly influencing performance, scalability and total cost of ownership. LPDRAM's unique combination of these attributes position it as a cornerstone solution for both AI and core compute servers in increasingly power and thermally constrained data center environments. Micron is collaborating with NVIDIA to co-design sophisticated memory for the needs of advanced AI infrastructure.

"Micron's 256GB SOCAMM2 offering enables the most power-efficient CPU-attached memory solution for both AI and HPC. Today's announcement highlights Micron's technology and packaging advancements to deliver the highest-capacity, lowest-power modular memory solution with the smallest footprint in the industry," said Raj Narasimhan, senior vice president and general manager of Micron's Cloud Memory Business Unit. "Our continued leadership in low-power memory solutions for data center applications has uniquely positioned us to be the first to deliver a 32Gb monolithic LPDRAM die, helping drive industry adoption of more power-efficient, high-capacity system architectures."

Designed for capacity, power efficiency and workload performance optimization
Micron's 256GB SOCAMM2 delivers higher memory capacity, substantially lower power consumption and faster performance for a variety of AI and general-purpose computing workloads.

- **Expanded memory capacity for AI servers:**
  With one-third more capacity than the prior highest capacity 192GB SOCAMM2, 256GB SOCAMM2 provides 2TB of LPDRAM per 8-channel CPU for larger context windows and complex inference workloads.
- **Lower power consumption and smaller footprint:**
  SOCAMM2 consumes one-third of the power compared with equivalent RDIMMs, while using only one-third of the footprint, improving rack density and reducing the total cost of ownership.[1]
- **Improved inference and core compute performance:**
  In unified memory architectures, 256GB SOCAMM2 improves time to first token by more than 2.3 times for long context, real-time LLM inference when used for KV cache offload compared to currently available solutions.[2] In standalone CPU applications, LPDRAM delivers more than 3 times better performance per watt than mainstream memory modules for high-performance computing workloads.[3]
- **Modular design for serviceability and scalability:**
  The modular SOCAMM2 design improves serviceability, supports liquid-cooled server architectures and enables future capacity expansion as AI and core compute memory requirements continue to grow.

"Advanced AI infrastructure requires incredible optimization at every layer to maximize performance and efficiency for demanding AI reasoning workloads," said Ian Finder, head of Product, Data Center CPUs at NVIDIA. "Micron's achievements in delivering massive memory capacity and bandwidth using less power than traditional server memory with 256GB SOCAMM2 is enabling the next generation of AI CPUs."

Driving industry standards and accelerating low-power memory adoption
Micron continues to play a leading role in the JEDEC SOCAMM2 specification definition and maintains deep technical collaborations with system designers to drive industry-wide improvements in power efficiency and performance for next-generation data center platforms.

Micron is now shipping customer samples of its 256GB SOCAMM2 and offers the industry’s broadest data center LPDRAM portfolio, spanning 8GB to 64GB components and 48GB to 256GB SOCAMM2 modules.

Additional resources:

- LPDDR at Scale: Enabling Efficient LLM Inference Through High-Capacity Memory
- Every watt matters: How low-power memory is transforming data centers
- SOCAMM2 webpage
- Data center memory webpage

About Micron Technology, Inc.

Micron Technology, Inc. is an industry leader in innovative memory and storage solutions, transforming how the world uses information to enrich life for all. With a relentless focus on our customers, technology leadership, and manufacturing and operational excellence, Micron delivers a rich portfolio of high-performance DRAM, NAND and NOR memory and storage products. Every day, the innovations that our people create fuel the data economy, enabling advances in artificial intelligence (AI) and compute-intensive applications that unleash opportunities — from the data center to the intelligent edge and across the client and mobile user experience. To learn more about Micron Technology, Inc. (Nasdaq: MU), visit micron.com.

Micron Product and Technology Communications Contact:
Mengxi Liu Evensen
+1 (408) 444-2276
productandtechnology@micron.com

Micron Investor Relations Contact:
Satya Kumar
+1 (408) 450-6199
satyakumar@micron.com

[1] One-third of the power consumption calculated based on watts of power used by one 128GB, 128-bit bus width SOCAMM2 module compared to two 64GB, 64-bit bus width DDR5 RDIMMs. One-third footprint calculation compares SOCAMM2 area (14x90mm) versus a standard server RDIMM.

[2] Results are based on Micron internal testing of real-time inference with Llama3 70B model (with FP16 quantization) using 500K context length and 16 concurrent users. The projected TTFT latency improvement is based on a latency of 0.12s for 2TB LPDRAM per CPU vs. 0.28s for 1.5TB LPDRAM per CPU. See our whitepaper published earlier this month for more detail on test conditions: LPDDR at Scale: Enabling Efficient LLM Inference Through High-Capacity Memory.

[3] Micron internal testing measuring Pot3D solar physics HPC code performance on identical capacities of LPDDR5X and DDR5.