Astera Labs Bolsters Leo Smart Memory Controller Family for Agentic AI and General-Purpose Cloud Workloads

September 15, 2026

As high-value AI applications drive explosive memory-footprint growth amid tight supply, new connectivity architectures enhance optimal memory utilization and reuse

Highlights

- **New Leo X-Series optimizes token economics:** Paired with Astera Labs' Scorpio X-Series Fabric Switches, Leo X-Series enables fabric-attached memory for long-context, KV-cache-intensive workloads. Platform-specific interfaces and customizable features improve inference performance with up to 62% faster time to first token (TTFT) and up to 22% more tokens per second (TPS). [1]
- **New Leo 2 E-Series and P-Series address CPU-attached capacity and rack-scale memory**: The expanded Leo family supports CPU-attached DDR4 & DDR5 memory expansion with PCIe® 6, single x16 and dual x8 host connectivity and CXL® 3.2 support. The new E & P-Series Smart Memory Controllers double memory bandwidth and capacity versus the previous generation.
- **Reliable memory reuse improves infrastructure TCO:** Hyperscale-grade RAS, memory-health management, on-chip hardware engines, and automated repair enable reliable reuse of previously deployed DDR4 DIMMs alongside DDR5 memory in new, high-volume cloud server fleets.
- **Growing customer engagement and design-win momentum:** The expanded Leo family is driving increased design activity and new design wins across a broadening set of AI labs, hyperscale and neocloud customers.

SAN JOSE, Calif., Sept. 15, 2026 (GLOBE NEWSWIRE) -- **Astera Labs, Inc. (Nasdaq: ALAB),** a leader in semiconductor-based connectivity solutions for rack-scale AI infrastructure, today announced three new memory connectivity solutions—Leo X-Series and the Leo 2 E & P-Series—extending its hyperscale-proven Leo Smart Memory Controller™ family across direct fabric-attached GPU memory and CPU-attached expansion and pooled/shared memory architectures.

"Agentic AI adoption is outpacing the industry's ability to provision memory for it, and that gap is widening every quarter," said Matt Kimball, vice president and principal analyst at Moor Insights & Strategy. "That growth is landing at the worst possible time for memory supply, with DDR5 tight and pricing climbing, so every gigabyte already deployed has to work harder. The new Leo X-Series and Leo 2 E and P-Series Smart Memory Controllers immediately address KV-cache-intensive AI workloads, improve memory utilization, and reliably reuse previously deployed memory across both AI and general-purpose cloud infrastructure, instead of simply buying more of it."

New Leo X-Series: direct fabric-attached memory for agentic AI
The new Leo X-Series is a fabric-attached Smart Memory Controller purpose-built to directly support AI accelerator-side memory demand. Paired with Astera Labs' Scorpio fabric switches, using PCIe and platform-specific protocols for scaling up GPUs, Leo X-Series enables direct connectivity to AI fabrics and a dedicated memory tier for offloading KV cache and agent context. As context windows grow and multi-turn sessions retain more prior tokens, larger KV-cache capacity with low-latency and high bandwidth access becomes critical for the highest-value agentic AI inference workloads, helping minimize response times and deliver a more responsive user experience.

By connecting memory directly to the scale-up fabric, Leo X-Series provides a lower-latency and higher bandwidth path for GPU-to-KV-cache access versus architectures that rely solely on CPU-attached memory or NVMe storage tiers. Leo X-Series' support for PCIe in addition to platform-specific custom interfaces allows hyperscalers and AI platform providers to build a customized memory companion architecture suited to their compute platform rather than a one-size-fits-all design. These capabilities are designed to improve KV-cache performance and token economics, delivering up to 62% lower TTFT and up to 22% more TPS. [1]

Enhanced Leo 2 E-Series and Leo 2 P-Series for memory expansion and pooling/sharing
The enhanced Leo 2 E & P-Series provide complementary approaches to CPU-attached capacity and rack-scale memory utilization that can support AI agents, in-memory databases, and general-purpose cloud workloads. These devices support four DDR4 or DDR5 memory controllers that double memory bandwidth and capacity versus the previous generation, while an optimized chip package fits add-in cards and other designs to optimize DIMM integration density.

- **Leo 2 E-Series** provides direct CPU-attached CXL 3.2 memory expansion via PCIe 6 x16 host connectivity. It delivers additional memory capacity without requiring another CPU socket and enables infrastructure providers to redeploy previously deployed memory in new, state-of-the-art cloud servers supporting high-volume workloads.

- **Leo 2 P-Series** enables disaggregated CXL memory architectures with pooling and sharing across hosts, supported by

dual-port PCIe 6 x8 connectivity and dynamic capacity management. Hosts can draw pooled memory on demand instead of over-provisioning every server, turning stranded DRAM into a rack-level resource increasing utilization which is critical in a memory supply constrained environment.

A common foundation for reliable, interoperable memory deployment and re-use
Across the Leo family, hyperscale-grade RAS and memory-health management support reliable operation for evolving compute and memory platforms. Purpose-built memory test engines and automated repair engines help extend maximum lifetime for existing memory DIMMs, assist with pre-deployment testing, and identify reliable DIMMs for reuse. Enhanced memory error reporting, event recording, scrubbing, customizable thermal management, and resilient firmware updates help protect workloads and extend memory service life. Workload monitoring, performance profiling, hotness tracking, software-defined data placement, and latency optimization help fine-tune memory placement and access latency for long-context agentic AI and general purpose workloads. Telemetry and management integration, together with Astera Labs' COSMOS software suite, provides fleet-wide visibility as operators combine previously deployed DDR4, newly deployed DDR5 in modern cloud server fleets, pooled, and accelerator-attached capacity across the rack.

"Agentic AI is where the economics of AI infrastructure are being decided, and those economics depend on putting every usable gigabyte of memory to work," said Thad Omura, senior vice president, Compute Connectivity Group at Astera Labs. "The enhanced Leo family gives infrastructure providers purpose-built ways to connect memory to accelerators, CPUs, and hosts across the rack—turning previously deployed and stranded capacity into a resource that new AI and cloud workloads can use. We're seeing that translate into broader customer engagement across AI labs, hyperscalers, and neoclouds."

The enhanced Leo family was developed by Astera Labs in close collaboration with CPU and GPU ecosystem partners, including AMD, Arm, Intel, and major memory suppliers, alongside additional hyperscale and OEM collaborators. The family is sampling today with hyperscaler customers.

Astera Labs will demonstrate Leo family capabilities—including direct fabric-attached memory for Agentic AI, DDR4 reuse, dynamic memory pooling, and Agentic AI memory tiering with Scorpio X-Series—at AI Infra Summit 2026 at the Santa Clara Convention Center, September 15–17.

Additional Resources:

- Leo Smart Memory Controllers webpage: https://www.asteralabs.com/products/leo-smart-memory-controllers/
- "Supercharging Agentic AI: How Leo X-Series Smart Memory Controllers Accelerate Inference With KV Cache Offload" blog: www.asteralabs.com/resources/blog/supercharging-agentic-ai-how-leo-x-series-smart-memory-controllers-accelerate-inference-with-kv-cache-offload/

Ecosystem Support:
Robert Hormuth, corporate vice president, Architecture and Strategy at AMD, said:
"Agentic AI is increasing demand for balanced systems that can move and process more data without sacrificing efficiency. AMD EPYC processors deliver leadership performance, exceptional memory bandwidth and capacity, and high-speed I/O across the broadly deployed x86 platform. Our collaboration with Astera Labs around CXL extends that foundation, giving customers greater flexibility to scale memory-intensive AI and cloud workloads and improve infrastructure utilization."

Eddie Ramirez, Vice President of Go-To-Market, Cloud AI, Arm, said:
"As AI infrastructure evolves, agentic AI, reinforcement learning and database workloads are driving demand for higher-capacity memory. Arm and Astera Labs are working together to pair Arm AGI CPU with Leo CXL Smart Memory Controllers, helping customers meet these growing requirements while maintaining scalable performance."

Srini Krishna, Fellow at Intel Data Center Group, said:
"CXL was created to bring greater flexibility and efficiency to memory across the data center. Technologies like Astera Labs' Leo Smart Memory Controllers help turn that vision into reality, demonstrating how ecosystem innovation can help customers get more out of their AI infrastructure and support the next generation of data-intensive workloads."

Jangseok Choi, Vice President of Product Planning Team at Samsung Electronics, said:
"We are contributing to the evolution of AI infrastructure through the development of CXL memory solutions designed to help address the growing memory demands of modern data centers. We are excited to collaborate with Astera Labs to help ensure our DRAM solution offer strong interoperability within the evolving CXL ecosystem, and remain committed to supporting a scalable, high-performance memory architecture for the AI era."

About Astera Labs
Astera Labs (Nasdaq: ALAB) provides rack-scale AI infrastructure through purpose-built connectivity solutions. By collaborating with hyperscalers and ecosystem partners, Astera Labs enables organizations to unlock the full potential of modern AI. Astera Labs' Intelligent Connectivity Platform integrates CXL®, Ethernet, NVLink Fusion, PCIe®, and UALink™ semiconductor-based technologies with the company's COSMOS software suite to unify diverse components into cohesive, flexible systems that deliver end-to-end scale-up and scale-out connectivity. The company's custom connectivity solutions business complements its standards-based portfolio, enabling customers to deploy tailored architectures to meet their unique infrastructure requirements. Discover more at www.asteralabs.com.

[1] As measured by Astera Labs internal testing on an Intel GNR server with 8 DDR5 memory channels populated, using a data center Gen 5 AI accelerator on the Qwen2.5-32B AI Model.

Forward-Looking Statements
This communication contains forward-looking statements within the meaning of the federal securities laws, including statements regarding the expected capabilities, performance, benefits, availability, customer adoption, commercial success and market opportunity of Astera Labs' expanded Leo Smart Memory Controller family of products, as well as related market trends and conditions. Forward-looking statements may be identified by words such as “allows,” “built to,” “can,” “climbing,” “could,” “designed,” “growth,” “vision,” “will” and similar expressions.

These forward-looking statements are based on current expectations and assumptions and involve risks and uncertainties that could cause actual results to differ materially from those expressed or implied by such statements. These risks and uncertainties include, among others, the risk that demand for agentic AI, AI inference, memory-intensive workloads and related infrastructure does not develop as anticipated; memory market conditions, including supply, availability and pricing trends, do not evolve as expected; customers do not adopt CXL, fabric-attached memory, memory pooling, sharing, expansion or other next-generation memory architectures; Astera Labs' products do not achieve the expected performance, utilization, reliability, interoperability, memory reuse, infrastructure efficiency or economic benefits in customer environments; benchmark, testing or simulation results are not indicative of real-world performance; customer evaluations, engagements, design activity, design wins or sampling programs do not result in production deployments or revenue; Astera Labs and its ecosystem partners are unable to develop, manufacture, qualify, deploy or support products and technologies on expected timelines or at expected cost; changes in technology, customer requirements, competitive offerings or industry standards; supply chain constraints or disruptions; litigation or disputes relating to our products or technologies; global economic, political and industry conditions, including those affecting the semiconductor and AI infrastructure markets; regulatory developments; international conflicts; and other risks and uncertainties described in Astera Labs' Annual Report on Form 10-K, Quarterly Reports on Form 10-Q and other filings with the Securities and Exchange Commission.

Forward-looking statements speak only as of the date they are made. Readers are cautioned not to place undue reliance on forward-looking statements, and Astera Labs undertakes no obligation to update or revise any forward-looking statement, whether as a result of new information, future events or otherwise, except as required by law.

Media Contact:
Peter Lo
Peter.lo@asteralabs.com