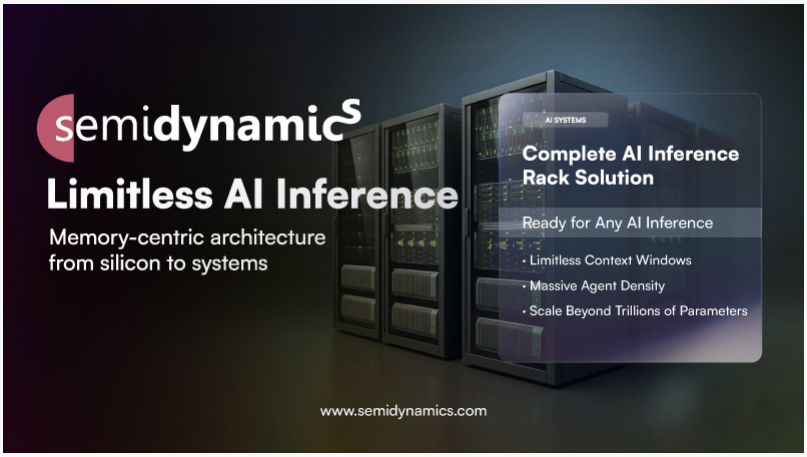
Semidynamics brings its full inference stack to ISC HPC 2026

AI compute is only as useful as the memory architecture feeding it



Semidynamics ISC HPC 2026

BARCELONA, SPAIN, June 22, 2026 /EINPresswire.com/ -- Building high-performance AI silicon is necessary but no longer sufficient: the compute is only useful if the memory system can keep it fed. That is the argument Semidynamics CEO Roger Espasa will bring to ISC High Performance 2026 in Hamburg (23–25 June, booth A22, Hall H), where the Barcelona-based company — which taped out its first 3nm chip with TSMC in December 2025 — will present its plans for a complete memory-centric inference platform, from RISC-V core to liquid-cooled, OCP-compliant rack, at a major HPC trade show for the first time.

> Compute you can't feed is wasted silicon. The hard, still-unsolved problem is the memory architecture that keeps every tensor unit working, and that is what we designed from the core up."
>
> *Semidynamics CEO Roger Espasa*

"You need serious compute to play in AI inference, and we built it," said Espasa. "But compute you can't feed is wasted silicon. The hard, still-unsolved problem is the memory architecture that keeps every tensor unit working, and that is what we designed from the core up. Our 3nm silicon is the proof it works."

Breaking the memory wall

Peak TOPS, the industry's default yardstick, tells you what a chip can do in principle, not what it delivers in production. As models grow and context lengths stretch, it is memory bandwidth and capacity that determine whether that compute is actually used — and what each token costs. Conventional infrastructure caps memory with expensive, supply-constrained HBM and pages KV-caches in and out as it runs short, leaving tensor units stalled and rated performance unavailable in practice.

"Every operator we talk to has the same problem: racks that look formidable on a spec sheet but

sit half-starved because the memory system can't feed them," said Espasa. "The TOPS are real; the utilisation isn't. We engineered the compute and the memory architecture together, so the tensor units stay fed, and that is why the economics work."

Crucially, the differentiator is not the memory itself — high-capacity LPDDR is a commodity any vendor can buy. It is the architecture that turns that memory into usable compute by keeping the tensor units continuously fed. Semidynamics' Gazzillion™ subsystem is engineered to tolerate memory latency end-to-end, which is what lets the platform be built around inexpensive, high-capacity LPDDR from the core up rather than premium HBM. That capability is designed into the RISC-V core and proven in the 3nm SOC — which is why the silicon milestone matters: the answer to the memory wall lives in the architecture, not in a faster memory part. Built this way, the platform delivers multiples of conventional rack memory capacity, very large resident KV-caches, and sustained throughput at high user concurrency.

One architecture, four levels of scale

At ISC HPC 2026, Semidynamics will present its inference platform across four levels: the Inference Engine, an out-of-order 64-bit RISC-V core with integrated vector and tensor units and the Gazzillion memory subsystem built in; the Inference SOC, a 3nm device combining multiple Inference Engines with the ability to run standard Linux workloads; the Inference Board, pairing a general-purpose host with Inference SOCs over a high-bandwidth fabric engineered for persistent KV-cache residency across long context lengths; and a liquid-cooled, OCP-compliant Inference Rack ready for standard data centre integration.

The December 2025 tape-out is among the first 3nm tape-outs achieved by a European semiconductor company and will be followed by a production tape-out later this year.

A European platform, at a European moment

ISC 2026 takes place as Europe accelerates investment in sovereign AI capacity through the EuroHPC AI Factory and Gigafactory programmes. Semidynamics recently announced a strategic cooperation with SiPearl, the European designer of high-performance CPUs, to develop an EU-sovereign, OCP-based rack-scale AI compute platform for large-scale cloud inference — combining SiPearl's Arm®-based CPU for host compute and orchestration with Semidynamics' RISC-V-based inference accelerator. The company has also secured a strategic investment from SK hynix to co-optimise its architecture with next-generation memory technologies.

"Europe has a real chance here," Espasa added, "but not by copying an architecture designed somewhere else and shipping it late. Sovereignty is about controlling the design, not just the location. Memory architecture is where the basis of competition can still change — and that is a contest Europe can win. Designed in Europe, taped out at 3nm, built on open standards: that is a platform Europe can actually own."

No migration, no lock-in

Semidynamics software stack — including the Aliado Orchestrator and the AKL (Aliado Kernel Library) — runs the tools AI teams already use, including vLLM, PyTorch and ONNX Runtime, with support for models such as Llama and DeepSeek directly from Hugging Face. Inference works out of the box, with no proprietary migration path.

David Harold
Foundational Marketing
email us here
Visit us on social media:
LinkedIn

This press release can be viewed online at: https://www.einpresswire.com/article/921332691

EIN Presswire's priority is source transparency. We do not allow opaque clients, and our editors try to be careful about weeding out false and misleading content. As a user, if you see something we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.
© 1995-2026 Newsmatics Inc. All Right Reserved.