US Semiconductors

AI Infra Summit 2026 Takeaways

Equities

Americas

Semiconductors

Timothy Arcuri
Analyst
timothy.arcuri@ubs.com
+1-415-352 5676

Natalia Winkler, CFA
Analyst
natalia.winkler@ubs.com
+1-415-352 4626

Seth Cho
Associate Analyst
seth.cho@ubs.com

Summary

This past week we attended the AI Infrastructure Summit at the Santa Clara Convention Center - the event brings together engineers and executives to discuss compute, networking, systems, and software to build next-generation AI factories with a presence from all major AI companies. We emerged thinking 1) NVDA's flexibility and system-level expertise are under-appreciated by the market amid the focus on individual compute architectures (ASIC, fast inference, etc); 2) memory is still the biggest bottleneck and it's not getting any better; and 3) AI is accelerating compute design cycles and helping hyperscalers ramp ASIC solutions faster - whether this will accelerate timelines to internalize these ASICs remains to be seen.

- **First, inference demand remains insatiable, but the problem for compute operators is becoming more complicated.** They increasingly need to optimize the economics, particularly cost/token, for today's model generation while preserving enough hardware fungibility to support workloads that may look very different several years from now. This tension between optimization and flexibility came up repeatedly, and we think it is becoming an increasingly important consideration in how AI infrastructure gets designed. At the same time, we continue to see up-and-coming solutions from vendors like CBRS, SambaNova, Etched increasingly in the spotlight.
- **Second, optimization is moving to the system level.** Engineers are no longer simply looking to add more compute. Memory bandwidth/capacity and storage are increasingly important bottlenecks, so the focus is shifting toward getting more out of the entire system rather than simply adding accelerators. This makes sense to us because there is little value in adding more compute if the rest of the system cannot keep the accelerators sufficiently utilized. The "memory wall" has been among the most common themes addressed with solutions ranging from increasing scale-up network throughput / reducing latency and novel compute architectures.
- **Third, networking is becoming much more important for inference.** As inference workloads become more distributed and latency-sensitive, moving data efficiently between accelerators and memory is becoming nearly as important as the compute itself. Better networking drives higher accelerator utilization and ultimately lowers cost/token. Against this backdrop, we are increasingly hearing that hyperscalers want multiple alternatives and protocols, making the scale-up market look less and less like a winner-take-all opportunity. Our conversations at the event suggested that ESUN has addressed many of Ethernet's historical latency issues through a shorter header and is now viewed as a credible scale-up alternative, while retaining the ecosystem and reliability benefits of Ethernet. At the same time, NVLink is viewed as the most mature/reliable non-Ethernet based option and UALink as an up-and-coming alternative. Importantly, we walked away thinking essentially everyone is open to experimenting with multiple protocols. In a UALink dedicated session, we heard from a Google engineer on the panel that although GCP continues to leverage OCS solutions broadly, it is also evaluating UALink as an option for particular latency-sensitive inference workloads, although our sense is these would initially be smaller deployments for very specific applications.
- **Fourth, CXL is emerging as an interesting way to extend the useful life of DDR4.** In multiple sessions at the event, hyperscalers were discussing CXL as a way to add DDR4 memory to the latest CPUs designed for DDR5. In one of the sessions, META highlighted that 43% of servers are memory-bound today and 69% of the



This report has been prepared by UBS Securities LLC. **ANALYST CERTIFICATION AND REQUIRED DISCLO**
INFORMATION ON THE QUANTITATIVE RESEARCH REVIEW PUBLISHED BY UBS, BEGIN ON PAGE 4.

carbon footprint is related to memory DIMMs. Importantly, a typical DIMM lifecycle is at least 2x longer than the CPU lifecycle. The obvious benefit is potentially reducing the amount of more expensive DDR5 required in the system. While CXL does introduce incremental latency versus local DIMMs, we heard that a smaller allocation, such as ~1/4 of system memory coming from CXL-connected DDR4, could still deliver comparable performance, effectively hiding the latency. We think this is another good example of the broader theme of operators increasingly optimizing AI infrastructure around total system economics rather than any single component.

- **Fifth, AI is starting to compress chip development and deployment cycles.** New AI tools are allowing engineers to parallelize parts of chip design that were traditionally sequential, which should shorten development timelines. But faster design cycles are also putting more pressure on how quickly new generations can actually be deployed. AWS described a new approach where it deploys and tests the racks before the compute silicon is available, then drops in the compute chip as essentially the last step. AWS said this can compress the ramp from the traditional ~9 months to several weeks. To us, this is an important second-order effect of faster AI chip roadmaps: it is not enough to design the next chip faster; the entire infrastructure stack increasingly needs to be ready for a much faster product cadence.

AI Infra Summit 2026 Overview

AI Infra is a 3-day annual summit in mid-September in the Bay Area for enterprise practitioners, data scientists, and ML engineers - attracting ~8,000 experts in the field, which we note more than doubled compared to last year's scale. This year's event featured a number of keynotes from key Compute, Networking, Memory and Storage players, with INTC's CEO and executives from AMD, MRVL, AVGO, Oracle, and AWS on stage. Beyond the keynotes, the event had 5 dedicated tracks: Compute, featuring companies such as d-Matrix, Cerebras, Etched, SambaNova, Rebellions, etc; Data Movement, dedicated to system-level data management and featuring Networking; Memory and Storage, with representatives from Samsung, SK Hynix, Astera Labs, etc.; Data & Models, covering model architectures and implications of agentic AI workloads; AI Data Center, focused on AI factory design and power; and Physical AI, addressing edge applications.

Valuation Method and Risk Statement

We use various valuation techniques such as P/E, EV/FCF for valuing the companies in this report. Risk factors include but are not limited to macroeconomic factors such as a downturn in the economy, a disruption of international trade, technological disruption due to new inventions, or business model innovation whereby structural changes in the industry alter the future course of unit sales, ASPs, and revenues.