26 Aug 2026 01:36:43 ET | 11 pages

US Semiconductors

Hot Chips Day 2: Networking is the New Compute

Atif Malik[AC]
James Bowlin
Papa Sylla
Kelsey Chia, CFA

CITI'S TAKE

If Hot Chips Day 1 was all about compute, Day 2 was fundamentally about connectivity. For the first wave of AI, the primary constraint was compute. The industry focused on building faster GPUs, larger accelerators, and more training capacity. Now as AI systems scale to trillions of parameters, massive context windows, and autonomous agent workloads, the limiting factor increasingly becomes the ability to move data efficiently across the system or networking. **This creates a particularly constructive outlook for Nvidia networking, Broadcom, Arista, co-packaged optics vendors Lumentum/Coherent, and CXL ecosystem companies like Marvell/Astera Labs, in our view.** We summarize key Day 2 themes below:

1. Networking Becomes the Critical Layer of AI

Nearly every major presentation pointed toward the same conclusion: future AI performance will be determined less by individual chips and more by how effectively those chips communicate. Broadcom's Thor Ultra, Nvidia's BlueField-4 DPU, Nvidia Spectrum-X, Nvidia Photonics, Meta's MTIA roadmap, Google's TPU architecture, Cerebras, SambaNova, Samsung's memory innovations, and XCENA's CXL-based processing all address different parts of the same challenge. Data movement is becoming the dominant bottleneck. Historically, compute was scarce and communication was relatively easy. Today, AI clusters contain thousands of accelerators generating enormous amounts of data that must be shared across servers, racks, and datacenters. As a result, network efficiency increasingly determines overall system utilization and performance. The implication is profound: The future AI winner is not necessarily the company with the fastest processor, but the company that can move data most efficiently throughout the entire system.

2. Memory Is Really a Networking Story

One of the most important insights from Day #2 is that the industry's "memory problem" is actually a specialized networking problem. Presentations repeatedly highlighted memory capacity, bandwidth, power consumption, and data transfer costs. Samsung's LPDDR5X-PIM and XCENA's CXL-PNM architecture both attempt to reduce the amount of data that must travel across the system. Cerebras emphasizes extreme on-chip memory bandwidth. The common objective is not simply adding memory; it is reducing communication overhead. Every byte that remains local is a byte that does not need to traverse increasingly congested

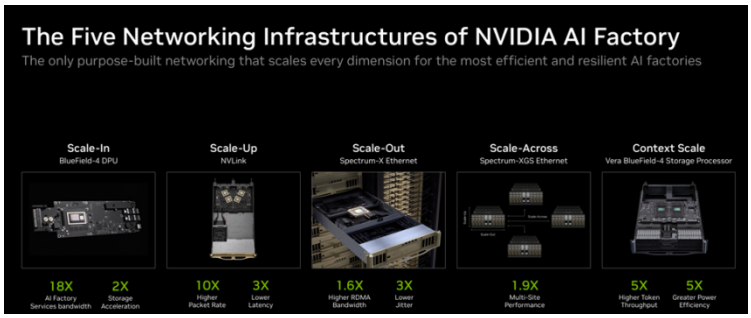
See Appendix A-1 for Analyst Certification, Important Disclosures and Research Analyst Affiliations

interconnects. Viewed through this lens, memory innovation becomes an extension of networking innovation.

3. AI Factories Are Network-Centric Architectures

Nvidia repeatedly framed future datacenters as AI factories rather than traditional cloud infrastructure. This is an important conceptual shift. Traditional datacenters were built around servers. AI factories are built around data flows. To function efficiently, AI factories require: Compute, Memory, Networking, Storage and Security to operate as a single integrated system. The strategic implication is that value creation is shifting away from standalone chips and toward providers that can optimize the entire data movement stack.

Figure 1. NVIDIA AI Factory



Source: NVIDIA

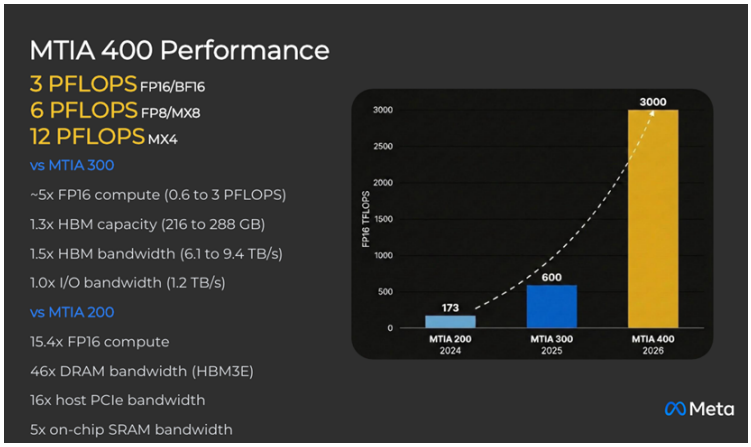
4. Inference Is Driving the Need for Better Networks

Another important Day #2 theme was the rise of AI inference, particularly agentic AI workloads. Multiple companies emphasized that future growth will come from running models rather than training them. Nvidia's LPX focuses on low-latency deployment and long-context workloads. **Groq 3 LPX rack produces 11,000 tokens per second, a new bar for inference AI. LPX delivers 4x faster long-context decode in 3rd party benchmarking.** SambaNova highlighted that agentic inference is increasingly decode-dominated. The industry conversation has shifted from: "How fast can we train a model?" to "How efficiently can we serve billions of AI interactions?".

5. Custom Silicon Is Increasingly a Network Optimization Strategy

Meta's MTIA roadmap and Google's TPU v8 demonstrate the accelerating trend toward custom AI silicon. **While this is often framed as a challenge to Nvidia, the deeper story is that hyperscalers are optimizing their entire infrastructure stack around their own workloads.** The goal is not simply cheaper compute. The goal is reducing system-level bottlenecks. In over a decade, Google has increased TPU shared-memory system performance by 1,000,000X. TPU 8t superpod will have 9,600 chips in aggregate and 2PB of shared memory.

Figure 2. Meta's Custom AI Silicon



Source: Meta

Companies Mentioned:
Alphabet Inc (GOOGL.O; US$346.96; 1; 25 Aug 26; 16:00) | Arista Networks (ANET.N; US$190.94; 1; 25 Aug 26; 16:00) | Astera Labs, Inc (ALAB.O; US$282.65; 1; 25 Aug 26; 16:00) | Broadcom Inc (AVGO.O; US$356.74; 1; 25 Aug 26; 16:00) | Cerebras Systems Inc (CBRS.O; US$183.92; 1; 25 Aug 26; 16:00) | Coherent Corp (COHR.K; US$288.14; 1; 25 Aug 26; 16:00) | Lumentum Holdings Inc (LITE.O; US$885.57; 1; 25 Aug 26; 16:00) | Marvell Technology Inc (MRVL.O; US$240.38; 1; 25 Aug 26; 16:00) | Meta Platforms Inc (META.O; US$570.05; 1; 25 Aug 26; 16:00) | NVIDIA Corp (NVDA.O; US$213.05; 1; 25 Aug 26; 16:00) | Samsung Electronics (005930.KS; W257000.0; 1; 25 Aug 26; 15:45)