

AI INFRA SUMMIT

September 9-11, 2025 | Santa Clara



DATA CENTER INSIGHTS REPORT

Read a snapshot of some of the key takeaways from the 2024 AI Hardware & Edge AI Summit (now the AI Infra Summit) featuring leading industry speakers as they discussed the infrastructure surrounding data centers.



Networking the AI Future: Ethernet's Evolving Role in Scaling AI Infrastructure



Hasan Siraj
Head of Software and AI,
Infrastructure Products
Broadcom

Key Takeaways

1. Broadcom's merchant silicon is powering large hyperscalers and enterprises, providing high-performance and open networking solutions.
2. Networking is a critical component for scaling AI infrastructure, with requirements for high bandwidth, efficient load balancing, low latency, and resilience.
3. Broadcom's switch and endpoint scheduling fabrics address these networking challenges, reducing job completion time by up to 10%.
4. Tomahawk 5 is Broadcom's latest 51.2 Tbps network silicon, with a roadmap to 100 Tbps and 200 Tbps, while improving energy efficiency by 90%.
5. Thor 2 is Broadcom's 400 Gbps AI-optimized NIC, designed for high RDMA performance and low power consumption.
6. Ethernet is the de facto standard for large AI cluster deployments, offering performance, scale, and widespread adoption.
7. The Ultra Ethernet Consortium is modernizing RDMA to support multipathing, out-of-order placement, and selective retransmit for scaling AI infrastructure.
8. Broadcom's Ethernet ecosystem includes silicon vendors, hardware vendors, software vendors, and operational expertise.
9. Optical I/O can provide dramatic improvements in profitability and interactivity for large AI models, both today and in the future.
10. Broadcom is investing in co-packaged optics to address thermal challenges as AI infrastructure scales to 100 Tbps and beyond.

Photonic Fabric™ based Scale-Up Network for Chip- to-Chip & Chip-to- Memory Connectivity



Preet Virk
Co-Founder & COO
Celestial AI

Key Takeaways

1. The key challenges in AI scaling are memory bandwidth, interconnect bandwidth, and memory capacity.
2. Celestial AI's photonic fabric solution can provide 86.4 TB/s or 390 TB/s of interconnect bandwidth, far exceeding the industry's needs.
3. The photonic fabric has a thermally stable optical modulator that can be integrated directly under the ASIC, eliminating the need for DSP and enabling low latency, low power, and low-cost interconnects.
4. Celestial's ASIC integrates PCIe, CXL, DDR, and HBM controllers, enabling a 33 TB memory appliance with 115 Tb/s of switching capacity.
5. This memory appliance can support a 10 TB recommendation engine using only 16 XPUs, compared to 56 XPUs in a traditional setup, improving compute density and efficiency.
6. For large language models like GPT-IV, the photonic fabric-based system can allocate more memory for context and batch size, improving inference quality and monetization.
7. Celestial's technology can enable a 128 XPU cluster with 792 TB of contiguous memory, addressing the growing memory capacity needs.
8. The photonic fabric's performance, power, and cost advantages can help democratize AI by lowering the barriers to deployment.
9. The technology can simplify the AI software stack by eliminating the need for model sharding and improving the efficiency of operations like broadcast and reduce.
10. The improved compute efficiency and density can lead to a lower carbon footprint for AI infrastructure.

Improving AI Infrastructure Performance and TCO With Optical I/O-Based Scale-Up Fabrics



 Ayar Labs

Mark Wade
CEO and Co-Founder
Ayar Labs

Key Takeaways

1. Ayar Labs is focused on improving the economics and performance of AI infrastructure, particularly around interconnect fabrics.
2. Massive investments (over \$1 trillion) are flowing into AI hardware and infrastructure, but the return on investment is being questioned.
3. The key challenges are improving the unit economics for AI application builders and accelerating the discovery of new killer AI applications.
4. Parallelizing AI models across multiple GPUs/accelerators can increase memory bandwidth and capacity, but the interconnect fabric becomes a bottleneck.
5. Ayar Labs has developed metrics like throughput, interactivity, and profitability to compare end-to-end AI system architectures.
6. Optical I/O in the scale-up fabric can significantly improve profitability (6x) and extend the range of interactivity for large language models like GPT-4.
7. As models grow larger (e.g., 14 trillion parameters), the benefits of optical I/O become even more pronounced.
8. Ayar Labs is shipping optical I/O chiplets today but scaling manufacturing to meet market demand remains a challenge.

Jabil Acquires Mikros Technologies to Support Demand for Liquid Cooling in AI Data Centers

2024 EVENT PARTNERS



Mikros Technologies' thermal management capabilities complement Jabil's existing data center solutions to enable the growth of AI infrastructure

ST. PETERSBURG, Fla. – October 3, 2024 – Jabil Inc. (NYSE: JBL), a global leader in design, manufacturing, and supply chain solutions, today announced the successful acquisition of Mikros Technologies LLC, a leader in the engineering and manufacturing of liquid cooling solutions for thermal management, completed October 1.

As the continued adoption of artificial intelligence, energy storage, and electric vehicles drives higher-power density systems across consumer and commercial industries, liquid cooling has emerged as a more energy-efficient alternative to air cooling for many applications. This acquisition will provide Jabil with critical capabilities for helping our customers manage the intense thermal requirements of their current and next-generation products while keeping sustainability and cost considerations top of mind.

"We are thrilled to welcome Mikros Technologies to the Jabil team," said Ed Bailey, Senior Vice President and Chief Technology Officer. "The thermal management capabilities they bring will allow Jabil to extend the range of services we provide to cloud service providers, hardware OEMs, and liquid cooling solutions



providers. In addition to the data center ecosystem, we see significant opportunities in other end-markets that require thermal management, including automated test equipment for semiconductors, batteries, energy storage systems, and electric vehicles."

For more than 30 years, Mikros Technologies has developed innovative microchannel liquid cooling solutions to address complex thermal management challenges. Their proprietary microchannel cold plate designs enhance energy efficiency by effectively cooling over one kilowatt per square centimeter. Mikros's exclusive technologies and capabilities will complement Jabil's extensive portfolio of data center lifecycle solutions, semiconductor test equipment solutions, and energy and transportation solutions.

"Joining Jabil is a tremendous path forward for Mikros Technologies. We can now scale our best-in-class liquid cooling capabilities to meet the thermal demands of the world's leading brands alongside a high-performing team of people we enjoy working with," said Drew Matter, CEO of Mikros Technologies.