Semiconductors

Hot Chips 2026 - Agentic AI Inference Reshapes Chip Architecture Across the Industry; Custom Silicon Proliferation Continues; Networking Taxonomy Expands with "Scale-In" Tier

Semiconductors & Semiconductor Capital Equipment / IT Hardware

Harlan Sur AC
(1-415) 315-6700
harlan.sur@jpmorgan.com

Mayur Ramdhani
(1-212) 622-1664
mayur.ramdhani@jpmorgan.com

Apoorva Kumar
(1-212) 270-0668
apoorva.kumar@jpmorgan.com
J.P. Morgan Securities LLC

We recently attended the Hot Chips 2026 conference and came away with an even more constructive view on the multi-year AI infrastructure buildout, with agentic AI-driven inference workloads now clearly reshaping chip architecture across the industry. Every session we attended opened with agentic AI as the design imperative, a marked shift from Hot Chips 2025, where training-optimized design was still the primary focus. Several key themes were prevalent at the conference: **(1)** agentic AI inference is fundamentally reshaping silicon design, with decode-dominated workloads, KV cache management, long context handling, and low-latency multi-chip execution now the first-order design constraints (highlighted across numerous presentations); **(2)** custom silicon proliferation continues (headlined at the event by OpenAI's Jalapeño ASIC) as performance and economics rapidly improve; **(3)** AMD reinforced the credibility of its 2H26/2027 competitive answer to Nvidia's Vera Rubin with expansive detail on its Instinct MI400 series and Helios rack solution; **(4)** the networking taxonomy continues to expand, with Nvidia introducing a "scale-in" tier via BlueField-4 and Broadcom detailing Thor Ultra as its Enhanced RoCE-based scale-out AI Ethernet NIC; and **(5)** HBM remains a gating supply constraint for AI silicon - every major roadmap presented at the conference showed dependency on HBM3E/4/4E, reinforcing expectations of prolonged memory tightness. One cross-cutting theme that we think was somewhat underappreciated but worth flagging was Broadcom's involvement across many of the most consequential products/technologies presented at the event — e.g., Google's TPU program, OpenAI's Jalapeño ASIC, Meta's MTIA ASIC program, SambaNova's RDU (ASIC) program, and the proliferation of ethernet for scale-up and scale-out (including Broadcom's own Thor Ultra 800G Ethernet NIC highlighted at the conference), underscoring what we view as AVGO's expanding dual role across both AI compute silicon (via custom AI ASICs) and AI infrastructure networking. Taken together, we continue to see AI as a durable multi-year driver of semiconductor demand, and we continue to view AVGO, NVDA, AMD, MRVL, MTSI, ALAB, MU, and SNDK as key beneficiaries of the ongoing AI infrastructure buildout.

- **Agentic AI-driven inference is now a major design imperative across the industry, with decode-dominated workloads and KV cache management fundamentally reshaping silicon architecture.** Agentic AI inference was a clear top-of-mind design imperative at this year's event. Every presentation we attended emphasized agentic workloads as a primary design driver, marking a shift from last year's event where training-optimized architectures still dominated. We noted a collective recognition that inference (particularly agentic multi-turn workloads with growing context lengths and mixture-of-experts architectures) is fundamentally memory-bandwidth-bound and decode-dominated rather than compute-FLOPS-bound and prefill-dominated, and we think this has meaningful implications for which architectures win in the next chip cycle. SambaNova's SN50 RDU presentation quantified this shift, showing that 75-97% of compute cycles across frontier models (DeepSeek, MiniMax, Kimi, Qwen, GLM) are now decode-dominated, and argued that memory bandwidth utilization (not peak FLOPS) is the binding

See page 5 for analyst certification and important disclosures.

constraint on inference throughput. Google's TPU v8 presentation (which featured training-optimized v8t and inference-optimized v8i) highlighted materially larger on-chip SRAM allocation for v8i specifically to address MoE inference latency. Intel's Crescent Island GPU was designed from the ground up for agentic AI inference, and even its memory choice (LPDDR5x instead of the traditional HBM) reflects the industry's evolving recognition that inference is more sensitive to capacity than raw bandwidth. OpenAI's Jalapeño similarly framed its performance metrics in terms of latency (TTLT, TBT), energy per request (tokens/joule), and Pareto frontier positioning rather than peak FLOPS. Taken together, we think this collective architectural pivot suggests the chip cycle will be increasingly bifurcated between training-optimized silicon (peak FLOPS + large scale-up interconnect bandwidth) and inference-optimized silicon (memory capacity + on-chip SRAM + low-latency multi-chip fabrics), and we see this as a broadly constructive dynamic that should benefit a broader ecosystem than the training-heavy prior cycle.

- **Custom silicon proliferation continues, headlined by OpenAI's Jalapeño ASIC alongside others signaling an ongoing commitment to internal silicon development.** OpenAI's presentation on its Jalapeño ASIC (OAI's first-ever internal chip designed in partnership with Broadcom and specifically optimized for inference workloads) emphasized both the speed of the development cycle (~9 months from initial RTL to tapeout - enabled by AI-accelerated hardware/software co-design) and OpenAI's ongoing commitment to internal silicon development, and we would expect Jalapeño-successor programs and adjacent silicon projects to follow in a fast-cadence pattern. Design metrics were presented in terms of Pareto frontier positioning (latency, energy per request, tokens/kW) rather than peak FLOPS, reflecting the inference-optimized nature of the design, and Jalapeño was framed as a substantial improvement vs. the leading merchant GPU comparison points on the workloads OpenAI cares most about. Joining Jalapeño in the custom silicon expansion narrative were Google's TPU family (where AVGO is the primary design partner on v8/9/10/11), Meta's MTIA program (AVGO also partnered on MTIA 300/400/450/500), Microsoft's MAIA (MRVL is MSFT's key ASIC partner), SambaNova's dataflow-architecture SN50 RDU, and Intel's Crescent Island GPU. It is also worth noting, though, that OpenAI has publicly committed to substantial NVDA infrastructure deployment through 2030 (~12 GW cumulative), hyperscalers continue to buy NVDA GPUs alongside in-house custom programs, and no frontier lab has signaled displacement of merchant GPU demand. In our view, custom silicon and merchant GPUs should continue to coexist as complementary rather than substitutive elements of the AI compute stack, with custom ASICs targeting specific inference workloads at scale where the customer already has the model, workload characteristics, and volume to justify NRE, while merchant GPUs (NVDA VR/VRU, AMD MI400/500) continue to serve the broader ecosystem's need for a fungible platform that runs every model. That said, we maintain the view that share of the overall AI accelerator TAM between merchant GPU (which currently commands >85% share, per JPMe) and custom XPU/ASIC will trend towards parity over the next several years as AI labs (including Anthropic), and hyperscalers rapidly scale custom XPU infrastructure.
- **AMD provided expansive architectural detail on the Instinct MI400 series and Helios rack solution, reinforcing the credibility of its competitive answer to NVDA's Vera Rubin, and positioning UALink as a viable open-standard scale-up alternative to NVLink.** AMD's MI400 series deep dive was the company's most detailed disclosure to date that, in our view, was helpful in framing/de-risking AMD's competitive positioning relative to Nvidia. On rack-scale, AMD's Helios reference architecture delivers 2.9 exaflops of AI compute and 31 TB HBM4 memory across 72 GPUs per rack (matching NVDA VR's 72 GPU/rack configuration), positioning AMD as a compelling full-system alternative (not just a GPU alternative) for hyperscaler and neocloud customers. The MI455X GPU itself features an enhanced modular chiplet architecture (compute chiplets on N2 process, I/O and cache dies on N3P) with materially higher

人在草木间

HBM4 capacity and bandwidth relative to prior generations, giving AMD a competitive memory profile for MoE inference and long context agentic workloads where memory has emerged as the binding constraint. Importantly, AMD is using UALink 128 as its scale-up fabric, providing an open-standard alternative to NVLink and positioning UALink as a rallying point for the broader ecosystem (hyperscalers that prefer open standards, alternative accelerator vendors seeking scale-up interoperability, etc.) beyond just AMD. The Helios rack also integrates AMD's Pensando Vulcano 800 AI NICs (up to 3 per GPU for scale-out) and Pensando Salina 400 DPU for front-end 800GbE networking, giving AMD a vertically-integrated stack across GPU + CPU + DPU + AI NIC that mirrors Nvidia's full-stack platform approach. Layered on the accelerating inference workload dynamic we discussed earlier (where AMD has demonstrated strong competitive positioning given attractive TCO and ROCm's frictionless deployment of NVDA-trained models), we think Hot Chips reinforced our thesis that AMD is well-positioned to capture market share in AI compute (off a low base) over the coming years, particularly in inference where demand growth is concentrated.

- **Networking taxonomy expands as NVDA introduces "scale-in" tier via BlueField-4, while Broadcom's Thor Ultra 800G positions Ethernet as the AI scale-out and scale-across standard.** Two announcements stood out to us at this year's event. First, NVDA's BlueField-4 DPU presentation introduced a new networking category - "Scale-In Network" - sitting between the traditional scale-up (NVLink) and scale-out (Spectrum-X Ethernet) tiers, specifically designed to connect GPU workloads to the surrounding data center infrastructure (storage, CPUs running agents, orchestration, security). We see this as a conceptual expansion of the AI networking framework and one that, in our view, further reinforces NVDA's positioning as a full-stack networking platform provider (spanning NVLink for scale-up, Spectrum-X for scale-out, Spectrum-XGS for scale-across, and now BlueField-4 for scale-in), well beyond its origins as a GPU vendor. Second, Broadcom's Thor Ultra presentation detailed the company's 800G Ethernet NIC purpose-built for AI scale-out and scale-across fabrics, with an Enhanced RoCE (eRoCE) feature set that specifically addresses the reliability, congestion control, and multipathing challenges of AI training and inference workloads on Ethernet. Thor Ultra positions Ethernet as a compelling alternative to InfiniBand for AI scale-out fabrics and reinforces the ongoing shift of AI networking share toward Ethernet-based solutions (a dynamic that benefits both AVGO and MRVL directly). The broader networking discussion at Hot Chips 2026 fundamentally reinforces a theme that has emerged over the past 12 months: AI networking is a rapidly expanding TAM (both scale-up and scale-out) that is competitively contested but where the leading incumbents (AVGO, NVDA) are pulling ahead with vertically-integrated solutions. We continue to see this as an underappreciated aspect of the AI infrastructure buildout that will benefit not only merchant NIC/DPU incumbents (AVGO, NVDA), but also the broader networking silicon ecosystem, including MRVL (custom scale-out silicon, DCI/DSPs, and coherent optics), ALAB (connectivity fabric and PCIe/CXL retimers), and MTSI (high-speed lasers and optical components), as cluster sizes and networking bandwidth requirements continue to expand.
- **Intel used Hot Chips 2026 to signal renewed hyperscaler CPU + AI ambitions via Diamond Rapids (next-gen Xeon) and Crescent Island (agentic inference GPU), indicating architectural progress even as roadmap execution risk remains elevated.** While Intel remains under significant execution pressure, its two Hot Chips 2026 presentations we think demonstrated meaningful architectural progress on both CPU and GPU roadmaps and reinforced the company's commitment to remaining a competitive force in AI infrastructure silicon. Diamond Rapids (DMR), Intel's next-gen Xeon CPU, features a 3D Mesh architecture representing a step forward from the Modular Mesh in Sierra Forest / Granite Rapids, with 2.5D + 3D disaggregation and function-based partitioning that we think is well-suited for hyperscaler-optimized DC CPU workloads. Framing of DMR was notably around "how the processor moves, processes, and protects

data at hyperscale," signaling Intel's DC CPU strategy has evolved from a pure compute performance narrative to one centered on AI factory data movement, security, and orchestration, which we view as an appropriate strategic pivot given how hyperscaler CPU workloads have evolved with the AI buildout. On the GPU side, Crescent Island is purpose-built for agentic AI inference specifically, and reflects Intel's recognition that the meaningful GPU opportunity for the company is on the inference side (not competing head-on with NVDA/AMD on frontier-model training). Its 480GB LPDDR5x memory (vs. traditional HBM) reflects the industry's increasing recognition that inference is more sensitive to memory capacity than raw bandwidth, its 350W TDP positions it as an energy-efficiency-focused offering, and its on-chip MoE speculative decoding acceleration is workload-relevant. We view DMR and Crescent Island as demonstrative of Intel's architectural progress, with ultimate competitive impact dependent on execution and customer commit signals into 2027 (which at least for DMR we expect will be strong given tight server CPU supply-demand dynamics).