

Otto Analytics

FRONTIER INFRASTRUCTURE · CONVICTION RESEARCH

WHAT I READ THIS WEEK

Who Owns the Rack When Everyone Builds a Chip?

Custom accelerators are multiplying. The scarce layer may be the architecture that turns them into usable systems.

OTTO ANALYTICS · THE WEEKLY READ · NO. 11 · SEPTEMBER 3, 2026

Programming note: We were on vacation last week, so today's issue catches up on the news from August 24 through August 31.

Bottom line up front

This was the week custom AI silicon stopped looking like a distant threat and started looking like a production market. OpenAI published the first measured results for its Jalapeño inference chip. AMD and Cisco put an alternative AI system into service in Saudi Arabia. Marvell reported accelerating demand in both custom silicon and connectivity.

NVIDIA's response was more interesting than defending every accelerator socket. It expanded the route for other companies' chips to enter NVIDIA-connected racks. AWS plans to combine its Trainium processors with NVIDIA's interconnect and memory technology. MediaTek will help customers bring custom accelerators into NVIDIA's rack-scale architecture.

The usual debate asks which chip wins. The better question is who owns the system when several chips do.

An accelerator becomes useful capacity only after it is connected to memory, networking, software, power and cooling, then qualified to run real workloads reliably. Chip choice may be diversifying while control of the AI-factory architecture becomes more concentrated.

That is the tension worth underwriting.

The week the custom chip became a system

OpenAI published the first operating results for Jalapeño, its custom inference accelerator, on August 25. The company reported 1.5 to 1.9 times more AI work per watt at peak throughput and 1.7 to 3.6 times lower end-to-end latency (how long the system takes to return an answer) across three public models than its comparison systems. It plans to begin deploying the chip inside its own infrastructure by year-end.

The figures are company-reported. Production yield, total deployment volume, availability, cost and performance across a broader workload mix remain unproved. But this is stronger evidence than a roadmap. OpenAI has working first-party silicon, measured results and a multigenerational plan.

The most important part of the disclosure was not one benchmark number. OpenAI said the advantage came from designing the chip, memory, network, serving software and rack-scale system together around its own workloads.

That is what custom silicon actually means at scale. The customer is not merely replacing one processor. It is trying to optimize the path from model to machine.

The chip is where differentiation begins. The system is where the economics have to survive.

An alternative stack reaches production

AMD, Cisco and HUMAIN supplied a second proof point on August 31. They said an AI system built from AMD Instinct MI355X GPUs, AMD EPYC CPUs, Cisco Silicon One networking and Cisco 800G optics is live in Saudi Arabia and serving customers.

The companies did not disclose the system's current capacity, utilization, customer mix or revenue. Their next phase calls for as much as 250 megawatts beginning in 2027, with a longer-term target of up to one gigawatt by 2030. Those are plans, not operating capacity.

Still, the distinction matters. A non-NVIDIA architecture moved from announced intent to production service. The customer now has an environment in which performance, reliability and software maturity can be tested under real workloads.

Marvell's fiscal second-quarter results showed where the surrounding demand is appearing. Data-center revenue reached \$2.17 billion, up 46 percent from a year earlier and 18 percent sequentially. It represented 79 percent of company revenue. Management said connectivity remained strong and its custom business should accelerate significantly in the second half of the fiscal year.

Custom compute and connectivity are not opposing stories. More accelerator designs can increase the need for high-speed links, switching, memory movement and system integration. The arithmetic becomes more specialized while the cost of making thousands of processors work together becomes more important.

NVIDIA opens the gate without surrendering the road

NVIDIA's strategy became clearer from two disclosures on either side of the week.

AWS and NVIDIA plan to deploy two million additional NVIDIA GPUs across AWS infrastructure during 2027 and 2028. At the same time, Amazon's Annapurna Labs is expanding its use of NVIDIA NVLink Fusion and NVIDIA's custom high-bandwidth memory technology for future Trainium processors. AWS can continue designing differentiated compute while placing those chips inside a common rack-scale architecture with NVIDIA systems.

Then NVIDIA and MediaTek announced that MediaTek will use NVLink Fusion as a design foundation for custom accelerators. NVIDIA also invested \$3.5 billion in MediaTek convertible bonds.

The proposed division of labor is revealing. Customers can focus on the compute that differentiates their workloads. MediaTek and NVIDIA offer the interconnect, memory architecture, packaging, manufacturing support and rack-scale qualification needed to turn that compute into a deployable system.

This is not proof that the market will standardize on NVIDIA. It is evidence of how NVIDIA intends to compete as accelerator choice broadens.

The company does not have to design every chip if it can define the environment those chips must enter.

A chip is not a system

The path from working silicon to productive capacity has six gates.

1. Workload fit

The accelerator must improve something the customer values, such as latency, throughput, cost per task or useful work per watt. A benchmark advantage that disappears outside one model or batch size is not a platform.

2. Memory movement

Modern inference can be limited by how quickly model weights and cached context reach the processor. Memory bandwidth, capacity, placement and software determine whether theoretical compute becomes useful output.

3. Scale-up fabric

Processors inside a rack or pod must exchange data quickly enough to behave like a larger machine. Latency, bandwidth and collective operations determine how much performance is lost while chips wait on one another.

4. Scale-out network

Racks must connect across the data center without congestion overwhelming the workload. Switches, optics, network software and topology become part of system performance.

5. The physical rack

Power delivery, liquid cooling, packaging, cabling and serviceability determine how densely the system can operate and how often it remains available. A processor that benchmarks well but cannot be deployed reliably at rack scale is not usable capacity.

6. Software and qualification

Compilers, kernels, orchestration, security, monitoring and model support determine how quickly a customer can move workloads into production. The final product is not a chip that turns on. It is a system that keeps running.

A benchmarked accelerator is capability. A qualified rack serving real work is capacity.

The standard war inside the machine

The strategic contest now has two layers.

Scale-up technology connects processors closely enough to act as one system inside a rack or pod. Scale-out networking connects those systems across a larger cluster. Both matter, but they solve different distances.

NVIDIA's advantage is coordination. NVLink, memory, GPUs, CPUs, networking, software and reference rack designs can be qualified together. A customer may accept more dependence on one vendor if that path reduces deployment time and integration risk.

The counter-case is an open standard.

The UALink Consortium includes companies such as AMD, AWS, Cisco, Google, Meta and Microsoft. Its specifications are intended to create an open scale-up fabric for accelerators from multiple suppliers. The consortium's 2026 update added in-network compute, chiplet definitions and standardized manageability.

That is credible technical organization, not proof of commercial adoption. A published specification still needs interoperable products, volume deployment, software support and operating evidence. Open standards win only when customers can procure systems that work.

The same test applies to NVIDIA. A prevalidated architecture is valuable only if it produces better deployment economics than an open alternative after hardware cost, power, utilization and switching risk.

The contest is not proprietary versus open as a matter of ideology. It is time to useful capacity, performance per watt, reliability and freedom to change suppliers.

Where the return migrates

If accelerator choice broadens, value does not disappear. It moves toward the layers whose importance rises with system complexity.

That can include high-bandwidth memory (the fast memory placed beside an accelerator), SerDes (circuits that move data on and off chips) and scale-up links (connections that let many processors work as one). It also includes Ethernet switching, optics, advanced packaging (the way processors and memory are assembled into one module), rack integration, cooling, power delivery and the software that schedules work across different processors.

Where demand can show up inside Otto's current coverage universe:

- **Memory and packaging:** MU, SSNLF and TSM.
- **Fabric, switching, optics and timing:** AVGO, MRVL, CRDO, ANET and SITM.
- **Rack integration, cooling and power:** CLS, VRT, MOD, MPWR and ABB.
- **Cross-stack architecture and software:** NVDA and AMD.

NVIDIA could lose part of an accelerator socket while preserving economics through interconnect, memory, networking and software. An open ecosystem could instead divide those economics among custom-silicon designers, networking suppliers and system builders. Large cloud companies could internalize more of the stack and reduce the margin available to both.

None of those outcomes follows automatically from more chips being announced.

The durable supplier is the one whose layer remains necessary across architectures and whose economics survive as customers gain choice.

The system-level underwriting test

We would evaluate the market through eight measures:

1. **Production deployment:** Is the system serving customer workloads, or is the evidence still a benchmark and roadmap?
2. **Useful work per kilowatt:** Does the architecture improve output at a matched latency and quality level?
3. **Performance per rack:** How much of the chip's theoretical performance survives memory, networking and thermal constraints?
4. **Time to service:** How long passes between tapeout, system qualification and billable capacity?

5. **Software portability:** What must a customer rewrite to move workloads or change accelerators?
6. **Fabric dependence:** Can custom chips reach production through more than one scale-up and scale-out architecture?
7. **Reliability and utilization:** Does the system remain available and busy enough to earn an acceptable return?
8. **Economic capture:** Which supplier turns technical control into revenue, margin and durable switching costs?

This is the next useful denominator for the AI infrastructure market.

Stop counting chips announced. Count qualified, connected and utilized systems.

What would strengthen, weaken or break the thesis

Strengthen: multiple custom accelerators enter production through the same rack architecture; interconnect, memory and networking revenue grows faster than accelerator units; qualification time falls; and customers adopt the architecture across several generations.

Weaken: custom chips remain limited to narrow internal workloads; system integration costs erase their silicon advantage; or architecture providers cannot preserve pricing as customers demand more interoperability.

Break: open fabrics reach broad production with comparable performance and materially lower switching costs, hyperscalers successfully internalize the complete system, or networking and rack components commoditize faster than heterogeneous compute expands their importance.

The chip market can diversify without making the surrounding architecture less valuable. It may do the opposite. More processor designs create more boundaries that memory, networking, software and physical systems must cross.

More companies can design a chip. Far fewer can turn one into a reliable factory.

The next architecture winner may not make every accelerator. It may define the system competing accelerators have to enter.

That is what we read last week.

Otto Analytics

Expect a fresh issue tomorrow covering this week's developments.

SOURCES AND ASSUMPTIONS

- [OpenAI, Jalapeño's first results, August 25, 2026](#)
- [OpenAI, The full stack behind abundant intelligence, August 25, 2026](#)
- [NVIDIA and AWS, two million additional GPUs and heterogeneous AI infrastructure, August 26, 2026](#)
- [Marvell, fiscal second-quarter 2027 results, August 27, 2026](#)
- [AMD, Cisco and HUMAIN, Saudi production deployment, August 31, 2026](#)
- [NVIDIA and MediaTek, expanded custom-silicon partnership, August 31, 2026](#)
- [UALink Consortium, UALink specifications and mission](#)
- [UALink Consortium, 2026 specification update](#)
- [Otto Analytics, current Coverage Universe](#)

Jalapeño performance figures are OpenAI-reported and compare selected systems and operating points using the InferenceX benchmark. The AMD, Cisco and HUMAIN operating status and expansion targets are company-reported; current capacity, utilization, revenue and customer mix were not disclosed. NVIDIA and partner deployment quantities, architecture benefits and timelines are company plans rather than completed outcomes. Marvell's non-GAAP measures are omitted from the argument. UALink is treated as an open technical standard and competitive counter-case, not as evidence of broad production deployment. The six-gate system framework and eight-part underwriting test are Otto Analytics frameworks.

OTTO ANALYTICS

Otto Analytics publishes this column for informational and educational purposes only. This commentary is impersonal analysis offered uniformly to all subscribers and the general public. It is not investment advice, not an offer or solicitation, and does not account for your individual circumstances. Any ratings or conviction scores reflect our editorial assessment and are not recommendations to buy, sell, or hold any security. The author or Company may hold positions in securities discussed in this column. All investing involves risk, including loss of principal.

© 2026 Otto Analytics, LLC. All rights reserved.