



UALink vs NVLink: Open Scale-Up Fabrics Explained

Plan and build private AI compute with GPU Smith. We connect workload requirements with GPU hardware, networking, deployment and operating decisions so your infrastructure fits the work it must perform.

Published by GPU Smith · 2026-10-29 · ualink vs nvlink · ualink 200g 1.0 specification · ualink 1024 accelerators · nvlink fusion vs ualink · scale-up vs scale-out ai fabric · ualink switch availability

Original article: <https://gpusmith.com/articles/en/ualink-vs-nvlink-scale-up-fabrics>



In this article

1. The short answer
 2. Scale-up vs scale-out in one page
 3. The UALink Consortium and its specifications
 4. What UALink 1.0 actually defines
 5. Reading the performance claims
 6. NVLink 6 and NVLink Fusion
 7. Where AMD stands: Infinity Fabric today, UALoE in Helios
 8. Side-by-side comparison (as of 2026-10-28)
 9. Availability reality check
 10. Procurement implications
 11. Decision guide
 12. Sources
-

The short answer

As of October 28, 2026, NVLink is a shipping scale-up fabric from a single vendor. UALink is an open specification, and no vendor has publicly confirmed shipping a native UALink switch. NVLink 6 connects 72 GPUs in a Vera Rubin NVL72 rack. UALink 1.0 defines 200G-per-lane links for pods of up to 1,024 accelerators ([UALink Consortium, 2025-04-08](#)). Vendors that have given dates for native UALink switches point to 2027. AMD Helios, the first "UALink" rack you can order, runs UALink over Ethernet (UALoE) on Broadcom Tomahawk 6 switches, not on native UALink silicon ([HPE, 2025-12-02](#)).

So for 2026 deliveries the real choice is NVLink racks or AMD's Ethernet-based UALoE racks. Native UALink is a decision for 2027 and later.

Scale-up vs scale-out in one page

A **scale-up** fabric joins accelerators into one memory domain. GPUs read, write and run atomics directly against each other's memory through purpose-built switches. UALink describes itself in these terms: "read, write, and atomic transactions" over "a simple load/store protocol" ([UALink Consortium, 2025-04-08](#)). Tensor-parallel and mixture-of-experts (MoE) all-to-all traffic runs on this layer. NVIDIA cites MoE all-to-all as the reason for the NVL72 topology ([NVIDIA NVLink](#)).

A **scale-out** fabric (InfiniBand, RoCE, Ultra Ethernet) connects those domains over RDMA across thousands of nodes. It is covered in our [scale-out fabric comparison](#). UALink stops at the pod boundary: "inter-Pod communication and control is not specified by the UALink specification" ([UALink 1.0 white paper, 2025-04](#)). Helios pairs UALoE for scale-up with Ultra Ethernet-aligned Ethernet for scale-out ([AMD, 2026-07-23](#)).

The UALink Consortium and its specifications

The consortium was incorporated in October 2024. Its board is Alibaba, AMD, Apple, Astera Labs, AWS, Cisco, Google, HPE, Intel, Meta, Microsoft and Synopsys; NVIDIA is not a member of the board ([UALink Consortium, 2026-04-07](#)). As of October 28, 2026, the specification page lists six documents, all available for download ([UALink Consortium specifications](#)):

Specification	Release d	What it adds
UALink 200G 1.0	2025-04-08	Base protocol, 200G/lane, up to 1,024 accelerators
UALink 128G Data Link and Physical Layers 1.0	2025	128G-per-lane option (the page's description reads "coming soon")
UALink Common Specification 2.0	2026-04-07	In-Network Compute
UALink 200G Data Link and Physical Layers 2.0	2026-04-07	Physical and data link layers split out, so new speeds don't require changes to the other specs
UALink Manageability 1.0	2026-04-07	Control and management planes using gNMI, YANG, SAI and Redfish
UALink Chiplet 1.0 (listed as 1.01)	2026-04-07	Chiplet integration, "fully compliant with the UCle 3.0 Specification"

The 2026 release dates and contents come from the 2.0 release ([UALink Consortium, 2026-04-07](#)). The chair's roadmap post describes 1.0 as "at 200G and 128G per lane" and schedules **UALink 3.0 for 2027**: more bandwidth, reach "across racks and rows" and better reliability ([UALink Consortium blog, 2026-03-06](#)). No "128G 2.0" document exists. Compliance and interoperability programs are still described as planned ([UALink Consortium, 2026-04-07](#)).

What UALink 1.0 actually defines

Source for all of the following: [UALink 1.0 white paper, 2025-04](#).

- **Lane rate:** up to 200 GT/s of data per lane, signaled at 212.5 GT/s to leave room for Ethernet FEC and encoding. A 106.25G option is also defined.
- **Links:** x1, x2 or x4. A four-lane "Station" carries at most 800 Gbps in each direction.
- **PHY:** based on IEEE 802.3 (200GBASE-KR1/CR1 through 800GBASE-KR4/CR4). The PMD is unchanged; lower-interleave FEC modes cut latency.
- **Flits:** 64-byte transaction flits are packed into 640-byte data-link flits. Each fits exactly one RS(544,514) codeword and carries a 32-bit CRC and link-level replay.
- **Scale:** switches "connect up to 1024 Accelerators" and can be partitioned into isolated Virtual Pods. UALinkSec adds encryption for confidential-computing tenants.

The 1,024 figure is an addressing ceiling. No announced system we found comes near it, and shipping racks stop at 72. Because UALink reuses the Ethernet PHY, it shares SerDes, cables and retimers with 200G Ethernet. Astera Labs, for example, sells retimers "for 224G Ethernet and UALink connectivity" ([Astera Labs Q2 2026 results, 2026-08-04](#)). Above the PHY, though, UALink is not Ethernet, and a UALink switch is not an Ethernet switch.

Reading the performance claims

The consortium says UALink offers "the same raw speed as Ethernet with the latency of PCIe switches" and is "designed for deterministic performance achieving 93% effective peak bandwidth" ([UALink Consortium, 2025-04-08](#)). Treat both as design targets:

- A consortium-hosted 2026 white paper defines 93% as protocol payload divided by physical-layer capacity. It says the targets were "validated through architectural modeling and protocol simulation" and that "empirical verification across multi-vendor hardware is planned for 2026" ([TASK Consultancy for UALink Consortium, 2026](#)). We found no published hardware measurement.
- The same paper gives "approximately 1 microsecond round-trip latency" over copper shorter than four meters. That is also a specification target.

Putting 93% next to an NVLink TB/s figure compares a paper target with a marketing figure. Neither is a measured collective bandwidth.

NVLink 6 and NVLink Fusion

NVLink is NVIDIA's proprietary GPU-to-GPU link, extended across the rack by NVLink Switch chips.

Generational detail is in [NVLink 5 vs NVLink 6](#). One point matters here: **the current NVLink 6 figure is 3 TB/s per GPU and 216 TB/s per NVL72, but one NVIDIA page still shows the higher launch figure, because NVIDIA revised it after launch.**

NVIDIA page	Per-GPU NVLink 6	72-GPU domain
NVLink ("Preliminary specifications; may be subject to change")	3,000 GB/s, 36 links	216 TB/s (NVL72)
Vera Rubin NVL72 spec table	3 TB/s	216 TB/s
NVLink Fusion (launch figure, not updated)	"3.6 TB/s per XPU"	"260 TB/s ... (NVL72)"

The current figure checks out: $72 \times 3 \text{ TB/s} = 216 \text{ TB/s}$. The Fusion page's launch figure measures the same thing, bidirectional bandwidth per GPU in a Vera Rubin NVL72, so the difference is a revision, not a definition. NVIDIA announced 3.6 TB/s and 260 TB/s at CES 2026 ([NVIDIA, 2026-01-05](#)), and the Vera Rubin NVL72 page still listed them in June 2026. The current product pages and the [Vera Rubin datasheet](#) (dated 2026-10-14) give 3 TB/s, 216 TB/s and "1.7x" over NVLink 5. The Fusion page and NVIDIA's developer blogs still carry the launch figure. Treat 3 TB/s as the current Rubin GPU specification, and get the number for your exact SKU in writing. For comparison, GB200 NVL72 is 1.8 TB/s per GPU and 130 TB/s per rack ([NVIDIA GB200 NVL72](#)).

- **Domain size:** NVLink 6 domains are listed as "8 | 72" ([NVIDIA NVLink](#)). The Fusion page cites "future roadmap domain sizes up to 1,152", a roadmap figure in the same way that UALink's 1,024 is a specification ceiling.
- **In-network compute:** every NVLink Switch has SHARP engines for reductions and multicast (same source). UALink added In-Network Compute on paper in April 2026.
- **Availability:** "Production shipments of Vera Rubin are set to begin starting this fall" ([NVIDIA, 2026-05-31](#)). The product page now reads "Available Now".

NVLink Fusion lets hyperscalers "deploy custom XPU and CPUs into NVIDIA's world-leading AI infrastructure platform" on MGX racks ([NVIDIA NVLink Fusion](#)). It opens the endpoint, not the fabric: NVIDIA still supplies the switch. Fusion-specific checks are in [NVLink Fusion due diligence](#).

Where AMD stands: Infinity Fabric today, UALoE in Helios

Until Helios, AMD's scale-up domain was an 8-GPU Infinity Fabric node; see [NVLink vs Infinity Fabric](#). AMD describes Helios this way: "Four scale-up cartridges leverage UALink over Ethernet (UALoE) to seamlessly connect up to 72 GPUs ... with up to an industry leading 260 TB/s aggregate scale-up bandwidth." Each switch tray holds "Two 512 lane 200G UALoE switch ASICs with 216 UALoE (x2) active links per ASIC", rated at "21.6 TB/s (Bi-Dir) bandwidth per switch" ([AMD Helios](#)). HPE builds the switching with Broadcom's Tomahawk 6, using six HPE Juniper Networking QFX5252 scale-up Ethernet switch trays per rack ([HPE, 2025-12-02](#); [HPE, 2026-09-30](#)).

HPE calls UALoE an "open ... standard", but no UALoE document appears among the consortium's published specifications ([UALink Consortium specifications](#)). Ask which document defines the UALoE you are buying.

Worked example: do the Helios numbers add up?

The calculation uses only AMD and HPE figures plus two assumptions of ours. First, "216 UALoE (x2) active links" means 216 two-lane links. Second, each QFX5252 tray is one two-ASIC switch tray. AMD's "four scale-up cartridges" is not defined. Four two-ASIC trays would give only 172.8 TB/s, so "cartridge" probably does not mean switch tray.

1. Active lanes per ASIC: $216 \times 2 = 432$ (out of 512).
2. One direction: $432 \times 200 \text{ Gb/s} = 86.4 \text{ Tb/s} = 10.8 \text{ TB/s}$.
3. Both directions: $2 \times 10.8 = 21.6 \text{ TB/s}$. This matches AMD's per-switch figure.
4. ASICs per rack: $6 \text{ trays} \times 2 = 12$.
5. Rack total: $12 \times 21.6 = 259.2 \text{ TB/s}$, rounded to "up to 260 TB/s".
6. Per GPU: $259.2 / 72 = 3.6 \text{ TB/s bidirectional}$, or 72 lanes of 200 Gb/s per GPU ($72 \times 72 = 5,184 = 12 \times 432$).

So 260 TB/s is the sum of raw lane rates. It matches NVIDIA's CES 2026 launch figure, which is still on the Fusion page, and is 20% above the 216 TB/s in NVIDIA's current Vera Rubin specification.

Native UALink at AMD. As of October 28, 2026, we found no AMD primary statement committing a GPU to native UALink switches. AMD says "Next-generation AMD Instinct MI500 Series GPUs are coming in 2027" with "next-gen Pensando 'Como' and 'Monza' networking", but does not name the scale-up protocol ([AMD, 2026-07-23](#)). On the Q1 2026 earnings call (third-party transcript; Astera's written results releases give no UALink switch date), Astera Labs' CEO said "Both Amazon as well as AMD have said that their ASIC and GPU will launch sometime in 2027, and we will certainly be prepared to intercept that launch with our UALink switch" ([Astera Labs Q1 2026 call, 2026-05-05](#)). That is a supplier describing its customers' plans, not a commitment from AMD.

Side-by-side comparison (as of 2026-10-28)

	NVLink 6	Native UALink	UALoE (Helios)
Governance	NVIDIA proprietary; Fusion licenses endpoints	Open consortium spec	Called an open standard; no consortium document found
Lane rate	Not published	200G (212.5 GT/s signaling); 128G option	200G on Ethernet ASICs
Per-GPU	3 TB/s (current spec; 3.6 TB/s at CES 2026 launch)	Implementation-defined	Up to 3.6 TB/s (derived)
Max domain	72; 1,152 on roadmap	1,024 (spec)	72
In-network compute	SHARP, shipping	Common Spec 2.0	Not stated
Switch suppliers	NVIDIA	Astera Labs, Upscale, Marvell custom (announced)	Broadcom Tomahawk 6
Status	Production shipments from fall 2026	No native switch confirmed shipping	In production; deployments from Q4 2026

Availability reality check

Statuses use each vendor's own wording.

Product	Status as of 2026-10-28
NVIDIA Vera Rubin NVL72 (NVLink 6)	Production: "ramping into full production", shipments "starting this fall" (NVIDIA, 2026-05-31)
AMD Helios (UALoE), OEM reference design	Production: "now in production"; OpenAI expects it online "beginning in the fourth quarter of 2026" (AMD, 2026-07-23); "volume deployments expected in 2H 2026" (AMD Helios)
HPE Juniper QFX5252 (Tomahawk 6 UALoE tray)	Ordered: first Helios order, \$1.2B from Vultr, 2026-09-30; no delivery date given (HPE)
Astera Labs native UALink switch	Announced: "poised to expand further into 2027 with the introduction of UALink-based products" (Q1 2026 call transcript)

Product	Status as of 2026-10-28
Astera Labs Scorpio X 320-lane	Shipping family, ramping part: Scorpio is "Now Shipping to Leading Hyperscalers" with "Open and Platform-Specific Protocols"; the X 320-lane has a "production ramp in 2H 2026" (Astera Labs, 2026-05-05). Not described as UALink.
Upscale SkyFabriX (115.2 Tbps)	Early access: supports ESUN, UALoE, SUE-T and Ethernet/IP, plus "evolving protocols including UALink"; GA "planned for early 2027" (Upscale, 2026-10-08)
Marvell custom UALink (SerDes, controller, switch-core IP)	Announced 2025-06-11, no date given (Marvell); bought XConn to "expand our UALink scale-up switch team" (Marvell, 2026-01-06)
Synopsys UALink IP	Licensable IP , status not stated (Synopsys); UALinkSec_200 runs on the "silicon-proven" 224G PHY (Synopsys, 2026-01-15)
Cadence UALink IP	IP in silicon: "UALink-200G controller and PHY running in 3nm" (Cadence, 2026-04-13)

The consortium-hosted white paper puts switching silicon at "Multiple vendors in development, evaluation hardware expected in late 2026" ([TASK Consultancy for UALink Consortium, 2026](#)). That matches what vendors say: controller and PHY IP runs in silicon, UALoE ships in Helios, and native switches arrive in 2027.

Procurement implications

- **Lock-in sits in the switch.** NVIDIA supplies the switch for NVLink, including under Fusion. Helios splits suppliers across AMD, Broadcom and the OEM. However, UALoE is not a published consortium standard, so check the documents behind any "multi-vendor" claim.
- **Native UALink buys optionality from 2027, not in 2026.** Interoperability between vendors stays unproven until compliance programs exist.
- **Upgrade paths.** Neither AMD nor HPE has said that Helios UALoE trays can be swapped for native UALink switches. Assume they cannot unless it is in the contract.

Questions to ask any scale-up vendor:

1. What is the per-GPU scale-up bandwidth for this exact SKU, in each direction, and where is it published?
2. Which document defines the protocol on the wire?
3. What all-reduce and all-to-all bandwidth have you measured on a full 72-GPU domain, and with which collective library version?
4. Is in-network reduction available now, and in which software release?
5. For native UALink: is the switch silicon announced, sampling or in production, and does it have compliance results and tested accelerator partners?
6. Can the domain be partitioned for multi-tenancy, and is UALinkSec or an equivalent encryption feature enabled?

Decision guide

This is our assessment, not a benchmark.

- **MoE at scale, delivered in 2026:** NVLink racks or Helios, both 72-GPU all-to-all domains. Compare measured all-to-all bandwidth and software maturity, not headline TB/s.
- **Memory-bound long-context inference:** HBM per domain matters more than the fabric. AMD lists 31 TB of HBM4 per Helios rack ([AMD Helios](#)); NVIDIA lists 20.7 TB per Vera Rubin NVL72 ([NVIDIA](#)).
- **Your own XPU, 2027-2028:** carry both NVLink Fusion and native UALink through design review. Fusion offers a fabric that already ships. UALink offers a choice of switch vendors, if Astera Labs, Upscale or Marvell hit their 2027 dates.
- **A few racks for enterprise or sovereign use:** the fabric comes with the GPU vendor. Do not pay extra for "UALink-ready" until a native switch ships.

Recheck after AMD's and Astera Labs' next earnings reports and any announcement of a UALink compliance program.

GPU Smith helps teams compare scale-up fabrics and turn vendor claims into acceptance tests. [Contact us](#).

Sources

1. UALink Consortium Releases the Ultra Accelerator Link 200G 1.0 Specification (news release), UALink Consortium, 2025-04-08. https://ualinkconsortium.org/wp-content/uploads/2025/04/UALink-1.0-Specification-PR_FINAL.pdf
2. Introducing UALink 200G 1.0 Specification (white paper), UALink Consortium, 2025-04. https://ualinkconsortium.org/wp-content/uploads/2025/04/UALink-1.0-White_Paper_FINAL.pdf
3. Specifications, UALink Consortium, accessed 2026-10-28. <https://ualinkconsortium.org/specification/>
4. UALink Consortium Publishes Four Specifications Defining In-Network Compute, Chiplets, Manageability and 200G Performance (news release), UALink Consortium, 2026-04-07. https://ualinkconsortium.org/wp-content/uploads/2026/04/UALink-2.0-Specification-PR_FINAL.pdf
5. UALink Roadmap Insights: Accelerating Open, Scalable AI Networking, UALink Consortium blog (Kurtis Bowman), 2026-03-06. <https://ualinkconsortium.org/blog/ualink-roadmap-insights-accelerating-open-scalable-ai-networking-1296/>
6. UALink: An Open, High-Efficiency Scale-Up Interconnect for AI (white paper), TASK Consultancy, hosted by UALink Consortium, 2026. https://ualinkconsortium.org/wp-content/uploads/2026/01/UALink_White_Paper_Publication_FINAL_UPDATED.pdf
7. NVIDIA NVLink, NVIDIA, accessed 2026-10-28. <https://www.nvidia.com/en-us/data-center/nvlink/>
8. NVIDIA Vera Rubin NVL72, NVIDIA, accessed 2026-10-28. <https://www.nvidia.com/en-us/data-center/vera-rubin-nvl72/>
9. NVIDIA NVLink Fusion, NVIDIA, accessed 2026-10-28. <https://www.nvidia.com/en-us/data-center/nvlink-fusion/>
10. NVIDIA GB200 NVL72, NVIDIA, accessed 2026-10-28. <https://www.nvidia.com/en-us/data-center/gb200-nvl72/>

11. NVIDIA Vera Rubin Ramps Into Full Production (press release), NVIDIA, 2026-05-31.
<https://nvidianews.nvidia.com/news/vera-rubin-full-production-agentic-ai-factory>
12. AMD Helios Rackscale Solution, AMD, accessed 2026-10-28.
<https://www.amd.com/en/products/rackscale-solutions/helios.html>
13. AAI 2026: AMD Launches AMD Helios Rackscale Solution for Frontier AI, AMD Newsroom, 2026-07-23.
<https://newsroom.amd.com/news/aai-2026-helios-update/>
14. AAI 2026: AMD Delivers Full-Stack Compute for the Agentic AI Era (press release), AMD, 2026-07-23.
<https://ir.amd.com/news-events/press-releases/detail/1294/aai-2026-amd-delivers-full-stack-compute-for-the-agentic-ai-era>
15. HPE accelerates AI deployments with first AMD "Helios" AI rack-scale architecture with open, scale-up networking built with Broadcom (press release), HPE, 2025-12-02.
<https://www.hpe.com/us/en/newsroom/press-release/2025/12/hpe-accelerates-ai-deployments-with-first-amd-helios-ai-rack-scale-architecture-with-open-scale-up-networking-built-with-broadcom.html>
16. HPE secures its first AMD Helios order in \$1.2 billion deal with Vultr (press release), HPE, 2026-09-30.
<https://www.hpe.com/us/en/newsroom/press-release/2026/09/hpe-secures-its-first-amd-helios-order-in-12-billion-deal-with-vultr.html>
17. Astera Labs Extends Leadership in Open, AI Scale-Up Networking with New 320 Lane Scorpio X-Series Smart Fabric Switch (press release), Astera Labs, 2026-05-05. <https://asteralabs.gcs-web.com/news-releases/news-release-details/astera-labs-extends-leadership-open-ai-scale-networking-new-320>
18. Astera Labs (ALAB) Q1 2026 Earnings Call Transcript, The Motley Fool, 2026-05-05.
<https://www.fool.com/earnings/call-transcripts/2026/05/05/astera-labs-alab-q1-2026-earnings-transcript/>
19. Astera Labs Reports Second Quarter 2026 Financial Results (Form 8-K Exhibit 99.1), Astera Labs via SEC EDGAR, 2026-08-04.
<https://www.sec.gov/Archives/edgar/data/0001736297/000173629726000033/q226exhibit991.htm>
20. Upscale Introduces Token Fabric (press release), Upscale, 2026-10-08.
<https://upscaleai.com/blogs/upscale-introduces-token-fabric-the-industrys-most-comprehensive-standards-based-networking-portfolio-for-ai-factories>
21. Marvell Expands Custom Compute Platform with UALink Scale-up Solution for AI Accelerated Infrastructure (press release), Marvell, 2025-06-11. <https://investor.marvell.com/news-events/press-releases/detail/94/marvell-expands-custom-compute-platform-with-ualink-scale-up-solution-for-ai-accelerated-infrastructure>
22. Marvell to Acquire XConn Technologies, Expanding Leadership in AI Data Center Connectivity (press release), Marvell, 2026-01-06. <https://investor.marvell.com/news-events/press-releases/detail/1004/marvell-to-acquire-xconn-technologies-expanding-leadership-in-ai-data-center-connectivity>
23. UALink IP, Synopsys, accessed 2026-10-28. <https://www.synopsys.com/designware-ip/interface-ip/ualink.html>
24. Securing UALink: Introducing Synopsys UALinkSec_200 Security Module, Synopsys, 2026-01-15.
<https://www.synopsys.com/articles/securing-ualink-security-module.html>
25. Industry's First UALink-200G Controller and PHY Running in 3nm, Cadence Community blog, 2026-04-13. https://community.cadence.com/cadence_blogs_8/b/data-center/posts/industry-first-ualink-200g-controller-and-phy-running-in-3nm

26. NVIDIA Kicks Off the Next Generation of AI With Rubin — Six New Chips, One Incredible AI Supercomputer (press release), NVIDIA, 2026-01-05. <https://nvidianews.nvidia.com/news/rubin-platform-ai-supercomputer>
27. NVIDIA Vera Rubin datasheet, NVIDIA, 2026-10-14. <https://dam-cdn.nvd.orangelogic.com/AssetLink/56p68o47y6f1yucpubl0ump1c0aif422.pdf>

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://ualinkconsortium.org/wp-content/uploads/2025/04/UALink-1.0-Specification-PR_FINAL.pdf
2. <https://www.hpe.com/us/en/newsroom/press-release/2025/12/hpe-accelerates-ai-deployments-with-first-amd-helios-ai-rack-scale-architecture-with-open-scale-up-networking-built-with-broadcom.html>
3. <https://www.nvidia.com/en-us/data-center/nvlink>
4. https://ualinkconsortium.org/wp-content/uploads/2025/04/UALink-1.0-White_Paper_FINAL.pdf
5. <https://newsroom.amd.com/news/aai-2026-helios-update>
6. https://ualinkconsortium.org/wp-content/uploads/2026/04/UALink-2.0-Specification-PR_FINAL.pdf
7. <https://ualinkconsortium.org/specification>
8. <https://ualinkconsortium.org/blog/ualink-roadmap-insights-accelerating-open-scalable-ai-networking-1296>
9. <https://www.sec.gov/Archives/edgar/data/0001736297/000173629726000033/q226exhibit991.htm>
10. https://ualinkconsortium.org/wp-content/uploads/2026/01/UALink_White_Paper_Publication_FINAL_UPDATED.pdf
11. <https://www.nvidia.com/en-us/data-center/vera-rubin-nv172>
12. <https://www.nvidia.com/en-us/data-center/nvlink-fusion>
13. <https://nvidianews.nvidia.com/news/rubin-platform-ai-supercomputer>
14. <https://dam-cdn.nvd.orangelogic.com/AssetLink/56p68o47y6f1yucpubl0ump1c0aif422.pdf>
15. <https://www.nvidia.com/en-us/data-center/gb200-nv172>
16. <https://nvidianews.nvidia.com/news/vera-rubin-full-production-agentic-ai-factory>
17. <https://www.amd.com/en/products/rackscale-solutions/helios.html>
18. <https://www.hpe.com/us/en/newsroom/press-release/2026/09/hpe-secures-its-first-amd-helios-order-in-12-billion-deal-with-vultr.html>
19. <https://ir.amd.com/news-events/press-releases/detail/1294/aai-2026-amd-delivers-full-stack-compute-for-the-agentic-ai-era>
20. <https://www.fool.com/earnings/call-transcripts/2026/05/05/astera-labs-alab-q1-2026-earnings-transcript>
21. <https://asteralabs.gcs-web.com/news-releases/news-release-details/astera-labs-extends-leadership-open-ai-scale-networking-new-320>
22. <https://upscaleai.com/blogs/upscale-introduces-token-fabric-the-industrys-most-comprehensive-standards-based-networking-portfolio-for-ai-factories>
23. <https://investor.marvell.com/news-events/press-releases/detail/94/marvell-expands-custom-compute-platform-with-ualink-scale-up-solution-for-ai-accelerated-infrastructure>
24. <https://investor.marvell.com/news-events/press-releases/detail/1004/marvell-to-acquire-xconn-technologies-expanding-leadership-in-ai-data-center-connectivity>

25. <https://www.synopsys.com/designware-ip/interface-ip/ualink.html>
26. <https://www.synopsys.com/articles/securing-ualink-security-module.html>
27. https://community.cadence.com/cadence_blogs_8/b/data-center/posts/industry-first-ualink-200g-controller-and-phy-running-in-3nm

THE PUBLISHER

About GPU Smith

GPU Smith provides independent engineering and research for private AI infrastructure. We help technical buyers and operators reason about GPU workload sizing, hardware selection, procurement, deployment and inference operations, from the compute system to the networking, power and cooling requirements around it.

Start with the workload

Useful infrastructure decisions begin with the models, data, throughput, latency and operating constraints a team actually has. GPU Smith helps connect those requirements with system design and deployment choices. The goal is a privately controlled AI environment whose capacity and operational demands are understood before a purchasing decision.

Hardware and supplier research

Our public reference library covers GPUs, complete systems and networking components. The server-vendor directory and manufacturing research help teams investigate suppliers, compare options and follow sources behind technical claims. We distinguish manufacturer specifications from measured benchmarks and advertised capabilities from tested configurations. Reference pages are research resources, not a live inventory listing or a binding equipment quote.

Deployment and operations

GPU Smith's engineering scope includes infrastructure integration, networking, power, cooling and the practical operation of inference workloads. Our published articles explain the tradeoffs behind deployment, procurement and ongoing operations so teams can ask better questions and document decisions.

Work with GPU Smith

Explore the [hardware library](#), [GPU references](#), [networking references](#) and [vendor research](#). [Contact GPU Smith](#) with your workloads and deployment constraints to discuss a project and confirm current scope, pricing and availability.

Inclusion of a manufacturer or product in our research does not imply a partnership, certification or endorsement.

Website: <https://gpusmith.com>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.