NVLink Fusion Semi-Custom AI Racks: Due Diligence

GPUSmith · GPU Smith · 2026-09-19 · nvlink fusion semi-custom ai racks · nvlink fusion · nvhbm · semi-custom ai · ai racks · custom accelerators · high-bandwidth memory · ai infrastructure · gpu procurement

Executive Summary

NVLink Fusion semi-custom AI racks are not a single purchasable system class. They are a framework for attaching a buyer's or partner's custom processor, called an XPU, to NVIDIA memory, scale-up interconnect, CPUs, networking, management software, and MGX rack infrastructure. NVIDIA announced the NVHBM extension on **August 26, 2026**, and named Amazon's Annapurna Labs as the first collaborator (blogs.nvidia.com). AWS describes the work as an expansion of support, not a shipping product (press.aboutamazon.com). A commitment decision should therefore be tied to qualification evidence and contractual milestones, not the umbrella brand.

NVHBM moves NVIDIA's memory controller into the high-bandwidth memory base die (blogs.nvidia.com). NVIDIA claims up to **30% more bandwidth per stack**, **15% lower HBM power**, and **25% more compute-die area** than standard HBM4E (developer.nvidia.com) (developer.nvidia.com) (developer.nvidia.com). Public materials do not disclose workload, silicon revision, sample size, measurement method, or whether those figures came from simulation or physical silicon. Buyers should require an agreed floor, test method, and layout denominator.

The vendor's **1 GW, 2,000 W XPU** illustration says memory-power savings could allow up to 15,000 more XPUs (developer.nvidia.com). Reconstructing it makes the denominator explicit. If 1 GW means IT power and every XPU draws 2 kW, the baseline is 500,000 XPUs. Adding about 15,000 units requires an XPU-level power reduction near 2.91%. A 15% cut to HBM alone produces that result only if HBM represents roughly **19.4% of XPU power**. If 1 GW instead means facility power, power usage effectiveness (PUE) must be applied because PUE is facility energy divided by IT energy (energy.gov). At PUE 1.2, the same nameplate supports about 416,667 baseline XPUs, not 500,000.

The prudent options are **wait**, **prototype**, **dual-source**, or **commit conditionally**. Most enterprises should prototype only after receiving a frozen interface package, memory-partner nomination, package and thermal model, firmware ownership map, volume schedule, and rack acceptance plan. NVIDIA specifies NVLink 6 at **3.6 TB/s per XPU across 72 XPUs**, while configurations up to **1,152 XPUs** are explicitly future roadmap items (www.nvidia.com) (www.nvidia.com). Reserve power or supplier capacity only against dated deliverables, remedies, and measured acceptance criteria.

Introduction and Background

The buyer's question is not whether NVIDIA has assembled a credible set of technologies. It is whether a particular **semi-custom AI rack design** has enough defined, qualified, and supportable content to justify architecture lock-in, engineering spend, supplier reservations, or data-center power. This report treats **NVLink Fusion rack architecture** and **custom AI accelerator rack due diligence** as one decision problem. That distinction matters because NVLink Fusion spans several layers and several companies. NVIDIA calls it connective technology and intellectual property for custom XPUs and CPUs (www.nvidia.com), while MediaTek says a finished deployment still requires chip-to-rack engineering, qualification, and supply-chain work (www.mediatek.com).

The original NVLink Fusion program was unveiled on **May 18, 2025** (nvidianews.nvidia.com). Its first custom-silicon ecosystem list included MediaTek, Marvell, Alchip, Astera Labs, Synopsys, and Cadence (nvidianews.nvidia.com). The **NVHBM architecture** announced in August 2026 adds a custom memory-controller and base-die option. NVIDIA says memory partners will validate and offer it, but its announcement does not name those suppliers (blogs.nvidia.com).

For an enterprise buyer, accelerator program leader, or investor, the meaningful unit of diligence is therefore the named configuration and responsibility chain. GPU Smith's published method emphasizes written acceptance criteria, as-built documentation, part numbers, link budgets, power, cooling, lead times, and pricing (gpusmith.com) (gpusmith.com). That independent-advisor perspective is useful here: convert each roadmap statement into an evidence request, acceptance threshold, owner, and date.

Key Changes

NVHBM relocates control into the memory base die

Conventional HBM procurement separates a standardized memory stack from more controller and physical-interface logic on the accelerator. **NVHBM meaning** is narrower and more architectural: NVIDIA designs a custom HBM base die and controller, then memory providers validate and manufacture the stack (www.nvidia.com). The

proposed benefit is co-design across the XPU and memory boundary.

The claimed area benefit uses a different denominator from interface-block area or total die area. Buyers should ask for three separate layouts: interface-block area, net usable logic area, and total compute-die area. The contract should say which metric determines acceptance.

The comparison baseline also needs precision. JEDEC's published HBM4 standard specifies up to **8 Gb/s across a 2,048-bit interface**, yielding up to **2 TB/s per stack** (jedec.org) (jedec.org). NVIDIA compares NVHBM with "standard HBM4E," but the public NVHBM material does not specify a JEDEC revision, stack height, capacity, speed bin, temperature, error-correction mode, or traffic pattern. Samsung reported shipping HBM4E samples in May 2026, which is evidence of sample-stage standard memory, not a like-for-like NVHBM benchmark (news.samsung.com).

NVLink 6 raises the announced scale-up ceiling

NVIDIA describes **NVLink 6** as a bidirectional **3.6 TB/s per GPU** link fabric (developer.nvidia.com). In a 72-accelerator domain, the company states **260 TB/s of rack-level bandwidth** (developer.nvidia.com). Its described topology is all-to-all, allowing any accelerator to communicate with any other (developer.nvidia.com).

These are interface and topology specifications, not application throughput. The diligence plan must distinguish lane or fabric bandwidth, collective-communication throughput, memory bandwidth, and workload output. NVIDIA's own diagnostics documentation says topology information does not test bandwidth (docs.nvidia.com), and passive link-error evidence does not validate a path under load (docs.nvidia.com). A rack acceptance test must exercise the intended collectives, message sizes, concurrency, and failure modes.

The ecosystem expands, but ownership remains distributed

The program makes components available to semi-custom designers, including **NVLink chiplets, NVLink-C2C, NVLink switches, and MGX racks** (blogs.nvidia.com). That availability does not make every possible XPU-memory-rack combination prequalified.

Partner announcements illustrate the division of labor:

- **AWS and Annapurna Labs:** The planned work combines Trainium, NVHBM, GPUs, and a common rack architecture, with unnamed memory suppliers involved (press.aboutamazon.com).
- **MediaTek:** Customers bring the XPU and tailor connectivity, memory, packaging, performance, and power characteristics (www.mediatek.com).
- **Marvell:** Its announced scope includes custom XPUs and compatible scale-up networking, while NVIDIA supplies supporting platform technologies (www.marvell.com).
- **Alchip:** Its design flow covers power delivery, die-to-die electrical interconnect, and thermal characterization (alchip.com).
- **Astera Labs:** It announced plans for custom connectivity developed with hyperscaler partners, using roadmap rather than general-availability language (asteralabs.com).
- **Fujitsu:** Its NVLink Fusion scope is a co-development program integrating MONAKA CPUs and NVIDIA GPUs, with a broader 2030 horizon (global.fujitsu).

- **d-Matrix:** Its announced Raptor XPU was expected to tape out by year-end 2026, with MGX rack availability targeted for Q4 2027 (d-matrix.ai) (d-matrix.ai).

Architecture and Responsibility Boundaries

The correct **NVLink Fusion vendor comparison** is not a single scorecard ranking chip companies. It is a responsibility matrix for the named program. *Table 1* separates what the platform advertises from what the buyer still has to contract and verify.

Layer	Likely design or supply owner	Evidence available by September 2026	Buyer acceptance evidence
Custom XPU	Buyer, hyperscaler, or accelerator partner	Customer supplies its design and workload targets (nvidianews.nvidia.com)	Frozen RTL and interfaces, workload benchmark, power states, errata process
NVHBM controller and base die	NVIDIA design, memory-partner validation and supply	Controller is in the HBM base die; suppliers are not publicly named (blogs.nvidia.com)	Named supplier, qualification vehicle, stack capacity, speed bin, yield, lifetime and second-source plan
Advanced package	XPU or ASIC partner, foundry and assembly chain	Alchip lists 2 nm and 3 nm platforms plus CoWoS variants (alchip.com)	Package drawing, warpage, power integrity, thermal model, known-good-die strategy and capacity allocation
Scale-up fabric	NVIDIA NVLink chiplet, NVLink switch, partner integration	72-way all-to-all and 3.6 TB/s per XPU are current specifications; 1,152 is roadmap (developer.nvidia.com)	Link margin, BER, collective throughput, degraded-mode behavior, service tooling
CPU coherent link	NVIDIA NVLink-C2C plus CPU or silicon partner	NVIDIA describes a coherent connection to custom silicon (www.nvidia.com)	Coherency matrix, cache and reset behavior, firmware-version compatibility
Rack and power	NVIDIA MGX, ODM, power and mechanical suppliers	OCP Open Rack V3 defines intermateability, but this does not certify a specific MGX build (opencompute.org)	Bill of materials, fault domains, redundant-power definition, short-circuit study, maintainability test
Cooling	Rack integrator, CDU supplier, facility engineer	OCP requires CDU designs to be tested and qualified (opencompute.org)	Coolant, flow, pressure, temperature, water quality, leak response, facility-water interface
Firmware and operations	XPU vendor, NVIDIA, BMC supplier, integrator	Fabric Manager must match the loaded driver stack (docs.nvidia.com)	Signed update path, rollback, telemetry ownership, version matrix, offline recovery and support boundary

The table shows why a platform announcement is not a rack warranty. Every row crosses at least two organizational boundaries. A buyer should name one accountable integration owner and attach the complete responsibility assignment matrix to the statement of work.

Standards compatibility is interface-specific

NVLink is proprietary, according to NVIDIA's own infrastructure documentation (docs.nvidia.com). That does not make the rack closed at every layer. It means buyers should not treat NVLink bandwidth or NVHBM interchangeability as if they carried the same multivendor compliance regime as PCI Express, Compute Express Link (CXL), Universal Chiplet Interconnect Express (UCIe), or Redfish.

- **PCIe:** PCI-SIG describes multivendor interoperability as an explicit goal (pcisig.com). Its Integrators List records products that completed compliance-workshop testing (pcisig.com).
- **CXL:** The consortium calls CXL an open standard, and CXL 4.0 retains backward compatibility with earlier major versions (computeexpresslink.org) (computeexpresslink.org).
- **UCIe:** UCIe covers package-level physical, protocol, software, and compliance layers (uciexpress.org). Its management mechanisms include firmware download, thermal management, error reporting, and telemetry, yet vendor-specific drivers remain possible (uciexpress.org) (uciexpress.org).
- **Redfish:** DMTF defines it as an interoperable, multivendor, remote-management interface, while allowing optional vendor elements (redfish.dmtf.org) (redfish.dmtf.org).

The procurement rule is simple: require a named standard and compliance artifact for each boundary that claims openness. For proprietary boundaries, require bilateral interoperability results, source and update rights where appropriate, and a supported version matrix.

Implementation Considerations and Process Changes

Convert claims into a qualification plan

An **AI rack procurement checklist** should be a gated evidence plan, not a feature list. At minimum, the buyer should require:

- **Architecture baseline:** One controlled drawing covering XPU, NVHBM stacks, package, CPU, switches, NICs, DPUs, storage, power shelves, cooling distribution units, and facility connections.
- **Claim dictionary:** Exact numerator, denominator, baseline, operating point, and revision for bandwidth, power, usable area, and end-to-end performance.
- **Silicon maturity:** Tape-out, bring-up, stepping, characterization, reliability qualification, and production-release dates.
- **Memory qualification:** Named provider, stack and base-die revision, capacity, speed, thermals, error correction, repair policy, and allocation.
- **Yield model:** XPU die yield, base-die yield, DRAM-stack yield, package yield, known-good-die screen, and final-system yield.
- **Thermal envelope:** Component power maps, coolant inlet range, pressure drop, flow, CDU approach temperature, residual air load, and throttling curve.
- **Electrical evidence:** Load steps, power excursions, rail margin, busbar rating, power-supply redundancy, breaker coordination, and rack grounding.
- **Interconnect tests:** Bit-error rate, margin, collective throughput, congestion, failover, and behavior after link isolation.

- **Firmware map:** Owner and supported version for XPU firmware, switch firmware, BMC, fabric manager, driver, orchestration, and secure boot.
- **Reliability plan:** Component lifetime, accelerated stress method, field-replaceable units, spare ratios, repair time, and fault-injection results.
- **Volume plan:** Wafer, HBM, substrate, assembly, test, switch, optics, rack, CDU, and commissioning capacity by quarter.
- **Support model:** Severity definitions, escalation chain, remote and air-gapped procedures, logs, diagnostic access, and end-of-support dates.
- **Acceptance test:** Workload-level throughput, latency, utilization, power, thermals, and recovery measured over a stated duration.
- **Commercial terms:** Milestone payments, allocation priority, price-adjustment formula, cancellation rights, warranty start, and remedies for missed acceptance.

ASHRAE says a power-supply nameplate is not actual in-use draw and recommends workload-based power modeling (handbook.ashrae.org) (handbook.ashrae.org). This supports two separate planning values: a protected electrical maximum and an empirically measured workload distribution.

Qualify the thermal and facility system as a system

High-density racks move risk from room cooling into liquid loops, controls, fittings, and water chemistry. Uptime Institute says liquid cooling is typically used above **50 kW per rack**, and near **150 kW** generally requires almost total liquid cooling (intelligence.uptimeinstitute.com) (intelligence.uptimeinstitute.com). Cold plates may still leave **5% to 30%** of rack heat to air (intelligence.uptimeinstitute.com).

The test program should include:

- **Hydraulic proof:** OCP guidance calls for hydrostatic testing at **1.3 to 1.5 times working pressure** (opencompute.org).
- **Coolant commissioning:** Final flushing should use the actual operating coolant (opencompute.org).
- **Water quality:** ASHRAE ties conformance to long-term reliability, so define chemistry limits, sampling, alarms, and remediation (handbook.ashrae.org).
- **Actual heat load:** Cooling capacity should match measured heat, including residual air cooling and CDU approach temperature (handbook.ashrae.org).
- **Failure sequences:** Exercise pump, power, sensor, valve, network, and controller failures under load, then verify automatic and manual recovery.
- **Maintainability:** Demonstrate service isolation without draining unrelated racks and define whether N+1 means component or complete-system redundancy.

Gate capacity reservations

Reservations should unlock in stages. A reasonable sequence is:

1. **Evaluation gate:** Architecture pack, workload model, preliminary package and thermal model, supplier letters.
2. **Prototype gate:** Frozen interfaces, emulator or FPGA evidence, memory test vehicle, initial firmware and rack design.

3. **Qualification gate:** Production-intent silicon, full package, environmental and reliability results, integrated rack tests.
4. **Volume gate:** Qualified bill of materials, yield and throughput evidence, allocation contracts, support readiness.
5. **Site gate:** Approved electrical and cooling design, commissioned facility loop, staged delivery and acceptance.

This structure aligns payments and power reservations to evidence. It also prevents a nominally available component from being mistaken for a qualified rack. The original 2025 availability statement covered design services and solutions from ecosystem vendors, not completed customer XPU racks (nvidianews.nvidia.com).

Decision Scenarios

Table 2 maps evidence maturity to a decision. The appropriate path depends less on the size of the announced ecosystem than on the buyer's ability to absorb integration and schedule risk.

Decision	When it fits	Required controls	Capital and power posture
Wait	Workload is not demonstrably bandwidth-bound; production timing or memory source is unnamed	Maintain interface watch, benchmark standard alternatives, preserve facility optionality	No dedicated power reservation; avoid nonrefundable supplier commitments
Prototype	Strategic workload and XPU differentiation are plausible, but silicon and rack evidence are incomplete	Small test vehicle, capped engineering budget, explicit exit criteria, reproducible benchmarks	Reserve lab capacity only; use milestone-priced nonrecurring engineering
Dual-source	Schedule matters and a standard-HBM or non-NVLink route can meet minimum requirements	Common workload harness, abstraction at management and scale-out layers, separate bills of material	Hold conditional capacity for both paths, then down-select after qualification
Commit conditionally	Named suppliers, production-intent silicon, full-rack qualification, yield and support evidence meet thresholds	Dated acceptance gates, allocation terms, remedies, spares and lifecycle plan	Phase site and supplier reservations against passed gates

The most defensible default is **prototype or dual-source**, not an unconditional fleet commitment. A full commitment becomes rational when the economic value of differentiated XPU throughput exceeds integration cost and schedule risk under conservative, measured inputs.

Data Analysis and Evidence

Reconstructing the 1 GW example

The basic equation is straightforward: **XPU count = IT power budget / per-XPU draw**. The ambiguity lies in whether 1 GW describes facility input or IT load, and how much of the 2,000 W XPU belongs to HBM. PUE is the ratio of facility energy to IT energy, not a measure of compute productivity (energy.gov). DOE uses **1.6** as an

average-data-center benchmark, while Uptime reported **1.58** for its 2023 industry series (energy.gov) (journal.uptimeinstitute.com). Those are context values, not forecasts for a particular new site.

Table 3 shows a sensitivity model for a **1 GW facility input**, **2 kW per XPU**, and three utilization levels. All figures are calculated, not observed NVHBM results.

PUE assumption	IT power, MW	Physical XPUs at 2 kW	Effective XPUs at 60% utilization	At 80%	At 90%
1.10	909.1	454,545	272,727	363,636	409,091
1.20	833.3	416,667	250,000	333,333	375,000
1.40	714.3	357,143	214,286	285,714	321,429
1.60	625.0	312,500	187,500	250,000	281,250

The sensitivity is material. Moving from PUE 1.10 to 1.60 reduces theoretical physical XPU capacity by 142,045 before utilization is considered. Raising sustained utilization from 60% to 80% at PUE 1.20 adds 83,333 effective-XPU equivalents without adding hardware. This is why utilization, facility overhead, and non-XPU IT loads must appear beside component efficiency.

Now isolate the vendor's 15,000-unit claim. With a 500,000-XPU baseline, adding 15,000 means the improved design draws 500,000 / 515,000, or about **97.09%**, of baseline XPU power. That is a **2.91% XPU-level reduction**. If the only saving is 15% of HBM power, HBM must account for approximately **19.4% of baseline XPU power**. At a 10% HBM share, the implied gain is about 7,614 XPUs. At 20%, it is about 15,464. At 30%, it is about 23,560. These are algebraic scenarios, not forecasts.

The **memory-bandwidth-per-watt** claim also needs like-for-like definitions. If bandwidth rises 30% while HBM power falls 15%, the mathematical ratio becomes 1.30 / 0.85, or **1.529**, implying up to 52.9% more stack bandwidth per HBM watt. That calculation applies only to the vendor's maxima, the same baseline, and HBM power alone. No disclosed test configuration supports extending this stack-only ratio to XPU, rack, or workload efficiency. It should not be called a benchmark.

TCO variables buyers can actually audit

An **NVLink Fusion total cost of ownership** model should separate capital, energy, site, software, and schedule effects:

- **Capital:** XPU, NVHBM, package, switches, CPUs, NICs, DPUs, racks, power, CDU, facility modifications, spares, and nonrecurring engineering.
- **Energy:** Measured XPU power by workload, other IT power, PUE by load and season, electricity price, and demand charges.
- **Availability:** Useful throughput after utilization, planned maintenance, failures, checkpoint or restart overhead, and degraded-mode operation.
- **Schedule:** Engineering labor, tape-out and qualification dates, facility reservation lead time, and cost of delayed service.
- **Lifecycle:** Firmware support, security updates, repair inventory, memory and package sourcing, refresh compatibility, and decommissioning.

For illustration only, the EIA reported a 2025 US all-sector average retail electricity price of **13.63 cents/kWh** (eia.gov). A project should replace it with site-specific energy, transmission, demand, tax, and curtailment terms. Likewise, liquid cooling cannot be assigned a generic saving: an LBNL retrofit study measured **4% overall data-center savings**, while estimating 15% to 20% only under a different heat-rejection condition (bies.lbl.gov) (bies.lbl.gov).

Implications and Future Directions

The near-term opportunity is **co-design**, not guaranteed commoditization. The central **NVLink Fusion ecosystem risks** are the larger qualification surface and dependence on an unnamed memory supply path. Moving the controller into the base die could allow a custom accelerator program to optimize memory scheduling, interface area, and power together. SK hynix has announced a multi-year next-generation-memory partnership with NVIDIA, including memory for Vera Rubin, but that announcement does not name NVHBM (news.skhynix.com) (news.skhynix.com). Buyers should not infer a qualified supplier from a general partnership.

Open standards will remain important around the proprietary scale-up island. PCIe 7.0 specifies **128 GT/s** and up to **512 GB/s bidirectional on x16**, but PCI-SIG said preliminary testing would begin in 2026 before an official compliance program (pcisig.com) (pcisig.com). UCIe 3.0 supports **48 and 64 GT/s**, yet implementation choices still matter (uciexpress.org). A buyer can preserve optionality through standard management, scale-out, storage, host, and facility interfaces even when the XPU scale-up fabric is proprietary.

Three verification milestones should change the investment case:

- **Production-intent NVHBM evidence:** Named memory provider, qualification revision, capacity and speed, error behavior, measured bandwidth and power across temperature, yield, and volume date.
- **Integrated rack evidence:** Production XPU and switches, full firmware stack, workload and collective tests, facility-representative power and cooling, fault injection, and maintainability.
- **Commercial evidence:** Firm allocations, support ownership, price and escalation rules, warranty, lifecycle, and remedies tied to acceptance.

Until those exist, forecasts should carry ranges and probabilities. The roadmap to **1,152 accelerators** is potentially significant, but it has no public delivery date on the cited NVIDIA page. The nearer 72-XPU domain is the appropriate basis for current architectural modeling.

Frequently Asked Questions (FAQs)

What is NVHBM?

NVHBM is NVIDIA's custom high-bandwidth memory architecture for NVLink Fusion. It places an NVIDIA-designed memory controller in the HBM base die and is intended to be validated and offered through memory partners. It is not simply another public JEDEC generation name, and public materials do not establish drop-in interchangeability with standard HBM4E.

Is the claimed 30% end-to-end gain a benchmark?

No public test conditions support calling it a benchmark. NVIDIA calls it an end-to-end performance increase arising from co-designed improvements, but does not publish workload, system configuration, silicon revision, sample size, or measurement method. Buyers should classify it as an announced vendor claim until reproducible evidence is available.

What does NVLink Fusion supply?

Depending on the engagement, it can supply or enable NVLink chiplets, NVLink-C2C, switches, NVHBM technology, NVIDIA CPUs and networking, management software, MGX systems, and rack designs. The XPU designer, silicon and packaging partners, memory provider, rack integrator, cooling supplier, and operator still have program-specific deliverables.

Should a buyer reserve power now?

Only conditionally. A reservation should be tied to named hardware, measured workload power, PUE and non-XPU load assumptions, a cooling design, shipment milestones, and exit rights. A 1 GW headline without those denominators is not a capacity plan.

How should a custom AI accelerator rack be compared with alternatives?

Use effective workload throughput per constrained resource, including rack power, facility power, capital, time, and support risk. Run the same model and acceptance harness against a standard-HBM design and at least one alternative scale-up or scale-out architecture. Do not compare interface maxima with measured application results.

Conclusion

NVLink Fusion and NVHBM offer a coherent proposition for organizations that already have a differentiated XPU thesis: combine custom compute with NVIDIA's memory, scale-up, networking, management, and rack ecosystem. The public evidence supports the existence of the program, named design partners, an announced 72-XPU NVLink 6 architecture, and a defined controller relocation into the HBM base die. It does not yet establish production NVHBM performance, supplier identity, price, yield, volume timing, or an independently measured end-to-end advantage.

The buying decision should therefore follow evidence maturity. **Wait** when workload differentiation is weak or supplier facts are unnamed. **Prototype** when the architecture is strategically relevant but silicon and rack qualification are incomplete. **Dual-source** when schedule matters and a standard alternative can preserve leverage. **Commit conditionally** only after production-intent memory, package, XPU, firmware, thermal, and full-rack results meet written thresholds.

The 1 GW reconstruction demonstrates the discipline required. “15% lower HBM power” does not mean 15% more rack capacity. Under the vendor's 500,000-XPU baseline, the claimed 15,000-unit increment implies about a 2.91% reduction in total XPU power and an HBM share near 19.4%. PUE, other IT loads, utilization, and cooling can move

the result materially. Treat every headline as an input to be decomposed, not as an investment output. That is the practical due-diligence standard for any NVLink Fusion semi-custom AI rack commitment.
