



Immersion vs Direct-to-Chip Cooling for GPU Clusters

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 · immersion cooling vs direct-to-chip for gpu clusters · is immersion cooling production-ready for a 20mw ai deployment · single-phase vs two-phase immersion cooling gpu · immersion cooling nvidia warranty · direct liquid cooling vs immersion cooling data center · immersion cooling fluid pfas two-phase · ocp immersion requirements

Original article: <https://gpusmith.com/articles/en/immersion-vs-direct-to-chip-cooling-gpu-clusters>



In this article

1. The short answer
 2. The taxonomy: six kinds of direct liquid cooling
 3. What immersion is, in OCP's terms
 4. Standards maturity
 5. Fluids, fire and PFAS
 6. Density, weight and floor loading
 7. Serviceability and operations
 8. Resilience and redundancy
 9. OEM support and warranty
 10. Facility water temperatures
 11. Is immersion production-ready at 20 MW?
 12. Decision matrix
 13. FAQ
 14. Sources
-

The short answer

For NVIDIA's current rack-scale platforms the choice is mostly made for you: they are cold-plate designs. NVIDIA describes the GB200 NVL72 as "a rack-scale, liquid-cooled design" connecting 36 Grace CPUs and 72 Blackwell GPUs ([NVIDIA GB200 NVL72](#)), and HPE's QuickSpecs for its build of that rack state that "the compute trays are cooled with cold plates that are mounted on the CPUs and GPUs" ([HPE GB200 NVL72 QuickSpecs](#)). The GB300 NVL72 "features a fully liquid-cooled, rack-scale architecture" ([NVIDIA GB300 NVL72](#)). Immersion remains an option for servers you can qualify yourself, or that a vendor sells as an immersion configuration with a written warranty.

Market data agrees. Dell'Oro says single-phase DLC "accounts for the vast majority of liquid-cooled capacity coming online today", while immersion "is finding its place through selective adoption" ([Dell'Oro, 2026-01-08](#)).

The taxonomy: six kinds of direct liquid cooling

Uptime Institute counts six commercially available direct liquid cooling (DLC) categories: three cold-plate types (water cold plates, single-phase dielectric cold plates, two-phase dielectric cold plates) and three immersion types (chassis immersion, single-phase immersion, two-phase immersion) ([Uptime Institute, 2022-07-06](#)).

Category	What touches the silicon	Heat leaves the IT as	Typical form
Water cold plate (single-phase DTC)	Metal plate, water-based coolant inside	Warm liquid in a secondary loop	Standard-width rack, manifolds, CDU
Two-phase dielectric cold plate	Metal plate, boiling dielectric inside	Vapor, condensed at the rack or CDU	Rack, refrigerant-style loop

Category	What touches the silicon	Heat leaves the IT as	Typical form
Chassis (enclosed) immersion	Dielectric fluid inside a sealed sled	Liquid via dripless connectors	Vertical rack of sealed chassis
Single-phase immersion (open bath)	Dielectric fluid bath	Pumped or convected liquid to a heat exchanger	Horizontal tank, servers lowered in from the top
Two-phase immersion	Boiling dielectric bath	Vapor, condensed at a coil above the bath	Sealed horizontal tank

"Direct-to-chip" in vendor language almost always means the first row: a cold plate on each GPU and CPU, fed by a secondary loop from a coolant distribution unit (CDU). HPE's GB200 NVL72 CDU passes that heat to the facility loop through a liquid-to-liquid heat exchanger ([HPE QuickSpecs](#)). For acceptance testing, see our [cold plate GPU rack commissioning checklist](#).

What immersion is, in OCP's terms

As of October 2026, the current Open Compute Project (OCP) revision is **Immersion Requirements Rev 2.10, last updated 2023-08-18**, after 1.00 (2019-05-20) and 2.00 (2022-01-21) ([OCP Rev 2.10](#)). That explains why some sources cite [Rev 1.0](#) and others, such as [Vertiv's 2022 blog](#), cite Rev 2.0. OCP's [immersion project page](#) lists Rev 2.10 (titled v2.1 there) as current and marks rev 2 and rev 1 as legacy. It is "a requirements document and not a specification" ([Rev 2.0](#)).

Its definitions (unchanged from Rev 2.0):

- **Single-phase immersion:** the liquid "is heated and cooled without changing phase"; circulation "can be done through natural or forced convection."
- **Two-phase immersion:** the liquid "is heated and changes to gas"; the vapor is condensed and returns to the bath "through a natural buoyancy-driven flow."
- **Enclosed chassis:** sealed sleds, usually stacked in a rack, "pulled out and inserted in horizontal plane for service."
- **Open bath:** a tank in which multiple IT systems share one body of fluid. Such tanks "can be fully sealed, but are always opened from the top to service IT equipment" ([Rev 1.0](#); [Rev 2.10](#)).

OCP has also published a [Base Specification for Immersion Fluids](#) (Rev 1.0, 2022-12-01) [Warranty Guidelines for Immersion Cooled Technology](#) (Rev 1.0, 2024-04-26) and [Server Component Immersion Material Compatibility Testing](#) guidelines (v1.5, 2026-06-05).

Standards maturity

Cold-plate DLC has its own standards gap. Uptime wrote that "there is no comparable standard coolant or line of demarcation for DLC" and that mechanical interface standards "will take years to materialize in product form" ([Uptime, 2022](#)). In April 2026 it added that "the industry also lacks a singular consensus on the division of ownership and maintenance for DLC components" ([Uptime, 2026-04-15](#)). What cold plates have and immersion lacks is a de facto reference: the GPU vendor's own rack design, built and warranted by OEMs.

Fluids, fire and PFAS

Single-phase fluids

OCP groups immersion liquids into "hydrocarbons (i.e. mineral, synthetic or bio oils) and fluorocarbons (i.e. fully engineered liquids)" ([Rev 1.0](#)). Rev 2.10 sets minimums for any immersion fluid: a flash point (closed cup) of at least 150 °C, an auto-ignition point of at least 300 °C, and volume resistivity of at least 2.00 GΩm when new ([Rev 2.10](#)). A flash point at that level still allows the fluid to burn. Uptime's June 2024 report on immersion fluids is titled "Immersion fluids hold promise, but fire risk a concern" ([Uptime Intelligence, 2024-06-25](#)). (The full report is paywalled.) Expect your fire engineer and insurer to ask for the fluid's SDS and TDS.

Two-phase immersion and the 3M exit

Two-phase immersion relied on fluorinated fluids. On 2022-12-20 3M announced it would "exit per- and polyfluoroalkyl substance (PFAS) manufacturing" and "work to discontinue the use of PFAS across its product portfolio by the end of 2025" ([3M, 2022-12-20](#)). Uptime says interest in immersion now "almost exclusively means single-phase immersion", which "is largely due to 3M's exit from the segment" ([Uptime Intelligence, 2024](#)).

OCP's fluid specification defines PFAS as "any substance that contains a -CF₂- or -CF₃ group" ([OCP immersion fluid spec](#)). At EU level, ECHA's risk assessment committee (RAC) adopted its opinion on the universal PFAS restriction in early March 2026 ([ChemLinked, 2026-03-06](#)). The socio-economic committee (SEAC) published a draft opinion in March 2026 for consultation until 2026-05-25. As of 2026-10-28 its final opinion has not been reported as adopted; it is expected by the end of 2026, after which the Commission drafts the restriction for the REACH Committee, with Parliament and Council scrutiny before it enters Annex XVII ([Linklaters, 2026-04-09](#)). For any fluorinated fluid, ask whether it contains those groups and what supply commitment the vendor gives.

Two-phase direct-to-chip

Two-phase cold plates boil a dielectric inside the plate rather than a bath. OCP's March 2026 guidelines for these racks (Revision 1.1) call for fluids that are non-flammable, ASHRAE 34 class A1, with GWP below 700, near-zero ODP, and a boiling point between -26.3 °C and 60 °C ([OCP two-phase cold plate guidelines](#)). Dell'Oro expects two-phase DLC "to expand gradually", with adoption accelerating "once chip-level TDPs and thermal flux exceed the practical limits of single-phase systems" ([Dell'Oro](#)). Apply the same PFAS check to these fluids.

Density, weight and floor loading

Filled tanks "can get quite heavy", says OCP's fluid spec, which caps fluid density at 2,000 kg/m³ at 25 °C, a ceiling chosen so as not to exclude any liquid ([OCP immersion fluid spec](#)). Rev 2.10 asks vendors to publish bare, full and IT static load in kN/m² ([Rev 2.10](#)).

Worked example (derived from vendor datasheets; illustrative, not a design calculation).

Inputs, single-phase tank: the GIGABYTE 25U EIA tank (A1P0-EB0) is rated 80 kW with 25 °C inlet water and 60 kW with 35 °C. The tank measures 0.91×1.16 m and weighs 450 kg empty and 1,065 kg with 750 L of coolant. Its CDU measures 0.86×0.48 m and weighs 325 kg ([GIGABYTE A1P0-EB0](#)).

Inputs, cold-plate rack: HPE's GB200 NVL72 is $1,068 \times 600$ mm, 3,245 lb fully loaded, with a rack TDP of "132 kW nominal" and peak electrical design power of about 192 kW ([HPE QuickSpecs](#)).

Derived figure	Immersion tank (no servers)	GB200 NVL72 rack (loaded)
Footprint	$0.91 \times 1.16 = 1.056 \text{ m}^2$	$1.068 \times 0.600 = 0.641 \text{ m}^2$
Mass	1,065 kg	$3,245 \text{ lb} \times 0.4536 = 1,472 \text{ kg}$
Distributed load	$1,065 / 1.056 = 1,009 \text{ kg/m}^2 \approx 9.9 \text{ kN/m}^2$	$1,472 / 0.641 = 2,297 \text{ kg/m}^2 \approx 22.5 \text{ kN/m}^2$
Heat per footprint	$80 / 1.056 = 76 \text{ kW/m}^2$ (55 kW/m ² incl. its CDU)	$132 / 0.641 = 206 \text{ kW/m}^2$ (CDU excluded)
Implied fluid density (derived, not a datasheet value)	$(1,065 - 450) / 750 \approx 0.82 \text{ kg/L}$	n/a

The tank excludes servers; the rack excludes its CDU and aisle space. Even so, the pattern holds: a horizontal tank spreads its load and gives up heat density per square metre, while a DLC rack concentrates both. Check both against slab rating and delivery path; our [GPU rack floor loading and delivery path worksheet](#) walks through that check.

Scaling to 20 MW of IT (same inputs, illustrative). At 80 kW per tank: $20,000 / 80 = 250$ tanks, $250 \times 750 \text{ L} = 187,500 \text{ L}$ of fluid, about $250 \times 615 \text{ kg} \approx 154 \text{ t}$ of fluid mass. If facility water is 35 °C instead of 25 °C, the rating falls to 60 kW: $20,000 / 60 \approx 334$ tanks and about 250,000 L. Fluid capex is then simply *fill volume* $\times$ *your quoted price per litre*. We found no independent public price, so get it quoted with a top-up allowance.

Serviceability and operations

Cold-plate trays are serviced much like air-cooled ones, if the quick-disconnects behave. Immersion changes the work:

- **Access.** Open baths are serviced from the top ([OCP Rev 1.0](#)). Lifting equipment and ceiling clearance become part of the design. Rev 2.10 asks vendors to state height clearance "incl. clearance for lid, servicing mechanisms" ([Rev 2.10](#)).
- **Drip time.** Higher-viscosity fluids increase the time equipment must drip before servicing ([OCP fluid spec](#)).
- **Spill and staffing rules.** OCP requires spill-absorption capacity for "a catastrophic spill from the full contents of the largest tank", at least one person trained in spill management, and operator training "on the properties of each dielectric liquid in use" ([Rev 1.0](#)).
- **Hardware changes.** OCP's design guide for immersion-cooled IT flags thermal pastes that "may dissolve" (Indium foil is the usual replacement), PSUs that need modification "to allow long-term operation without fans", and optical connectors where fluid replacing the air gap causes "signal

reflection loss". Mitigations: air-gap-free or sealed transceivers, or port extenders ([OCP Design Guidelines for Immersion-Cooled IT Equipment, Rev 1.01](#)).

- **Fluid hygiene.** Rev 2.10 says the fluid supplier or solution provider "shall deliver the fluid hygiene/maintenance best practices" ([Rev 2.10](#)), and the warranty guideline notes that "some fluid vendors are starting to offer fluid quality monitoring service" ([OCP Warranty Guidelines](#)).

Resilience and redundancy

Uptime's April 2026 analysis: "DLC closely couples the facility infrastructure with the IT", redundant CDUs and manifolds "need to be carefully engineered and operated", and "acceptable ride-through time typically requires UPS power for CDU pumps, at a minimum" ([Uptime, 2026-04-15](#)). Coolant volume is "much less than that of the air in the data hall", so there is little thermal buffer.

Immersion has more thermal mass in the bath. OCP notes that dielectric ride-through can delay overheating "for seconds or days, depending on the technology", and that quantifying it is still to be defined. Its "High Availability" class requires N+1 or 2N cooling (dual heat exchangers or CDUs), which may be supplemented by ride-through providing "at least 30 minutes of operation without cooling" ([Rev 2.10](#)). Test ride-through at real GPU load during acceptance.

OEM support and warranty

OCP's warranty workstream exists to address "why warranties are not consistently offered for immersion cooling solutions". It notes that "almost all warranties for IT equipment today assume air cooling... and often the only approved use case" ([OCP Warranty Guidelines, 2024-04-26](#)).

The standard to look for is an explicit rider. Intel, for example, offers a "Xeon Processor Single-Phase Immersion Warranty Rider" for 4th and 5th Gen Xeon with Shell immersion fluids on Supermicro and Submer hardware ([Intel, 2025-05-13](#)). As of October 2026 we did not find a comparable public NVIDIA statement either supporting or excluding immersion for its data center GPUs, so we make no claim either way. Before buying, get answers in writing from the server OEM (and, through them, NVIDIA) to these questions:

1. Is immersion in *this named fluid* a supported configuration for *this exact SKU*, including GPU baseboard, NICs, optics and PSUs?
2. Which components must be removed or swapped (fans, thermal paste, labels, transceivers), and who does it?
3. What fluid temperature range and fluid-quality limits keep the warranty valid, and who monitors them?
4. How are RMA units cleaned and shipped, and who pays?

If the answer to question 1 is not an explicit yes on paper, plan for cold plates.

Facility water temperatures

Both architectures end at a heat exchanger. ASHRAE's fifth edition renamed the liquid-cooling classes by their upper facility-water limit: W17, W27, W32, W40, W45 and W+ (above 45 °C) ([Upsite, 2024-07-31](#)). OCP Rev 2.10 asks immersion vendors to report density at W-class temperatures and requires a W32 footprint figure ([Rev 2.10](#)). The GIGABYTE tank above loses 25% of its rating between 25 °C and 35 °C water.

On the cold-plate side, NVIDIA said at CES in January 2026 that Vera Rubin can be cooled with 45 °C water "with no chillers necessary", and that the Vera Rubin NVL72 rack will be 100% liquid cooled ([DCD, 2026-01-07](#)). Treat it as a vendor claim until the final rack specification is published. Warmer supply water reduces reliance on chillers for either architecture. For the water-consumption side see [how much water data centers use](#).

Is immersion production-ready at 20 MW?

Single-phase immersion is a commercial product, but at 20 MW of GPU load the evidence points away from it as the default:

- The largest share of new liquid-cooled capacity is single-phase DLC, and immersion is in "selective adoption" ([Dell'Oro, 2026](#)).
- NVIDIA's rack-scale systems are cold-plate products ([HPE QuickSpecs](#)).
- Warranty coverage for immersion is "not consistently offered" ([OCP](#)).
- Two-phase immersion lost 3M as a fluid supplier ([Uptime Intelligence](#)).

Immersion at that scale works when one party owns server design, fluid, tank and warranty, as in the Intel/Shell/Supermicro/Submer model. It is a poor fit for NVL72-class hardware whose warranty is your main asset protection.

Decision matrix

Situation	Lean towards	Why
GB200/GB300 NVL72 or successor racks	Single-phase cold plate	Vendor's own liquid-cooled design; OEM-built and warranted
HGX 8-GPU servers, new build	Cold plate (liquid-cooled SKU)	Liquid-cooled HGX SKUs exist, e.g. Cisco/Supermicro's 4U HGX B300 (Cisco datasheet); immersion needs written SKU+fluid approval
Mixed fleet in a legacy hall	Cold plate for GPUs, air for the rest	Keeps one air-side design; no tank retrofit
Certified immersion stack with written warranty rider	Single-phase immersion	Warranty and fluid risk sit with the vendor
Little ceiling height, no lifting gear, small ops team	Cold plate	Top-access servicing and fluid handling add work

Situation	Lean towards	Why
Long ride-through needed, hardware immersion-qualified	Single-phase immersion	Bath thermal mass; test it
Any fluorinated fluid	Defer, or get supply and PFAS commitments	3M exit; EU restriction pending
Hot facility water (W40/W45) is a design goal	Either; compare at equal W-class	Check derated tank kW and rack supply limits

For chiller plant and containment, see [GPU data center cooling design guide](#).

FAQ

Does immersion void an NVIDIA GPU warranty? We found no public NVIDIA statement either way as of October 2026. Get the OEM's answer in writing for the exact SKU and fluid.

Is two-phase immersion still an option? It exists, but mainstream interest is "nascent" after 3M's PFAS exit ([Uptime Intelligence](#)), and PFAS rules are still being finalised in the EU.

Do immersion tanks need a reinforced floor? Check them against the slab like any rack. In the example above a filled tank (before servers) loads about 9.9 kN/m², versus about 22.5 kN/m² for a loaded GB200 NVL72; the bigger differences are total fluid mass and spill containment.

For an independent review of a cooling architecture before you commit, [GPU Smith can help](#).

Sources

- [NVIDIA GB200 NVL72](#) — NVIDIA, accessed 2026-10-28
- [NVIDIA GB300 NVL72](#) — NVIDIA, accessed 2026-10-28
- [NVIDIA GB200 NVL72 by HPE QuickSpecs](#) — Hewlett Packard Enterprise, version 6, 2026-09-08, accessed 2026-10-28
- [Data Center Liquid Cooling Market to Approach \\$7 Billion by 2029](#) — Dell'Oro Group, 2026-01-08, accessed 2026-10-28
- [Direct liquid cooling: pressure is rising but constraints remain](#) — Uptime Institute, 2022-07-06, accessed 2026-10-28
- [Liquid cooling will not outgrow its high-density niche](#) — Uptime Institute, 2026-04-15, accessed 2026-10-28
- [Immersion fluids hold promise, but fire risk a concern](#) — Uptime Institute Intelligence, 2024-06-25 (public summary), accessed 2026-10-28
- [Immersion Requirements Rev 2.10](#) — Open Compute Project, 2023-08-18, accessed 2026-10-28
- [Immersion Requirements Rev 2.0](#) — Open Compute Project, 2022-01-21, accessed 2026-10-28
- [Immersion Requirements Document Rev 1.0](#) — Open Compute Project, 2019-05-20, accessed 2026-10-28
- [OCP Immersion project page](#) — Open Compute Project, accessed 2026-10-28

- [Updated immersion cooling requirements from OCP pave the way for increased adoption](#) — Vertiv, 2022-08-24, accessed 2026-10-28
- [Base Specification for Immersion Fluids Rev 1.0](#) — Open Compute Project, 2022-12-01, accessed 2026-10-28
- [Server Component Immersion Material Compatibility Testing v1.5](#) — Open Compute Project, 2026-06-05, accessed 2026-10-28
- [Warranty Guidelines for Immersion Cooled Technology Components, Equipment and Systems Rev 1.0](#) — Open Compute Project, 2024-04-26, accessed 2026-10-28
- [Design Guidelines for Immersion-Cooled IT Equipment Rev 1.01](#) — Open Compute Project, 2020-12-07, accessed 2026-10-28
- [Guidelines for Using Dielectric Heat Transfer Fluids in Two-Phase Cold Plate-Based Liquid-Cooled Racks Rev 1.1](#) — Open Compute Project, March 2026, accessed 2026-10-28
- [3M to Exit PFAS Manufacturing by the End of 2025](#) — 3M, 2022-12-20, accessed 2026-10-28
- [EU universal PFAS restriction process enters its decisive phase: SEAC draft opinion](#) — Linklaters, 2026-04-09, accessed 2026-10-28
- [ECHA Risk Assessment Committee Adopts Opinion on Universal PFAS Restriction](#) — ChemLinked, 2026-03-06, accessed 2026-10-28
- [Intel and Shell Advance Immersion Cooling in Xeon-Based Data Centers](#) — Intel, 2025-05-13, accessed 2026-10-28
- [AI Server for NVIDIA HGX B300, Liquid Cooled \(data sheet\)](#) — Cisco, accessed 2026-10-28
- [GIGABYTE 25U EIA Tank A1P0-EB0](#) — GIGABYTE, accessed 2026-10-28
- [Major changes to ASHRAE's fifth edition of thermal guidelines, part 3: liquid cooling chapter updates](#) — Upsite Technologies, 2024-07-31, accessed 2026-10-28
- [Vera Rubin hot water cooling reveal triggers HVAC share drop](#) — DatacenterDynamics, 2026-01-07, accessed 2026-10-28

Source links

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

1. <https://www.nvidia.com/en-us/data-center/gb200-nvl72>
2. <https://www.hpe.com/us/en/collaterals/collateral.a50009224enw.html>
3. <https://www.nvidia.com/en-us/data-center/gb300-nvl72>
4. <https://www.delloro.com/news/data-center-liquid-cooling-market-to-approach-7-billion-by-2029-as-ai-deployments-accelerate>
5. <https://journal.uptimeinstitute.com/direct-liquid-cooling-pressure-is-rising-but-constraints-remain>
6. <https://www.opencompute.org/documents/ocp-ac-immersion-requirements-rev-2-1-00-pdf>
7. <https://www.opencompute.org/documents/ocp-ac-immersion-requirements-specification-1-00-pdf>
8. <https://www.vertiv.com/en-us/about/news-and-insights/articles/blog-posts/updated-immersion-cooling-requirements-from-ocp-pave-the-way-for-increased-adoption>
9. <https://www.opencompute.org/community/immersion>
10. <https://www.opencompute.org/documents/ocp-ac-immersion-requirements-rev-2-v1-00-pdf>
11. <https://www.opencompute.org/documents/ocp-base-specification-for-immersion-fluids-20221201-pdf>

12. <https://www.opencompute.org/documents/2024-04-26-ocp-warranty-guidelines-for-immersion-cooled-technology-components-v1-00-pdf-1>
13. <https://www.opencompute.org/documents/ocp-component-compatibility-testing-guidelines-v1-5-final-june-5-2026-pdf>
14. <https://journal.uptimeinstitute.com/liquid-cooling-will-not-outgrow-its-high-density-niche>
15. <https://intelligence.uptimeinstitute.com/resource/immersion-fluids-hold-promise-fire-risk-concern>
16. <https://news.3m.com/2022-12-20-3M-to-Exit-PFAS-Manufacturing-by-the-End-of-2025>
17. <https://chemical.chemlinked.com/news/chemical-news/echa-risk-assessment-committee-adopts-opinion-on-universal-pfas-restriction>
18. <https://sustainablefutures.linklaters.com/post/102mp62/eu-universal-pfas-restriction-process-enters-its-decisive-phase-seac-draft-opin>
19. <https://www.opencompute.org/documents/guidelines-for-using-dielectric-heat-transfer-fluids-in-two-phase-cold-plate-based-liquid-cooled-racks-final-pdf>
20. <https://www.gigabyte.com/Enterprise/Accessory/A1P0-EB0-rev-200>
21. <https://www.opencompute.org/documents/design-guidelines-for-immersion-cooled-it-equipment-revision-1-01-pdf>
22. <https://www.intel.com/content/www/us/en/newsroom/news/data-center/intel-shell-advance-immersion-cooling-xeon-based-data-centers.html>
23. <https://www.upsite.com/blog/major-changes-to-ashraes-fifth-edition-of-thermal-guidelines-part-3-liquid-cooling-chapter-updates>
24. <https://www.datacenterdynamics.com/en/news/vera-rubin-hot-water-cooling-reveal-triggers-hvac-share-drop>
25. <https://www.cisco.com/c/dam/en/us/solutions/collateral/global-partners/supermicro/nvidia-hgx-b300-liquid-cooled-ds.pdf>

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.