# AI Data Center Rack Profiles

## A field guide to how AI racks are built, powered, and interconnected — by role

*Explainer · June 4, 2026*

*A category overview of how AI data center racks are organized by role, what drives their power consumption, and how interconnect and network-fabric generations have advanced year over year.*

| | |
|---|---|
| **TOPIC** | AI data center rack architecture and power density |
| **SCOPE** | Compute-heavy, balanced, storage-heavy, networking, and liquid-cooled profiles |
| **HORIZON** | 2020–2022 baseline through 2027–2028 vendor design targets |


**ABSTRACT**

AI data center racks are best understood as role-based profiles rather than a single standard form factor: compute-heavy, balanced, storage-heavy, networking, and a liquid-cooled cooling treatment most often applied to the compute-heavy role. Their defining characteristic is rapidly escalating per-rack power density, which has jumped roughly 5–10x from a pre-ChatGPT norm near 10–12 kW toward 120+ kW today, with vendor design targets reaching 600 kW–1 MW by 2027–2028. The "throughput by year" dimension is answerable only as interconnect and network-fabric generation roadmaps — NVLink, PCIe, and 400G/800G/1.6T Ethernet — not as a clean per-rack application-throughput series, because those figures are per-GPU or per-link, not rack aggregates. The practical implications: liquid cooling is now effectively mandatory above ~20 kW/rack, AMD competes directly with NVIDIA at the accelerator level, and the fabric layer is mid-transition between InfiniBand and Ethernet/RoCE.



*Keywords:* AI racks; power density; rack profiles; liquid cooling; NVLink; Ethernet fabric; DGX SuperPod; interconnect roadmap


## 1. Snapshot

This is a category explainer, not a company readout, so the orientation is about the object itself. An AI data center rack is the physical cabinet that packages the servers, storage, networking, and accelerators that run AI training and inference. The useful way to organize the space is by **role-based profile**: compute-heavy, balanced, storage-heavy, networking, and liquid-cooled. Treat that five-way grouping as a synthesized analytical model, not a published industry standard. There is no authoritative taxonomy of "the most common AI rack profiles" and no source provides a market-share ranking of which profile is most frequently deployed, so any claim that one profile is "most common" is unsupported. Note one structural point up front: "liquid-cooled" is a cooling *method* applied to a profile (usually compute-heavy), not a workload role.

## 2. Thesis: Why This Matters Now

The single fact that organizes everything is power density. AI rack power density has jumped roughly 5–10x, from a pre-ChatGPT (November 2022) norm of about 10–12 kW per rack to a current AI range commonly cited at 40–110 kW, with some deployments exploring 200+ kW. Traditional, non-AI racks still run 5–15 kW. That escalation is what forces every other design decision in the cabinet.

The reason "now" matters is that the curve is steepening, not flattening. Vendor reference architectures already point to 120+ kW racks shipping, and design targets reach 600 kW–1 MW by 2027–2028. Two cautions belong here. First, the upper end of today's range is a range, not a point value — sources differ on whether figures are per-system, per-rack, or include networking and overhead. Second, the far-horizon numbers rest on design roadmaps and analyst commentary, not shipping products, and should be read as targets rather than firm forecasts.

## 3. The Core Idea in Plain English

Think of an AI rack the way you would think of a race car versus a delivery van built on the same chassis. Same outer dimensions, radically different internals tuned to the job: a compute-heavy rack is packed with accelerators and runs hot, a storage-heavy rack trades silicon for capacity, a networking rack is mostly switches and optics, and a balanced rack splits the difference.

The old world ran general-purpose racks at 5–15 kW with simple air cooling. The new world packs the same footprint with GPUs drawing hundreds of watts each, pushing rack power into the tens or hundreds of kilowatts. Once you cross roughly 20 kW per rack, air can no longer carry the heat away, which is why liquid cooling stops being optional.

## 4. The Technical Space

AI racks are built around four functional building blocks: high-performance servers, storage, networking infrastructure, and specialized accelerators. Across the category, power and cooling are the dominant pressure points. A practical complication compounds this: AI servers are physically deeper than conventional ones, which reduces the cabinet space left for power distribution units (PDUs — the strips that feed and meter power to the equipment in a rack), cabling, and cooling hardware. You are fitting more heat-generating silicon into a tighter mechanical envelope.

"Good" in this space is measured on the dimensions that actually bind:

- **Power delivery.** Can the rack feed and distribute 40–110 kW (and rising) without starving the PDU and cabling budget the deep servers already squeeze.
- **Heat removal.** Above ~20 kW/rack, direct-to-chip liquid cooling (which circulates coolant directly to the chip rather than relying on airflow through the chassis) is the recommended approach; most AI training deployments now exceed that threshold, making liquid cooling effectively mandatory for high-density compute.
- **Interconnect bandwidth.** Both intra-rack (GPU-to-GPU) and inter-rack (fabric) links must keep accelerators fed, since a starved GPU is wasted capital.
- **Density-per-footprint.** How much useful compute fits in the constrained mechanical envelope.

This is the yardstick against which any specific rack design is judged.

## 5. The Profiles and the Power Math

The cleanest way to make the profiles concrete is to build them from component power and from a single vendor's high-end reference series.

**Component building blocks.** You can reconstruct rack totals from the parts. GPU thermal design power (TDP — the heat a chip is designed to dissipate, a close proxy for its power draw) runs from A100 at 400 W, to H100 SXM5 at 700 W, to B200 at up to 1,200 W; AMD's Instinct MI300X sits at 750 W and competes directly with the H100, so the market is not NVIDIA-only. Server CPUs add 150–200 W each, 400G optics draw roughly 10–14 W per port, and 800G optics 20 W+ per port.

**The DGX SuperPod power-by-year series.** This is the cleanest quantitative "power by year" anchor available, but it is one vendor's high-end reference architecture (NVIDIA's DGX SuperPod), not a market-wide average. Typical enterprise and cloud deployments lag this curve by one to two generations.

| Generation | Year | Per-rack power (approx.) | Notes |
|---|---|---|---|
| DGX A100 | 2020–2022 | ~24–26 kW | 6.5 kW/system; A100 TDP 400 W |
| DGX H100 | 2022–2023 | ~40–44 kW | 10.2 kW/system; H100 SXM5 TDP 700 W |
| GH200 | 2024 | ~72 kW | Grace Hopper; liquid-cooled |
| GB200 NVL72 | 2024–2025 | ~120–132 kW | B200 TDP up to 1,200 W; liquid-cooled |
| VR200 (projected) | 2026 | ~240 kW | Design target; not yet shipping |
| Rubin Ultra / Kyber (projected) | 2027–2028 | ~600 kW | ~576 GPUs/rack; vendor roadmap, speculative |

The 120-versus-132 kW spread for the GB200 NVL72 reflects genuine source variation on whether figures include networking overhead and PDU losses, so treat it as a range. The 2026 and 2027–2028 figures rest on unreleased design roadmaps and analyst commentary, not shipping hardware; read them as outer-bound design targets, with an average AI rack density near 50 kW by 2027 as the lower-confidence base case.

What is genuinely proprietary here is system integration and the NVLink interconnect, not the rack form factor itself. The cabinet, PDUs, and cooling manifolds are increasingly standardized and multi-vendor; the differentiated layer is the accelerator and its in-rack fabric.

## 6. Throughput by Year: Interconnect and Fabric Roadmaps

The honest answer to "throughput by year" is that there is no clean per-rack application-throughput series. What the evidence supports is two roadmaps of *generations*, and these are per-GPU or per-link figures that mix announcement and deployment dates, not rack aggregates.

**Intra-rack GPU-to-GPU interconnect.** NVLink (NVIDIA's proprietary high-bandwidth GPU-to-GPU link) defines the bandwidth between GPUs inside a rack or NVLink domain:

| Generation | Year | Bandwidth per GPU | Notes |
|---|---|---|---|
| NVLink 4.0 | 2022 (Hopper) | 900 GB/s | Intra-rack fabric for DGX H100 |
| NVLink 5.0 | 2024 (Blackwell) | 1.8 TB/s | GB200 NVL72 links 72 GPUs in one domain |
| NVLink 6.0 | ~2026 (Rubin) | ~3.6 TB/s | Announced roadmap; not yet shipping |

On the host-to-device side, PCIe (Peripheral Component Interconnect Express, the bus connecting CPUs to accelerators) 5.0 delivers 128 GB/s at x16; PCIe 6.0 (256 GB/s) began deploying from 2024; PCIe 7.0 hardware is expected around 2027.

**Inter-rack network fabric.** The 400G → 800G → 1.6T Ethernet path describes the dominant direction of travel, but it is not a clean single-track migration:

| Generation | Status |
|---|---|
| 400G Ethernet | Mainstream 2022–2023 |

| Generation | Status |
|---|---|
| 800G Ethernet | 802.3df ratified, accelerating 2024; production-standard for new builds 2025 |
| 1.6T Ethernet | 802.3dj finalizing ~2026; initial deployments 2027–2028 |

The caveat is essential. InfiniBand (a low-latency, high-bandwidth interconnect long dominant in HPC and AI clusters) remains active for latency-sensitive training, while Ethernet/RoCE (RDMA over Converged Ethernet, which delivers InfiniBand-style remote direct memory access over standard Ethernet) adoption is driven by operator economics as much as by raw speed. These are coexisting fabric choices, not a settled migration.

## 7. Where the Profiles Sit Relative to Each Other

The five profiles are best read as positions on a tradeoff surface rather than competing products.

**Compute-heavy is the power and cooling frontier.** It carries the densest accelerator load, drives the 40–110 kW (and rising) figures, and is the profile that pushes past the ~20 kW liquid-cooling threshold first. The liquid-cooled "profile" is, in practice, this role with direct-to-chip cooling applied; the GB200 NVL72 ships liquid-cooled by design because it cannot be air-cooled at 120–132 kW in a standard cabinet.

**Networking racks are bandwidth, not heat.** Their power comes from switch silicon and optics (10–14 W per 400G port, 20 W+ per 800G port), and their relevance is set by which fabric generation a build standardizes on.

**Storage-heavy and balanced fill the middle.** Storage-heavy trades silicon for capacity and runs cooler; balanced splits compute, storage, and networking for mixed workloads.

No durable moat attaches to any single profile, and no source ranks one as most common. Because the profiles sit in different power-stress regimes, disaggregated architectures — separating cooler storage and networking racks from the dense compute racks, potentially across different facility tiers — are a natural design response. The shared structural constraint is the deeper AI server chassis, which steals room from PDUs, cabling, and cooling, so density gains compound the power-delivery and heat-removal problems at once.

## 8. Risks and Open Questions

A few unknowns would materially change how you read this category. Each is specific enough to push on:

- **Will the 2027–2028 power figures (250 kW, 600 kW, 1 MW+) actually ship?** They originate in vendor design roadmaps — including NVIDIA "Rubin Ultra"/"Kyber"-class systems (~576 GPUs/rack) targeting ~600 kW around 2027 — and analyst commentary, not shipping hardware.
- **InfiniBand or Ethernet/RoCE for inter-rack fabric through 2026–2028?** One reading treats InfiniBand as historically dominant in AI clusters (cited near 80% share in 2023, from commercial material); another reports Ethernet/RoCEv2 reaching parity and hyperscalers such as Meta, Microsoft, and AWS converging on it. The market is transitioning, not settled.
- **What does "throughput" even denote for a rack?** Inter-rack Gbps, intra-rack GB/s per GPU, or computational FLOPS (floating-point operations per second) / tokens-per-second are three different answers, and no consistent year-by-year rack-level series exists in any single definition.
- **Is the five-profile taxonomy authoritative?** It is a synthesized model; treat the profile labels as organizing structure, not standard.

## 9. Bottom Line

Three takeaways. First, AI racks are role-based profiles, not a standard form factor, and per-rack power density — not application throughput — is the metric that actually defines them. Second, the category works because escalating density (10–12 kW pre-2022 toward 120+ kW today) forces liquid cooling above ~20 kW and reshapes every component budget in the cabinet. Third, the thing to watch is whether the 600 kW–1 MW design targets for 2027–2028 convert from vendor roadmaps into shipping hardware, and which fabric — InfiniBand or Ethernet/RoCE — carries the inter-rack load.

## 10. For the Nerds

The deeper tension is that "throughput by year" cannot be assembled into a clean rack-level series because the three candidate definitions live at different layers of the stack. NVLink and PCIe figures are per-GPU or per-link bandwidth; Ethernet and InfiniBand figures are per-port fabric speeds; FLOPS and tokens-per-second are application metrics. None aggregates cleanly to a single rack number without assumptions about topology, oversubscription, and parallelism strategy. A GB200 NVL72 quoting 1.8 TB/s of NVLink per GPU is describing the intra-rack scale-up domain, a different physical and economic regime than the 800G scale-out fabric stitching racks into a cluster. The practical consequence: planning around a specific GB/s figure for a given year means committing to a specific generation's deployment date, and that commitment carries more uncertainty the further out you project.

The open question underneath the power-density curve is thermodynamic, not just electrical. Direct-to-chip liquid cooling handles today's 120+ kW racks, but the 600 kW–1 MW design targets imply heat fluxes that may push toward immersion or two-phase cooling. Whether that transition arrives on the vendors' 2027 timeline is the bet that determines whether those power figures are forecasts or aspirations.

## WHERE SOURCES DISAGREE

**Whether InfiniBand or Ethernet/RoCE will dominate AI inter-rack fabric through 2026–2028**

One reading treats InfiniBand as the historically dominant AI-cluster fabric (cited at roughly 80% share in 2023) with Ethernet as the challenger. A competing set of sources reports Ethernet/RoCEv2 has reached parity on latency and bandwidth, that hyperscalers such as Meta, Microsoft and AWS are converging on RoCEv2, and that one commercial source claims Ethernet led back-end AI network deployments by mid-2025

*Unresolved: the share figures come from commercial/vendor material rather than primary research, and the sources agree the market is actively transitioning rather than settled*

**Whether the "400G → 800G → 1.6T Ethernet" timeline describes a clean single-track migration**

One reading presents it as an orderly generational roadmap with dates tied to IEEE ratification milestones. Another account shows InfiniBand remaining active for low-latency training while Ethernet adoption is driven as much by operator economics as by raw throughput, implying parallel tracks rather than one migration path

*Unresolved: the sources describe coexisting fabric choices and do not reconcile the roadmap dates with real-world adoption mix*

**What "throughput" means in this context**

One reading interprets it as inter-rack network bandwidth (Gbps/rack), which the sources document via Ethernet/InfiniBand generations. Another reading points to intra-rack interconnect bandwidth (NVLink/PCIe GB/s per GPU), and a third to computational throughput (FLOPS, tokens/sec), which standard accelerator specifications report but the sources do not assemble into a year-by-year rack-level series

*Unresolved: the available evidence supplies interconnect and fabric generation timelines but no consistent quantitative per-rack throughput series in any single definition*

**Whether there is any authoritative taxonomy of "the most common AI rack profiles"**

One reading offers the five-role grouping (compute-heavy, balanced, storage-heavy, networking, liquid-cooled) as a usable organizing structure. Another reading notes this is a synthesized model and that the networking-focused sources address fabric choices rather than any standardized rack taxonomy

*Unresolved: no source provides a published, industry-standard profile list or a ranking of which profile is most frequently deployed*

**Whether the 2027–2028 power figures (250 kW, 600 kW, 1 MW+) are credible forecasts**

One reading cites them as concrete vendor design targets and analyst projections for the Rubin/Kyber generation. Another reading treats them as speculative, since they rest on unreleased design roadmaps and commentary rather than shipping hardware

*Unresolved: the figures cannot be checked against shipping-product data and originate in vendor roadmaps and analyst forecasts*