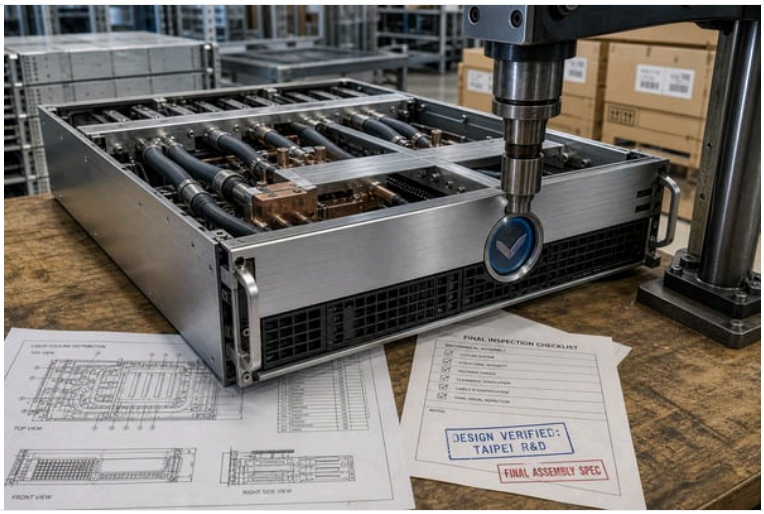
Which Server OEMs Design Their Own NVIDIA Servers?

Published July 15, 2026 42 min read



Executive Summary

No single company designs "the" NVIDIA server. Instead, three distinct layers of the supply chain share the work: **NVIDIA** itself designs the GPU silicon and the **HGX** baseboard reference design (Source: nvidia.com), a small group of branded **original equipment manufacturers (OEMs)**, chiefly **Dell Technologies**, **Supermicro**, **Hewlett Packard Enterprise (HPE)**, and **Lenovo**, design the chassis, cooling, power delivery, and firmware around that baseboard and sell the finished system under warranty, and a set of Taiwan-based **original design manufacturers (ODMs)**, including **Foxconn** (through its Ingrasys subsidiary), **Quanta Cloud Technology (QCT)**, **Wistron**, **Wiwynn**, and **Inventec**, do the actual rack-scale engineering and physical assembly, often for hyperscalers buying direct. As of July 2026, IDC's Worldwide Quarterly Server Tracker put total server market revenue at **$112.4 billion** for the third quarter of 2025 alone, up 61.1% year over year (Source: my.idc.com), with "ODM Direct" sales, meaning hyperscalers bypassing branded OEMs entirely, capturing **59.4%** of that revenue versus Dell's **8.3%**, Supermicro's **4.0%**, and HPE's **3.0%** (Source: my.idc.com).

Dell is the clearest "designs its own NVIDIA servers" story among the branded OEMs. Its **PowerEdge XE9680** and newer **XE9712** and **XE8712** platforms build custom chassis, liquid-cooling, and iDRAC firmware around NVIDIA's HGX baseboard (Source: delltechnologies.com), and Dell reported a record **$113.5 billion** in full fiscal 2026 revenue with **$64 billion** in AI-optimized server orders closed and a **$43 billion** AI backlog entering fiscal 2027 (Source: investors.delltechnologies.com). Supermicro is the most vertically integrated of the branded OEMs, doing more manufacturing in-house through its San Jose, Taiwan, and Netherlands facilities than rivals that lean on contract assembly, and it has repeatedly claimed "first-to-market" status on new NVIDIA platforms such as the liquid-cooled **GB200 NVL72** rack (Source: ir.supermicro.com); it posted fiscal 2025 revenue of **$22.0 billion**, up from **$15.0 billion** the prior year (Source: ir.supermicro.com). But Supermicro has also been the subject of an SEC accounting enforcement action that resulted in a **$17.5 million** penalty in 2020 (Source: sec.gov) and a March 2026 Department of Justice indictment of a Supermicro co-founder over an alleged $2.5 billion scheme to divert NVIDIA-equipped servers to China in violation of export controls (Source: theregister.com), which the report treats as material context, not settled fact.

Below the OEM layer sit the ODMs, which frequently do more of the actual engineering than buyers realize. NVIDIA's own MGX modular reference architecture explicitly names Foxconn's Ingrasys unit, Quanta's QCT, Wistron, and Wiwynn as design and manufacturing partners alongside Dell, HPE, Lenovo, and Supermicro (Source: nvidia.com), and research on Taiwan's electronics sector estimates Taiwanese firms handle roughly 90% of global AI server assembly, though that figure measures production volume, not profit share (Source: penchan.co). NVIDIA does not modify the HGX GPU baseboard for any partner; GPU count, NVSwitch layout, and NVLink wiring are fixed by NVIDIA, while OEMs and ODMs compete on chassis, cooling, CPU choice, and firmware (Source: amcompute.com). Real-world deployments illustrate the churn in this market: xAI's 100,000-GPU Colossus cluster was built with Supermicro hardware (Source: supermicro.com) before Elon Musk's companies redirected new orders to Dell, benefiting Dell's ODM partners Wistron and Inventec (Source: networkworld.com), while CoreWeave, Meta, and Tesla have each taken different paths between buying branded OEM hardware and co-designing systems directly with NVIDIA and ODM partners.

At the rack-scale frontier, NVIDIA's own GB200 NVL72 reference design, which connects 36 Grace CPUs and 72 Blackwell GPUs into a single 130 TB/s NVLink domain (Source: nvidia.com), is manufactured under license by both branded OEMs and ODMs simultaneously, most visibly Supermicro on the OEM side and Foxconn's Ingrasys on the ODM side, blurring the line between "who designed it" and "who builds it" even further (Source: ingrasys.com). For buyers researching this market, the practical takeaway is that brand recognition and design ownership are only loosely correlated: the companies with the most name recognition (Dell, Supermicro, HPE, Lenovo) collectively hold a minority of total server market revenue, while the companies doing a majority of the physical and much of the rack-level engineering work operate largely behind non-disclosure agreements with hyperscaler customers.

Introduction and Background

The question "which server OEMs design their own NVIDIA servers" sounds simple but hides a genuinely three-layered supply chain that most buyers, and even much of the trade press, conflate into a single category. At the top, **NVIDIA Corporation** designs the GPU die, fabricates nothing itself, and defines two reference platforms that server makers build around: **HGX**, a fixed baseboard holding eight SXM-format GPUs wired together with NVLink and NVSwitch, and **MGX**, a more flexible modular rack and chassis specification for mixed CPU and GPU configurations (Source: nvidia.com). With MGX, "OEM and ODM partners can build tailored solutions for different use cases while saving development resources and reducing time to market," according to NVIDIA's own product page (Source: nvidia.com). NVIDIA also sells its own turnkey systems under the **DGX** brand, built on the identical HGX baseboard but with a CPU, memory, storage, and support stack fixed entirely by NVIDIA rather than left open to a partner; NVIDIA's own DGX B200 documentation specifies eight NVIDIA B200 GPUs "that provide 1,440 GB total GPU memory" alongside two "Intel Xeon 8570 PCIe Gen5 CPUs with 56 cores each" and six 3.3 kW power supplies configured for redundancy (Source: docs.nvidia.com) (Source: docs.nvidia.com).

Below that reference layer sit the companies most buyers recognize as "server OEMs": Dell Technologies, Supermicro, HPE, Lenovo, Cisco, Gigabyte, and ASUS.NVIDIA's own certified-partner program lists dozens of companies across servers, workstations, and edge systems, including Advantech, Altos, ASRock Rack, Atos, BOXX, Fujitsu, H3C, Hitachi, Lanner, Leadtek, Nettrix, QCT, xFusion, Wistron, and Pegatron, with **more than 400 NVIDIA-Certified Systems** available across that ecosystem as of the current program listing (Source: nvidia.com). A server only reaches that certified list after passing NVIDIA's own qualification process, which NVIDIA describes as testing whether "the accelerated hardware is fully functional in that server design" across thermal, mechanical, power, and signal integrity criteria before certification review even begins (Source: nvidia.com). When the certification program launched its first wave in the A100 generation, NVIDIA named **Dell Technologies**, **GIGABYTE**, **Hewlett Packard Enterprise**, **Inspur Electronic Information**, and **Supermicro** as the initial shipping partners, with 14 certified servers from six systems makers among roughly 70 systems from at least 11 makers already engaged in the program (Source: blogs.nvidia.com) (Source: blogs.nvidia.com).

Underneath the recognizable OEM brands sits the layer that does much of the actual engineering: the ODMs. **Original design manufacturer (ODM)** is a specific supply-chain term meaning the contractor both designs and builds a product that ships under someone else's brand, which differs from a classic **original equipment manufacturer (OEM)** relationship in which the brand owns the design and the contractor only manufactures to spec (Source: penchan.co). As Penchan's supply-chain research puts it, "the brand, such as a cloud provider or computer company, defines the requirements and places the order," while the ODM carries the product through design, validation, and mass production (Source: penchan.co). In AI servers, this distinction matters enormously because the ODMs, primarily Taiwan-based firms such as Foxconn (via its Ingrasys unit), Quanta (via QCT), Wistron, Wiwynn, and Inventec, frequently perform more rack-level engineering than the branded OEM whose logo appears on the chassis. NVIDIA is itself a fabless semiconductor company, meaning it designs chips but outsources fabrication to TSMC and outsources full system-and-rack assembly to these Taiwan-based partners (Source: penchan.co (Source: penchan.co).

The stakes of the OEM relationship are high enough that NVIDIA's own leadership has publicly courted specific partners. Independent analyst coverage from SemiAnalysis recounts that at NVIDIA's GTC 2024 conference, CEO Jensen Huang visited Dell's booth and repeatedly praised the company, then called out Michael Dell by name from the keynote stage, a level of public favoritism SemiAnalysis attributes to Dell's rapid catch-up in shipping 8-way HGX platforms after having reportedly been "the lowest priority partner for Nvidia" during the earlier A100 generation, when Dell had not yet shipped an 8xA100 HGX server at all (Source: [newsletter.semianalysis.com](https://newsletter.semianalysis.com/p/how-dell-is-beating-supermicro#:~:text=OEM%E2%80%99s%20(Dell%2C%20Supermicro%2C%20HPE%2C%20Lenovo%2C%20etc%20and%20ODM%E2%80%99s%20(Quanta%2C%20FII%2C%20Inventec%2C%20Wistron%2C%20Wiywynn)
That single anecdote illustrates why the OEM-versus-ODM question is not purely technical: NVIDIA has commercial incentives to cultivate multiple credible branded partners so that no single OEM, and no single point of supply-chain failure, controls its go-to-market for enterprise and sovereign AI customers. This report walks through each layer in turn, quantifies market share and financial scale with data current as of July 2026, and closes with named case studies showing how real buyers, from xAI to Meta to Tesla, have navigated the OEM-versus-ODM decision in practice.

Dell Technologies

Capabilities

Dell's flagship AI platform is the **PowerEdge XE9680**, a 6U rack server that Dell's own specification sheet describes as its first eight-way GPU platform, supporting "8 NVIDIA HGX H100 80GB 700W SXM5 GPUs, fully interconnected with NVIDIA NVLink technology," or alternatively H200 or H20 SXM5 GPUs, AMD Instinct MI300X, or Intel Gaudi 3 accelerator options (Source: delltechnologies.com), alongside "up to 64-core 5th generation Intel Xeon processors" and up to 4 TB of DDR5 memory (Source: delltechnologies.com). Dell has since extended the line with the liquid-cooled **XE9712** for NVIDIA's B300 HGX platform and the **XE8712** supporting GB200 NVL4 rack-scale configurations. Notably, Dell explicitly sells an "OEM-ready" version of its own hardware to other resellers, stating that "from bezel to BIOS to packaging, your servers can look and feel as if they were designed and built by you" (Source: delltechnologies.com), a reminder that even branded OEMs sometimes act as white-label suppliers one layer further down the chain. Dell's engineering ownership covers the chassis, thermal design (including Multi-Vector Cooling and direct liquid cooling options), power delivery, and the **iDRAC9** baseboard management controller and OpenManage software stack, all built around NVIDIA's fixed HGX baseboard.

Adoption

Dell's AI server business accelerated sharply through fiscal 2026. The company's **Infrastructure Solutions Group (ISG)** reported record full-year revenue of **$60.8 billion**, up 40% year over year, with AI-optimized servers alone contributing **$24.68 billion** for the fiscal year, up 166%, and a fourth-quarter run rate of **$9.0 billion**, up 342% year over year (Source: investors.delltechnologies.com). ISG operating income reached **$7.1 billion** for the fiscal year, up 27% year over year (Source: investors.delltechnologies.com) and represented 72% of Dell's total reportable segment operating income, a sharp increase from 65% the prior year, underscoring how central AI infrastructure has become to Dell's overall profitability. Overall company revenue hit a record **$113.5 billion** for fiscal 2026, up 19% year over year, while Dell's Client Solutions Group, the traditional PC and laptop business, grew a comparatively modest 5% to **$51.0 billion** (Source: investors.delltechnologies.com) (Source: investors.delltechnologies.com). Chief operating officer Jeff Clarke stated the company "closed more than $64 billion in AI-optimized server orders" during the year and is "entering FY27 with record backlog of $43 billion" (Source: investors.delltechnologies.com) (Source: investors.delltechnologies.com). Named customers using Dell's NVIDIA-based platforms include **CoreWeave**, which Dell announced in December 2023 would use "PowerEdge XE9860 servers with NVIDIA H100 Tensor Core GPUs" as key infrastructure for its cloud, describing CoreWeave as "a specialized cloud provider for large-scale NVIDIA GPU-accelerated workloads" (Source: dell.com), with Dell describing the underlying PowerEdge XE9680 platform as "built for extreme acceleration of AI, ML and deep learning training" (Source: dell.com). Independent analysis from SemiAnalysis notes that Dell has more recently "gained sockets at CoreWeave, Tesla, and x.ai," three of the largest GPU buyers that use the OEM supply chain rather than buying direct from ODMs (Source: newsletter.semianalysis.com). Per IDC's Q3 2025 tracker, Dell led the branded OEM market with an **8.3%** revenue share, the highest of any named vendor (Source: my.idc.com).

Strengths and Limitations

- **Enterprise integration**: Dell's ProSupport Plus and OpenManage Enterprise stack give large enterprise buyers a single point of contact for firmware, warranty, and lifecycle management across a fleet.
- **Backlog visibility**: A **$43 billion** disclosed AI backlog gives buyers a clear read on Dell's forward demand and capacity commitments, as reported alongside the company's other AI-server order metrics (Source: investors.delltechnologies.com).
- **Late HPC entry**: SemiAnalysis has documented that Dell was slow to ship 8-way HGX A100 and H100 platforms, initially ceding early large-scale AI training deployments (including infrastructure used for GPT-3.5 and GPT-4) to rivals (Source: [newsletter.semianalysis.com](https://newsletter.semianalysis.com/p/how-dell-is-beating-supermicro#:~:text=OEM%E2%80%99s%20(Dell%2C%20Supermicro%2C%20HPE%2C%20Lenovo%2C%20etc%20and%20ODM%E2%80%99s%20(Quanta%2C%20FII%2C%20Inventec%2C%20Wistron%2C%20Wiyw
- **Premium pricing**: Third-party tier comparisons position Dell's list pricing as the baseline against which leaner competitors like Supermicro undercut on comparable configurations, reflecting the cost of Dell's larger services and support organization.

Supermicro

Capabilities

Super Micro Computer, Inc., commonly known as Supermicro, differentiates itself from Dell and HPE primarily through vertical integration and speed to market. The company states plainly that "our products are designed and manufactured in-house (in the US, Taiwan, and the Netherlands)," leveraging global operations for scale rather than relying purely on third-party ODM assembly (Source: ir.supermicro.com (Source: ir.supermicro.com). Its **Server Building Block Solutions** approach lets customers mix and match chassis, power supplies, and cooling around a common architecture, and by June 2025 Supermicro was marketing "more than 30 solutions designed for air or liquid-cooled NVIDIA HGX B200, liquid-cooled NVIDIA GB200 NVL72, and NVIDIA RTX PRO 6000 Blackwell Server Edition" (Source: ir.supermicro.com). Its **DLC-2 direct-liquid-cooling technology** and in-rack coolant distribution units are, in the company's own words, "capable of removing 250kW of heat per rack," a figure Supermicro cites as enabling denser deployments within existing facility power envelopes (Source: ir.supermicro.com). Trade coverage of Supermicro's manufacturing model additionally describes a related-party arrangement in Taiwan, under which affiliated companies including Ablecom handle a portion of subassembly work for Supermicro; this arrangement has drawn scrutiny in prior short-seller reports and is a factor buyers evaluating supply-chain transparency may wish to weigh alongside Supermicro's cost advantages (Source: amcompute.com).

Adoption

Supermicro's fiscal 2025 (ended June 30, 2025) net sales reached **$22.0 billion**, up from **$15.0 billion** the prior fiscal year, a 47% increase, with net income of roughly **$1.0 billion** (Source: ir.supermicro.com). The company guided fiscal 2026 net sales of at least **$33.0 billion**. More recent trade press coverage from The Register reported Supermicro "recently revealed $12.7 billion revenue in its most recent quarterly earnings" as of early 2026, evidence the company has continued growing in absolute terms even amid its compliance controversies (Source: theregister.com). Per IDC, Supermicro held the second-largest branded OEM revenue share in Q3 2025 at **4.0%**, though that was down from 7.4% a year earlier as competition intensified (Source: my.idc.com). Supermicro's most visible deployment is **xAI's Colossus** cluster in Memphis, Tennessee, which Supermicro's own case study describes as connecting "100,000 NVIDIA Hopper Tensor Core GPUs" using NVIDIA's Spectrum-X Ethernet networking platform in a liquid-cooled SuperCluster architecture (Source: supermicro.com).

Strengths and Limitations

- **Speed and flexibility**: Supermicro has repeatedly been first among branded OEMs to ship systems on new NVIDIA platforms, including the GB200 NVL72 rack.
- **Price competitiveness**: Because Supermicro avoids some of the ODM margin layer that Dell and HPE pay, and runs a leaner sales and support organization, comparable configurations often carry lower list prices than Dell or HPE equivalents.
- **Governance and compliance history**: Supermicro was fined **$17,500,000** by the SEC in August 2020 after the agency found the company "violated federal securities laws by engaging in improper accounting," prematurely recognizing revenue and understating expenses from fiscal 2015 through 2017, and as a result "filed with the Commission materially misstated financial statements in its annual, quarterly and current reports during that period" (Source: sec.gov) (Source: sec.gov) (Source: sec.gov). More recently, in March 2026 the Department of Justice charged Supermicro co-founder Yih-Shyan "Wally" Liaw and two associates with conspiring to divert servers containing NVIDIA GPUs worth **$2.5 billion** to Chinese customers in violation of US export controls, allegedly using falsified documents and staged "dummy" servers to pass inspections (Source: theregister.com) (Source: theregister.com); Supermicro itself was not named as a defendant and stated the conduct violated its own compliance policies.

- **Service model limits**: Community discussion on enterprise IT forums generally describes Supermicro hardware as good value but with a smaller support and parts-availability footprint than Dell or HPE, a tradeoff examined further in the Case Studies section.

HPE

Capabilities

Hewlett Packard Enterprise split from HP Inc. on November 1, 2015, taking the server, storage, networking, and services businesses, and later strengthened its high-performance computing pedigree with a **$1.3 billion** acquisition of supercomputer maker Cray in 2019. Its flagship NVIDIA-based AI platform is the **HPE Cray XD670**, a 5U chassis supporting "8x NVIDIA H200 SXM 700W TDP GPUs with 141 GB HBM each," paired with "2x 5th Gen Intel Xeon Scalable processors CPUs, up to 400W TDP" and up to 32 DDR5 DIMMs per socket, available with air cooling or a direct liquid cooling option (Source: hpe.com) (Source: hpe.com). HPE markets the XD670's benchmark credentials directly, stating it "delivered six #1 results in the MLPerf Inference v5.1 tests including in computer vision and LLMs chat Q&A and text generation" (Source: hpe.com).

Adoption

HPE reported fourth-quarter fiscal 2025 revenue of **$9.7 billion**, up 14% year over year, though its Server segment specifically was down 5% to **$4.5 billion** in the same quarter, reflecting a broader portfolio mix shift toward Networking following the Juniper Networks acquisition (Source: hpe.com). Within the same quarter, HPE's own reporting shows "Networking revenue was $2.8 billion, up 150% from the prior-year period" and Hybrid Cloud revenue of $1.4 billion, down 12%, while Financial Services contributed $889 million, roughly flat year over year, illustrating that AI-driven networking demand, not the server line specifically, is currently HPE's fastest-growing business (Source: hpe.com). HPE's consumption-based **GreenLake** model posted an annualized revenue run-rate of **$3.2 billion**, up 63% year over year (Source: hpe.com). By IDC's Q3 2025 count, HPE ranked fifth among branded server OEMs with a **3.0%** revenue share, behind Dell, Supermicro, IEIT Systems, and a statistically tied Lenovo (Source: my.idc.com).

Strengths and Limitations

- **HPC heritage**: HPE's Cray lineage gives it deep experience with the cooling, networking, and job-orchestration challenges of exascale-class systems, which increasingly resemble the demands of large AI training clusters.
- **Liquid cooling leadership**: HPE was among the earliest branded OEMs to standardize direct liquid cooling across its AI server line, ahead of the Blackwell generation's higher per-GPU thermal design power.
- **Consumption pricing flexibility**: GreenLake lets enterprises deploy on-premises NVIDIA hardware under cloud-like, pay-for-what-you-use economics rather than a straight capital purchase.
- **Slower server segment growth**: HPE's core Server revenue declined year over year even as the company's overall business grew, suggesting NVIDIA-based AI systems have not yet offset softness elsewhere in its traditional server book.

Lenovo and the Rest of the Certified OEM Field

Capabilities

Lenovo's server business traces to two IBM acquisitions: the ThinkPad PC line in 2005 for **$1.75 billion** and IBM's x86 server business, including System x, BladeCenter, and Flex System, in October 2014 for **$2.1 billion**, which brought roughly 7,500 IBM employees to Lenovo. Its current flagship NVIDIA platform is the **ThinkSystem SR675 V3**, which in its SXM5 GPU configuration supports NVIDIA HGX H100 or H200 4-GPU modules, alongside up to two AMD EPYC 9004 or 9005 series processors "scalable up to 128 cores per socket, 256 cores in total" and, with two processors installed, memory capacity Lenovo's own documentation lists at "Maximum: 3 TB" of DDR5 (Source: pubs.lenovo.com) (Source: pubs.lenovo.com). Lenovo's **Neptune** liquid-cooling technology, first developed for supercomputing customers, now extends across its Blackwell-class ThinkSystem SR680b platform. Lenovo's own support organization, Premier Support, provides "advanced technical support 24x7x365 in more than 100 markets" according to the company's own service description, a global reach few competitors below Dell and HPE can match (Source: techtoday.lenovo.com).

Beyond the "big four" of Dell, Supermicro, HPE, and Lenovo, **Cisco** entered the server market in 2009 with its Unified Computing System (UCS) line and remains an NVIDIA-certified server partner, while **Gigabyte** and **ASUS**, both Taiwan-founded PC and component makers, have built out dedicated GPU server divisions that participate in both the HGX certified-partner program and NVIDIA's MGX modular architecture. NVIDIA's certified-systems partner list separately names **Fujitsu**, **H3C**, **Hitachi Vantara**, **Atos**, **Nettrix**, and workstation-focused **BOXX** and **Z by HP** as smaller-volume participants in the same ecosystem (Source: nvidia.com).

Adoption

IDC's Q3 2025 tracker recorded a statistical tie between **IEIT Systems** (3.7%) and **Lenovo** (3.6%) for third place in worldwide server revenue share, both ahead of HPE's 3.0% and both trailing Dell and Supermicro (Source: my.idc.com). Lenovo's own segment reporting separately shows its Infrastructure Solutions Group, the division housing ThinkSystem servers, as a substantial and growing contributor to a total company revenue base measured in the tens of billions of dollars. This underscores a pattern the report returns to repeatedly: the branded, Western-recognizable OEM tier (Dell, Supermicro, HPE, Lenovo, Cisco, Gigabyte, ASUS) collectively controls a modest fraction of overall server revenue relative to ODM-direct sales to hyperscalers.

Strengths and Limitations

- **Global support footprint**: Lenovo's Premier Support covers more than 100 markets, one of the broadest geographic reaches of any server OEM, useful for multinational enterprises with distributed data centers (Source: techtoday.lenovo.com).
- **Second-tier scale**: None of the Cisco, Gigabyte, or ASUS server lines approach Dell or Supermicro in AI server revenue, meaning buyers may face narrower configuration catalogs and slower firmware update cadences.
- **China-market exposure**: IEIT Systems (formerly Inspur), Lenovo, and other Chinese-founded or China-headquartered vendors face a distinct set of export-control and buyer-perception considerations that Western-headquartered OEMs do not, particularly for government and defense-adjacent procurement.

The ODM Layer: Who Actually Builds the Box

Capabilities

The companies that physically design and assemble a meaningful share of the world's NVIDIA-based AI servers are not household names to enterprise buyers but are essential to understanding "who actually builds NVIDIA GPU servers." **Original design manufacturer (ODM)** describes a contractor that takes a product "through design, development, validation, mass production, and shipment under the customer's brand," distinct from a pure assembly relationship (Source: penchan.co). NVIDIA's own MGX documentation names its system-partner ecosystem explicitly, and the roster spans both OEMs and ODMs: Aetina, ASRock, ASUS, Cisco, Compal, Gigabyte, HPE, **Ingrasys** (Foxconn's AI-server subsidiary), **Inventec**, Lanner, Lenovo, MiTAC, MSI, Pegatron, **Quanta Cloud Technology**, Supermicro, **Wistron**, and **Wiwynn** are all listed as MGX system partners (Source: nvidia.com).

Foxconn (Hon Hai Technology Group) runs its AI server design work through **Ingrasys**, which markets itself around building the "world's first AI server lighthouse factory" and describes the GB200 NVL72 rack it manufactures as connecting "36 Grace CPUs and 72 Blackwell GPUs through fifth-generation NVLink" to function as a single coherent accelerator domain, with each compute tray featuring "2 GB200 Grace Blackwell Superchips" and power shelves individually rated to "provide a maximum power output of 33kW" (Source: ingrasys.com) (Source: ingrasys.com) (Source: ingrasys.com). **Quanta Cloud Technology (QCT)**, the enterprise arm of Quanta Computer, states plainly that it "designs, manufactures, integrates, and services cutting-edge offerings for 5G Telco/Edge, AI/HPC, Cloud, and Enterprise infrastructure" (Source: qct.io), and QCT's president Mike Yang has described the company's QuantaGrid line as "built upon QCT's expertise in server design and NVIDIA's proven MGX and HGX system architectures," with its liquid-cooled HGX B200 systems rated to "deliver up to 15X faster real-time inference performance, 12X lower cost, and 12X less energy" compared to prior-generation deployments (Source: qct.io) (Source: qct.io), an explicit ODM claim to genuine engineering ownership rather than mere assembly. **Wiwynn**, spun out of Wistron specifically to focus on cloud and data-center ODM work, showcased its GB300 NVL72-ready systems at GTC 2025 in partnership with Wistron, stating the two companies "are among the first in line with NVIDIA GB300 NVL72" readiness, a platform Wiwynn describes as carrying "over 20TB of HBM3e memory" (Source: wiwynn.com) (Source: wiwynn.com; Wiwynn's president William Lin described the work as requiring "comprehensive integration at the rack level" spanning GPUs, compute systems, cooling, and networking (Source: wiwynn.com.

Adoption

According to IDC's Q3 2025 tracker, "ODM Direct" sales, meaning hyperscalers and large cloud providers buying rack-scale systems straight from ODMs without a branded OEM in between, generated **$66.79 billion** of the quarter's **$112.4 billion** total server revenue, a **59.4%** share, up from 45.1% a year earlier (Source: my.idc.com). Analysis from Taiwan-focused research firm Penchan estimates Taiwanese firms account for roughly 90% of global AI server production, though it stresses that figure "is a production or assembly figure, not brand share, revenue share, or profit share" (Source: penchan.co). SemiAnalysis frames the same split more bluntly: "the market for Nvidia servers can generally be broken out into OEM's (Dell, Supermicro, HPE, Lenovo, etc) and ODM's (Quanta, FII, Inventec, Wistron, Wiywynn, ZT Systems, etc)," with hyperscalers preferring ODMs because they need a lower level of white-glove service (Source: [newsletter.semianalysis.com](https://newsletter.semianalysis.com/p/how-dell-is-beating-supermicro#:~:text=OEM%E2%80%99s%20(Dell%2C%20Supermicro%2C%20HPE%2C%20Lenovo%2C%20etc%20and%20ODM%E2%80%99s%20(Quanta%2C%20FII%2C%20Inventec%2C%20Wistron%2C%20Wiywynn)
Those margins are thin: SemiAnalysis notes ODMs typically capture only low single-digit percentage margins from building server hardware (Source: newsletter.semianalysis.com), a structural reality Penchan corroborates, describing "system-assembly margins" as "often in the single digits" because customers are concentrated among a handful of cloud giants with strong bargaining power (Source: penchan.co).

Strengths and Limitations

- **Cost efficiency at scale**: ODMs strip out the branded OEM's sales, marketing, and premium-support overhead, which is why hyperscalers with in-house operations teams gravitate toward buying direct.
- **Genuine design ownership**: Contrary to the assumption that ODMs merely assemble someone else's blueprint, companies such as QCT and Wiwynn explicitly market their own server and rack-level design expertise, not just manufacturing capacity.
- **Thin, volatile margins**: Because ODM economics depend on scale and yield rather than brand premium, the segment is unusually sensitive to customer concentration, demand cycles, and shifts in production location.
- **Limited enterprise service model**: ODMs generally do not offer the phone-support, 4-hour parts replacement, or global services organization that branded OEMs like Dell and HPE sell as part of the premium, making direct ODM purchasing impractical for most buyers below hyperscaler scale.

Feature Comparison

Buyers evaluating "which NVIDIA GPU server vendor" to use need to weigh design ownership, financial scale, and flagship platform against each other simultaneously. *Table 1* below summarizes the primary branded OEMs and the leading ODM tier across these dimensions, using the most recent verifiable figures as of July 2026.

COMPANY	LAYER	FLAGSHIP NVIDIA PLATFORM	MOST RECENT REPORTED REVENUE	Q3 2025 SERVER MARKET SHARE (IDC)	DESIGN OWNERSHIP
Dell Technologies	OEM	PowerEdge XE9680 / XE9712 (HGX H100/H200/B300)	$113.5B FY26 total; $24.68B FY26 AI-optimized servers (Source: investors.delltechnologies.com)	8.3% (Source: my.idc.com)	Chassis, cooling, iDRAC firmware, CPU/memory choice around fixed HGX baseboard
Supermicro	OEM (most vertically integrated)	SYS-821GE-TNHR (HGX B200) / GB200 NVL72	$22.0B FY25 (Source: ir.supermicro.com)	4.0% (Source: my.idc.com)	In-house manufacturing in US, Taiwan, Netherlands; Building Block modular chassis
HPE	OEM	Cray XD670 (HGX H200)	$9.7B Q4 FY25 total (Source: hpe.com)	3.0% (Source: my.idc.com)	Liquid-cooling engineering, Cray HPC networking stack, GreenLake consumption model
Lenovo	OEM	ThinkSystem SR675 V3 / SR680b (HGX H100/H200/B300)	$14.5B FY24/25 Infrastructure Solutions Group	3.6% (statistical tie) (Source: my.idc.com)	Neptune liquid cooling, XClarity firmware, chassis and CPU choice
Foxconn / Ingrasys	ODM	GB200/GB300 NVL72 rack integration	Not separately disclosed	Counted within 59.4% "ODM Direct" (Source: my.idc.com)	Full rack-scale mechanical, thermal, and power design (Source: ingrasys.com)
Quanta / QCT	ODM	QuantaGrid D75x (MGX/HGX B100/B200)	Not separately disclosed	Counted within 59.4% "ODM Direct" (Source: my.idc.com)	Full server and rack-level design, direct-to-chip liquid cooling (Source: qct.io)
Wistron / Wiwynn	ODM	GB300 NVL72 / HGX B300 NVL16	Not separately disclosed	Counted within 59.4% "ODM Direct" (Source: my.idc.com)	Rack-level integration of GPUs, cooling, power, and networking (Source: wiwynn.com)

The table underscores a pattern visible across every layer of this market: revenue disclosed by name does not track cleanly with design responsibility. Dell and Supermicro report the largest branded revenue figures, but IDC's own tracker shows the "ODM Direct" category, which is not broken out by individual ODM company, dwarfs every named branded OEM combined, at nearly six times Dell's share alone. That asymmetry is the single most important fact for anyone trying to understand who "really" designs NVIDIA-based AI infrastructure: the companies most buyers can name control a minority of the market, while companies most buyers have never heard of, working directly for hyperscalers under non-disclosure terms, build the majority of it. A buyer's practical path through this table depends on scale: enterprises and sovereign AI buyers purchasing dozens to low hundreds of servers per year almost always land in the four branded-OEM rows, while only organizations operating at thousands-of-GPUs scale, with dedicated data center operations staff, can realistically negotiate direct with the ODM tier.

Performance and Benchmarks

Because every HGX-based OEM server uses the identical NVIDIA GPU baseboard, with GPU count, memory configuration, NVLink wiring, and NVSwitch layout "all fixed by NVIDIA" regardless of which brand sells the finished system (Source: amcompute.com), raw GPU compute throughput does not meaningfully differ between a Dell, Supermicro, HPE, or Lenovo server running the same NVIDIA generation. NVIDIA's HGX Rubin NVL8 baseboard, the current top-of-line platform as of mid-2026, delivers "up to 10x more token factory throughput versus HGX B200" according to NVIDIA's own published specifications, driven by 400 PFLOPS of NVFP4 inference compute, 176 TB/s of memory bandwidth, and 28.8 TB/s of NVLink Switch bandwidth (Source: nvidia.com). Because this baseboard-level performance is identical across every OEM's system, the meaningful differentiators shift to thermal management, power delivery headroom, and firmware efficiency, exactly the areas branded OEMs compete on. HPE has marketed its Cray XD670 platform's independent MLPerf Inference v5.1 benchmark results directly, claiming "six #1 results" including in computer vision and large-language-model chat and text-generation workloads (Source: hpe.com), the clearest independently-audited benchmark claim among the OEMs surveyed for this report.

Rack-scale systems raise the stakes further because the "server" itself becomes an entire liquid-cooled rack rather than a single chassis. NVIDIA's own published specifications for the GB200 NVL72, which connects 36 Grace CPUs and 72 Blackwell GPUs into one NVLink domain, are the clearest baseline available for comparing every OEM and ODM implementation of that design. *Table 2* below reproduces NVIDIA's published GB200 NVL72 and DGX B200 specifications side by side, since these are the two reference points every branded OEM and ODM rack-scale system is built to match or approach.

SPECIFICATION	NVIDIA GB200 NVL72 (RACK-SCALE)	NVIDIA DGX B200 (SINGLE SYSTEM)
GPU configuration	72 Blackwell GPUs across 18 compute trays	8x NVIDIA B200 GPUs (Source: docs.nvidia.com)
CPU configuration	36 Grace CPUs (2,592 Arm Neoverse V2 cores total)	2x Intel Xeon 8570, "56 cores each" (Source: docs.nvidia.com)
GPU memory	13.4 TB HBM3e	"1,440 GB total GPU memory" (Source: docs.nvidia.com)
NVLink bandwidth	130 TB/s total domain bandwidth (Source: nvidia.com)	"5th generation NVLink switches that provide 14.4 TB/s aggregate bandwidth" (Source: docs.nvidia.com)
LLM inference performance	30x faster than NVIDIA H100 GPU-based systems (Source: nvidia.com)	72 PFLOPS FP8 training with sparsity
Power supplies	Distributed across 1RU power shelves, "33kW per power shelf" (Source: ingrasys.com)	Six power supply units, "5+1 redundancy" (Source: docs.nvidia.com
Who ships it	NVIDIA DGX GB200, plus OEM/ODM equivalents from Supermicro, Dell, and Ingrasys	NVIDIA direct only; not rebadged by third parties

The comparison illustrates why "rack-scale" has become the unit of competition rather than the individual server chassis. A single GB200 NVL72 rack delivers roughly 30 times the real-time trillion-parameter inference throughput of an equivalent H100-generation deployment, according to NVIDIA's own published benchmarks, and NVIDIA's successor GB300 NVL72, which swaps in Blackwell Ultra GPUs and over 20 TB of HBM3e memory, is rated to deliver up to a 50x overall increase in AI factory output performance compared to Hopper-based platforms. Because DGX represents NVIDIA's own reference implementation with a fixed CPU, NIC, DPU, and firmware stack, it functions as the performance and configuration baseline against which every branded OEM and ODM alternative is implicitly measured, even though NVIDIA does not publish list prices for DGX B200, DGX B300, or GB200 NVL72 systems. The practical performance gap between vendors, where it exists, shows up almost entirely in sustained thermal behavior under continuous multi-day training loads rather than in peak theoretical FLOPS, which is precisely why liquid cooling investment (HPE's Cray heritage, Lenovo's Neptune, Supermicro's DLC-2, and Dell's Multi-Vector Cooling and direct liquid cooling options) has become the primary battleground for OEM differentiation as Blackwell-generation GPUs draw up to 1,000 watts each, up from 700 watts for Hopper-generation H100 SXM GPUs.

Data Analysis and Evidence

The clearest quantitative signal in this market is the growth rate itself. IDC's Worldwide Quarterly Server Tracker recorded the server market reaching **$112.4 billion** in revenue during the third quarter of 2025, a **61.1%** year-over-year increase, with revenue from servers containing an embedded GPU growing 49.4% year over year and "representing more than half of the server market revenue" for the first time (Source: my.idc.com). Within that total, IDC separately reports that "revenue generated from x86 servers increased 32.8% in 2025Q3 to $76.3 billion while Non-x86 servers increased 192.7% YoY to $36.2 billion" (Source: my.idc.com), meaning non-x86 architectures, including NVIDIA's Grace-based Arm systems, are growing roughly six times faster than the traditional x86 server base. The United States was the fastest-growing region, up 79.1% year over year, driven by a 105.5% increase in accelerated server spending specifically, while China (referred to in IDC's release as PRC) "is growing at 37.6% year-over-year growth in 2025Q3 accounting for almost a fifth of the quarterly revenue worldwide" (Source: my.idc.com). Beyond the United States and China, IDC's regional breakdown showed "Canada grew 69.8% pushed by the same reason" on the accelerated-server dynamic, while "APeJC, EMEA and Japan also showed very healthy doble digit growth with 37,4%, 31.0% and 28,1% respectively, while Latin America showed a low single digit growth with 4.1% increase in the quarter" (Source: my.idc.com) (Source: my.idc.com), indicating the AI server buildout remains heavily concentrated in North America and China even as it becomes a genuinely global phenomenon.

Table 3 below reproduces IDC's exact Q3 2025 vendor standings, the single most authoritative snapshot available of how server revenue is actually distributed across the branded OEM tier versus direct hyperscaler purchasing.

VENDOR / CATEGORY	Q3 2025 VENDOR REVENUE	Q3 2025 MARKET SHARE	Q3 2024 MARKET SHARE	YOY CHANGE
Dell Technologies	$9,301.62M	8.3%	9.7%	+37.2%
Super Micro (Supermicro)	$4,498.11M	4.0%	7.4%	-13.2%
IEIT Systems	$4,140.48M	3.7% (statistical tie)	6.6%	-10.5%
Lenovo	$4,004.44M	3.6% (statistical tie)	4.6%	+26.1%
Hewlett Packard Enterprise	$3,398.15M	3.0%	5.0%	-2.3%
ODM Direct	$66,790.83M	59.4%	45.1%	+112.2%
Rest of Market	$20,310.95M	18.1%	21.6%	+34.7%
Total	$112,444.59M	100.0%	100.0%	+61.1%

Source: IDC Worldwide Quarterly Server Tracker, published December 11, 2025 (Source: my.idc.com).

This table is the most direct available answer to "who actually builds NVIDIA GPU servers" in dollar terms: the ODM Direct category alone, growing 112.2% year over year, more than doubled its market share contribution and now represents nearly six times Dell's share and almost fifteen times Supermicro's. IDC notes that this shift is being driven by hyperscalers and cloud service providers adopting GPU-embedded servers at a pace that "almost doubled in size compared to 2024," with revenue for the first three quarters of 2025 alone reaching **$314.2 billion** (Source: my.idc.com). On the financial side, the branded OEMs are still growing in absolute dollar terms even as their share of total server revenue shrinks: Dell's revenue climbed 37.2% year over year even as its market share slipped from 9.7% to 8.3%, because the ODM Direct segment is expanding faster than the entire market. Supermicro and HPE, by contrast, saw both share and year-over-year revenue decline in the same quarter, reflecting intensified competition and, in Supermicro's case, the reputational drag of its ongoing compliance issues discussed above. On the manufacturing-geography side, Penchan's supply-chain analysis estimates that Taiwan handles the large majority of physical AI server assembly, but stresses that the widely repeated "roughly 90%" figure measures production volume, not the profit captured by Taiwanese firms, since branded OEMs and hyperscalers retain most of the margin even when a Taiwanese contractor does the physical build (Source: penchan.co).

Case Studies and Real-World Examples

xAI's Colossus: A Supplier Shift in Real Time

Elon Musk's AI company **xAI** built its 100,000-GPU Colossus supercomputer in Memphis, Tennessee, using Supermicro hardware, with Supermicro's own case study describing the cluster as connecting "100,000 NVIDIA Hopper Tensor Core GPUs" through NVIDIA's Spectrum-X Ethernet platform in a liquid-cooled configuration (Source: supermicro.com). By November 2024, however, trade press reporting confirmed that xAI had "redirected its AI server orders from Supermicro to Dell, delivering a significant lift to Dell and its key suppliers, Inventec and Wistron" (Source: networkworld.com). The report cited Supermicro's auditor resignation and delayed SEC filings as contributing factors in the shift, and noted that Wistron, "responsible for manufacturing motherboards and assembling Dell's AI servers," was expanding production at three Hsinchu, Taiwan facilities and its Mexico operations to absorb the new demand, while Inventec, described as "one of Dell's top three server assembly providers," was similarly positioned to benefit (Source: networkworld.com) (Source: networkworld.com). Analyst commentary in

the same report noted that "both Wistron and Inventec have been increasing their inventories throughout 2024, with Wistron holding over $5 billion in inventory and Inventec over $2 billion," positioning them to absorb the reallocated demand quickly (Source: networkworld.com). This single case illustrates all three supply-chain layers simultaneously: an NVIDIA GPU baseboard, a branded OEM (first Supermicro, then Dell) taking the customer relationship, and Taiwan-based ODMs (Wistron and Inventec) doing the underlying assembly regardless of which brand's logo shipped on the box.

CoreWeave and the Neocloud OEM Relationship

CoreWeave, a specialized GPU cloud provider, announced in December 2023 that it would use "Dell PowerEdge XE9860 servers with NVIDIA H100 Tensor Core GPUs" as key infrastructure powering its cloud platform, with Dell's press release describing CoreWeave as "a specialized cloud provider for large-scale NVIDIA GPU-accelerated workloads" (Source: dell.com). The deal also included Dell ProSupport services and dedicated account management, illustrating why some "neocloud" buyers, despite operating at a scale that could justify direct ODM purchasing, still choose branded OEM relationships for the support and lifecycle guarantees ODMs typically do not offer. SemiAnalysis independently confirms CoreWeave as one of three flagship accounts, alongside Tesla and x.ai, where "Dell specifically has gained sockets" against Supermicro's earlier incumbency (Source: newsletter.semianalysis.com).

Tesla's In-House Supercomputer

Tesla built one of the earliest large-scale NVIDIA-based AI training clusters for its Autopilot and self-driving development, unveiled by senior AI director Andrej Karpathy at a CVPR conference session. NVIDIA's own blog post on the system describes the cluster as using "720 nodes of 8x NVIDIA A100 Tensor Core GPUs" for a total of "5,760 GPUs total" to achieve "an industry-leading 1.8 exaflops of performance" (Source: blogs.nvidia.com). Karpathy described it at the time as roughly the world's fifth-most-powerful supercomputer by raw FLOPS. Tesla's approach illustrates a pattern distinct from both the pure-OEM and pure-ODM models: heavy involvement from Tesla's own hardware engineering team in cluster design, working with server suppliers (reported elsewhere to include both Dell and Supermicro across different cluster generations) rather than relying entirely on either a branded OEM's off-the-shelf platform or a hyperscaler-style direct ODM relationship.

Meta's Grand Teton: A Hyperscaler Co-Designs Its Own NVIDIA Platform

Meta Platforms represents the clearest example of a hyperscaler designing its own NVIDIA-based server rather than buying an OEM's or even an ODM's pre-existing design. Announced at the 2022 OCP Global Summit, Meta's **Grand Teton** platform used NVIDIA H100 Tensor Core GPUs to train and run AI models, with NVIDIA's own blog describing the announcement as "Meta today announced its next-generation AI platform, Grand Teton, including NVIDIA's collaboration on design" (Source: blogs.nvidia.com). NVIDIA's vice president of hyperscale and high-performance computing, Ian Buck, framed the release as opening Meta's design to the broader industry: "system builders around the world will soon have access to an open design for hyperscale data center compute infrastructure" (Source: blogs.nvidia.com). Grand Teton packed twice the network bandwidth and four times the host-to-GPU bandwidth of Meta's prior Zion platform, and Meta's own infrastructure hardware VP Alexis Bjorlin said integrating everything into a single server "dramatically simplifies deployment," letting Meta install and provision its fleet faster and more reliably. Meta subsequently expanded Grand Teton's design to support AMD's MI300X accelerators and unveiled a Catalina rack architecture for NVIDIA Blackwell chips, underscoring that hyperscalers with sufficient in-house hardware engineering capacity treat OEM and ODM relationships as optional rather than mandatory.

Supermicro's Compliance History: A Cautionary Case in OEM Selection

Beyond product design, Supermicro's regulatory history is itself a relevant case study for buyers evaluating OEM risk. The SEC's August 2020 enforcement action found the company "violated federal securities laws by engaging in improper accounting" from fiscal 2015 through 2017, prematurely recognizing revenue on undelivered goods and understating expenses, resulting in a **$17.5 million** civil penalty against the company and additional penalties against its then-CFO (Source: sec.gov). More recently, the March 2026 Department of Justice indictment alleged that Supermicro co-founder Yih-Shyan "Wally" Liaw and two associates conspired to divert servers "fitted with Nvidia GPUs worth $2.5 billion to Chinese customers in violation of US export controls," allegedly using a Southeast Asian pass-through company, falsified paperwork, and staged non-functional "dummy" servers to deceive both Supermicro's own auditors and Commerce Department inspectors (Source: theregister.com) (Source: theregister.com). Supermicro itself was not charged, and it stated the alleged conduct contravened its own compliance policies; the individuals involved face up to 30 years in prison if convicted, and one remains a fugitive as of the indictment's unsealing. This case underscores that "who designs your NVIDIA server" carries governance and export-control implications beyond pure engineering capability, a factor sovereign, government, and regulated-industry buyers weigh explicitly in vendor selection.

NVIDIA's GB200 NVL72: When the Rack Itself Becomes the Product

The GB200 NVL72 case is worth treating separately because it shows how thoroughly rack-scale design has displaced the single-chassis server as the unit buyers actually purchase. NVIDIA's own specification describes the platform as connecting "36 Grace CPUs and 72 Blackwell GPUs in a rack-scale, liquid-cooled design" that functions as a single 130 TB/s NVLink domain, delivering 30 times the real-time trillion-parameter inference performance of an equivalent NVIDIA H100 deployment (Source: nvidia.com). Foxconn's Ingrasys markets its own manufacturing capability around this exact design, describing its Taiwan operation as the "world's first AI server lighthouse factory" (Source: ingrasys.com), while Supermicro simultaneously markets its own GB200 NVL72 configurations as part of its "first-to-market" Blackwell portfolio and Wiwynn and Wistron jointly market GB300 NVL72 readiness (Source: wiwynn.com. In practice this means a buyer asking for a "GB200 NVL72" can receive functionally similar hardware from at least three different corporate sources, an OEM, an ODM, or NVIDIA directly as DGX GB200, each implementing the identical NVIDIA-fixed compute and NVLink specification but differing in support terms, delivery lead time, and price.

Implications and Future Directions

The data assembled in this report point toward three durable structural trends. First, the gap between branded OEM revenue share and ODM Direct revenue share is widening, not narrowing: ODM Direct grew from 45.1% to 59.4% of total server market revenue in a single year even as the overall market itself grew 61.1%, which means hyperscaler-direct purchasing is capturing a disproportionate share of already explosive growth (Source: my.idc.com). For non-hyperscaler buyers, this trend does not eliminate the branded OEM's relevance, since Dell, Supermicro, HPE, and Lenovo remain the only practical channel for enterprises, sovereign AI initiatives, and mid-sized neoclouds that need warranty support and predictable lead times rather than hyperscaler-scale purchasing power, but it does mean list pricing and lead times for branded OEM hardware will increasingly be set at the margin by how much capacity NVIDIA and its ODM partners have left over after satisfying direct hyperscaler orders.

Second, NVIDIA's own reference-architecture strategy, splitting MGX (flexible, modular, multi-vendor) from HGX (fixed, high-bandwidth, training-optimized) from DGX (fully NVIDIA-owned turnkey), is deliberately designed to let OEMs and ODMs compete on secondary attributes like cooling, firmware, and CPU choice while NVIDIA retains full control over the GPU baseboard itself. MGX in particular, which according to NVIDIA's original 2023 announcement "can slash development costs by up to three-quarters and reduce development time by two-thirds to just six months," is pulling smaller and mid-tier partners like Gigabyte, ASUS, ASRock Rack, Pegatron, and QCT into the AI server market at a pace that would have been impractical if each vendor had to engineer a rack-scale liquid-cooled system from scratch (Source: nvidia.com). Expect this modular-reference-design strategy to keep expanding the roster of viable OEM and ODM brands even as GPU generations (Blackwell, Rubin, and beyond) get more thermally demanding and mechanically complex. The move from GB200 NVL72 to GB300 NVL72, rated at up to a 50x overall AI factory output increase over Hopper-based platforms, illustrates how quickly the rack-scale reference design itself is iterating, and every OEM and ODM in this report must re-qualify its chassis, cooling, and power delivery against each new generation roughly annually.

Third, the Supermicro compliance episodes documented in this report, the 2020 SEC accounting settlement and the 2026 DOJ export-control indictment, are likely to accelerate a "flight to compliance" among regulated buyers (government, defense, financial services, and multinational sovereign AI programs), even where Supermicro's underlying engineering and pricing remain competitive. Buyers in these categories should expect increased due-diligence requirements, additional contractual export-control attestations, and potentially longer vendor-approval cycles when evaluating any OEM or ODM with material China-manufacturing exposure, a category that, per NVIDIA's own MGX partner list, includes nearly every ODM discussed in this report. Taken together, these three trends suggest that over the next several product generations the practical distinction buyers should track is not "which brand's logo is on the chassis" but "which layer of the supply chain, NVIDIA's fixed baseboard, an OEM's chassis and firmware engineering, or an ODM's rack-scale integration, actually determines the price, lead time, and support terms of the system in front of them."

Frequently Asked Questions (FAQs)

What is the difference between an OEM and an ODM server manufacturer? An OEM (original equipment manufacturer), in the branded-server context, designs the chassis, cooling, firmware, and support model around NVIDIA's fixed GPU baseboard and sells the finished system under its own brand and warranty. An ODM (original design manufacturer) both designs and physically builds the hardware, typically shipping it under the buyer's own brand or directly to a hyperscaler with no third-party brand at all, and captures a much thinner margin for doing so (Source: penchan.co).

Which companies rebadge NVIDIA DGX servers? None do, strictly speaking. DGX is NVIDIA's own fixed-configuration product line, sold directly by NVIDIA with NVIDIA's own support and software stack; it is not rebadged by third parties. What OEMs and ODMs build instead are separate HGX-based systems using the identical GPU baseboard as DGX but with OEM-chosen CPUs, memory, storage, and firmware, and NVIDIA's GB200 and GB300 NVL72 rack-scale designs are offered both as NVIDIA-branded DGX systems and as equivalent rack configurations sold by partners like Supermicro, Dell, and the ODM tier.

How does Supermicro compare to Dell and HPE on NVIDIA server design? All three build systems around the identical, NVIDIA-fixed HGX baseboard, so raw GPU compute performance does not meaningfully differ between them. The differences are in vertical integration (Supermicro manufactures more in-house), pricing (Supermicro configurations often undercut Dell and HPE list prices), support model (Dell and HPE offer more extensive global service organizations), and governance history (Supermicro has faced SEC and DOJ actions that Dell and HPE have not).

Is there a list of NVIDIA-certified server OEMs? Yes. NVIDIA maintains a public NVIDIA-Certified Systems partner directory listing more than 400 certified systems from partners spanning Dell, HPE, Lenovo, Supermicro, Cisco, ASUS, Gigabyte, Fujitsu, H3C, Hitachi, QCT, Wistron, Inventec, Pegatron, and more, queryable through NVIDIA's Qualified System Catalog (Source: nvidia.com).

What is an original design manufacturer in the context of NVIDIA AI servers? It is a Taiwan-headquartered company, typically Foxconn (via Ingrasys), Quanta (via QCT), Wistron, Wiwynn, or Inventec, that performs full rack-scale mechanical, thermal, and power-delivery engineering around NVIDIA's GPU baseboard, then either ships the finished product under a hyperscaler's own name or, in the "ODM Direct" category IDC tracks separately, with no third-party brand at all.

How is NVIDIA's own reference design different from an OEM's custom server design? NVIDIA's reference designs, HGX, MGX, and the GB200/GB300 NVL72 rack architecture, fix the GPU count, NVLink wiring, and NVSwitch topology; nothing downstream can change that layer. An OEM's custom design work happens entirely around that fixed core: choice of CPU vendor, memory capacity and speed, storage configuration, network interface cards, chassis height and cooling method (air, hybrid, or full liquid), baseboard management controller firmware, and the software and support stack layered on top. Two servers can therefore share an identical NVIDIA GPU baseboard while differing substantially in real-world thermal headroom, serviceability, and total cost of ownership.

Which NVIDIA GPU server vendor is best for a given buyer? There is no universal answer; the report's evidence suggests Dell suits buyers who need enterprise-grade global support and predictable lead times, Supermicro suits buyers prioritizing price and speed to new NVIDIA platforms while accepting a leaner support model and Supermicro's compliance history as a risk factor, HPE suits buyers needing proven liquid-cooling engineering and HPC pedigree, and Lenovo suits multinational buyers needing broad geographic support coverage, while true hyperscale buyers with in-house hardware teams increasingly bypass all four in favor of direct ODM relationships.

Conclusion

The honest answer to "which server OEMs design their own NVIDIA servers" is that all of the major branded OEMs, Dell, Supermicro, HPE, and Lenovo, genuinely design meaningful portions of their NVIDIA-based systems (chassis, cooling, power delivery, firmware, and CPU and memory selection), while NVIDIA itself fixes the GPU baseboard for every one of them, and a separate, less visible tier of Taiwan-based ODMs, Foxconn's Ingrasys, Quanta's QCT, Wistron, Wiwynn, and Inventec, does design and physical-assembly work that in dollar terms now dwarfs the combined revenue of every branded OEM. Dell leads the branded tier on both revenue scale and disclosed AI backlog, Supermicro leads on vertical integration and time-to-market despite a documented history of accounting and export-control controversies, HPE leads on liquid-cooling and HPC pedigree, and Lenovo leads on global support reach among the second tier. Beneath all four, ODM Direct sales to hyperscalers, now 59.4% of total server market revenue and growing faster than the market overall, represent the segment doing the most actual engineering work relative to its public visibility. Buyers choosing among these options in the second half of 2026 should weigh not just raw GPU performance, which is functionally identical across every HGX-based system regardless of brand, but support model, compliance history, liquid-cooling maturity, and lead-time exposure to the same NVIDIA allocation constraints every vendor in this market shares.

Tags: nvidia server oems, dell vs supermicro, nvidia gpu servers, oem vs odm servers, hgx dgx mgx, supermicro, dell poweredge, hpe cray, ai server manufacturers, gb200 nvl72
