NVIDIA AI Server OEM Comparison: Supermicro vs Dell vs HPE

Published July 21, 2026 34 min read

Executive Summary

Enterprises and cloud providers evaluating **NVIDIA** AI infrastructure no longer buy chips directly from NVIDIA; they buy racks assembled by original equipment manufacturers (OEMs) and original design manufacturers (ODMs) that license NVIDIA's reference architectures. As of July 2026, the market for **NVIDIA GB200 NVL72** and **GB300 NVL72** rack-scale systems is dominated by a small group of branded OEMs, principally **Dell Technologies**, **Supermicro**, **Hewlett Packard Enterprise (HPE)**, and **Lenovo**, alongside Taiwanese original design manufacturers **Foxconn** (Hon Hai Precision Industry), **Quanta Cloud Technology (QCT)**, **Wiwynn**, and **Wistron** that build the bulk of hyperscaler-direct racks. According to IDC's Worldwide Quarterly Server Tracker as reported by Electronics Weekly, **Dell Technologies** claimed the top position in the first quarter of 2026 with a **16.5%** worldwide server revenue share on **$20.28 billion** in quarterly revenue, up **244.1%** year over year, while **Super Micro Computer** held second place with **7.6%** share and **Lenovo** ranked third at **4.6%** (Source: electronicsweekly.com). "ODM Direct" vendors, the unbranded Taiwanese manufacturers that sell straight to hyperscalers such as Meta, Microsoft, and Google, still controlled the largest single share of the market at **50.2%**, though that was down from **64.1%** a year earlier as branded OEMs captured a growing share of AI infrastructure spending (Source: electronicsweekly.com).

Each OEM has converged on the same underlying NVIDIA silicon but differentiates on liquid-cooling engineering, supply chain scale, and services. Dell reported **$64 billion** in cumulative AI-optimized server orders for fiscal year 2026, with **$25 billion** shipped and a **$43 billion** backlog entering fiscal 2027, and raised its FY27 AI server revenue outlook to **$60 billion** after booking **$24.4 billion** in new AI orders in a single quarter (Source: investors.delltechnologies.com (Source: investors.delltechnologies.com). Supermicro, which built its reputation on early **direct liquid cooling (DLC)**, posted **$10.2 billion** in fiscal third-quarter 2026 net sales with more than **80%** of revenue tied to AI GPU platforms, though gross margin remained thin at **9.9%** (Source: ir.supermicro.com). HPE's Cloud & AI segment revenue grew **22.9%** year over year to **$7.7 billion** in its fiscal second quarter of 2026, with server revenue up **32.7%** to **$5.5 billion**, and CEO Antonio Neri described the company's backlog as "record-breaking" (Source: hpe.com (Source: investors.hpe.com). Foxconn, which assembles roughly **40%** of global AI rack volume by its own chairman's account, reported a **40%** jump in quarterly sales tied to AI server demand, with June 2026 revenue alone reaching approximately **$45 billion** (Source: thenextweb.com).

The **NVIDIA GB200 NVL72** connects **36 Grace CPUs** and **72 Blackwell GPUs** in a liquid-cooled, rack-scale design delivering **130 terabytes per second** of NVLink bandwidth and up to **30x** faster real-time trillion-parameter LLM inference than the prior H100 generation (Source: nvidia.com). All major OEMs license this design, plus the successor **GB300 NVL72** and forthcoming **Vera Rubin NVL72**, through NVIDIA's open **MGX** modular reference

architecture, which more than **50 partners** are adopting for the Vera Rubin generation alone (Source: blogs.nvidia.com). Rising Vera Rubin rack prices, estimated at **$5 million to $8.8 million** and potentially higher as HBM4 memory costs climb, are compressing OEM margins even as unit revenue grows, since NVIDIA increasingly ships more pre-integrated content itself (Source: tomshardware.com). For buyers, the practical choice is rarely about raw GPU access, since all OEMs ship the same **NVIDIA silicon**, but about liquid-cooling maturity, delivery speed, services and financing, regional support, and each vendor's position in the order queue as **Blackwell and Vera Rubin supply remains constrained through 2026 and into 2027**.

Introduction and Background

When an enterprise, cloud provider, or government agency decides to deploy **NVIDIA** graphics processing units (GPUs) at scale, it does not purchase chips from NVIDIA directly. Instead, it buys a complete rack-scale system, comprising GPUs, CPUs, memory, networking, **power delivery**, and liquid cooling, from one of a small number of qualified server manufacturers. This report examines the competitive landscape of NVIDIA AI server original equipment manufacturers (OEMs) as of July 2026, comparing **Dell Technologies**, **Supermicro**, **Hewlett Packard Enterprise**, **Lenovo**, and the Taiwanese original design manufacturers (ODMs), principally **Foxconn** (Hon Hai Precision Industry), **Quanta Cloud Technology**, and **Wiwynn**, that assemble the majority of hyperscaler-direct AI infrastructure.

The distinction between an OEM and an ODM matters for buyers. Branded OEMs such as Dell and HPE sell fully supported, serviced systems under their own name, typically to enterprises, sovereign governments, and mid-tier cloud providers that need integration services, financing, and a single point of accountability. ODMs such as Foxconn and Quanta build largely unbranded racks to a hyperscaler's own specifications, a category IDC terms "ODM Direct," and this segment still supplied **50.2%** of worldwide server revenue in the first quarter of 2026 even as its share compressed from **64.1%** a year earlier (Source: electronicsweekly.com). Total worldwide server market revenue reached **$122.6 billion** in the first quarter of 2026, up **30.4%** year over year, with GPU-accelerated servers alone generating **$68.9 billion**, or **56.2%** of the total (Source: electronicsweekly.com).

The underlying hardware that all of these companies build around is NVIDIA's rack-scale reference architecture. The **GB200 NVL72**, introduced in 2024, connects **36 Grace CPUs** and **72 Blackwell GPUs** into a single liquid-cooled rack that functions as one giant GPU domain, delivering **130 TB/s** of NVLink bandwidth (Source: nvidia.com). Its successor, the **GB300 NVL72**, integrates **72 Blackwell Ultra GPUs** and delivers up to **50x** the AI factory output of Hopper-generation systems (Source: nvidia.com).NVIDIA does not manufacture these racks in volume itself for most customers; instead it licenses the electro-mechanical design and publishes the **MGX** modular reference architecture so that OEMs and ODMs can build compliant systems faster and more cheaply (Source: nvidia.com). NVIDIA itself continues to post record results from this ecosystem: fiscal first-quarter 2027 (ended April 26, 2026) revenue reached **$81.6 billion**, up **85%** year over year, with Data Center revenue of **$75.2 billion**, up **92%** (Source: investor.nvidia.com (Source: investor.nvidia.com). This report walks through each major OEM's capabilities, adoption, and strengths and limitations, presents a feature comparison matrix, reviews available performance data, quantifies the market with data drawn from IDC, TrendForce, and public filings, profiles five named deployments, and concludes with implications for buyers navigating a still supply-constrained market.

Dell Technologies

Capabilities

Dell's AI server line centers on the **PowerEdge XE9680**, an air-cooled 6U server supporting eight NVIDIA HGX H100 or H200 GPUs, and the newer **PowerEdge XE9712**, a rack-scale, direct-liquid-cooled platform built around NVIDIA's GB200 and GB300 NVL72 architecture that unifies up to **72 Blackwell GPUs** per rack with **1.8 TB/s** of GPU-to-GPU NVLink bandwidth (Source: dell.com (Source: delltechnologies.com). Dell markets these as part of its Integrated Rack Scalable Systems (IRSS) program, which delivers fully built and pre-tested racks rather than individual servers, reducing customer integration time (Source: delltechnologies.com). In June 2026, Dell became the first vendor to ship rack systems built on NVIDIA's next-generation **Vera Rubin** platform, delivering **PowerEdge XE9812** servers in a **Dell PowerRack** configuration to CoreWeave, which Dell says delivers up to **10x** lower cost per token than Grace Blackwell NVL72 for large-scale agentic inference (Source: dell.com). Dell's AI Data Platform additionally bundles unstructured data engines built with Elasticsearch and a federated SQL engine from Starburst, positioning the offering as a full-stack AI factory rather than raw compute alone (Source: siliconangle.com).

Adoption

Dell's Infrastructure Solutions Group (ISG) posted record full-year fiscal 2026 revenue of **$60.8 billion**, up **40%** year over year, and record quarterly ISG revenue of **$19.6 billion** in the fourth quarter, up **73%** (Source: investors.delltechnologies.com). Across fiscal 2026, Dell closed more than **$64 billion** in AI-optimized server orders, shipped over **$25 billion**, and entered fiscal 2027 with a record **$43 billion** backlog, according to COO Jeff Clarke (Source: investors.delltechnologies.com). Momentum accelerated further in the first quarter of fiscal 2027 (ended roughly May 2026), when Dell booked **$24.4 billion** in new AI orders, recognized **$16.1 billion** of AI server revenue, and raised its FY27 AI server revenue guidance to **$60 billion**, with total ISG revenue up **181%** year over year to **$29 billion** (Source: investors.delltechnologies.com (Source: fool.com). In IDC's Q1 2026 vendor tracker, Dell held

the number-one position in the overall server market with a **16.5%** revenue share, more than double any other named vendor (Source: electronicsweekly.com). Dell's marquee AI infrastructure customer is CoreWeave, the AI-focused cloud provider, which received the first Dell PowerEdge XE9712 GB200 NVL72 racks in December 2024, the first GB300 NVL72 racks in mid-2025, and the first Vera Rubin-based PowerRack systems in June 2026 (Source: investors.delltechnologies.com (Source: dell.com). Dell was also confirmed by Elon Musk and Michael Dell as one of two server suppliers, alongside Supermicro, for xAI's Colossus supercomputer in 2024 (Source: datacenterdynamics.com).

Strengths and Limitations

Dell's principal strength is scale combined with enterprise sales and services reach: its **$43 billion** backlog and status as the first vendor to ship each successive NVIDIA rack generation, GB200, GB300, and Vera Rubin, to a major cloud customer signal deep engineering coordination with NVIDIA (Source: investors.delltechnologies.com). Dell's global services organization and financing options, useful for enterprises without hyperscaler-grade data center operations teams, differentiate it from ODM-direct alternatives. The chief limitation is that AI server gross margins are thin industry-wide; Dell's ISG segment growth has come with gross margin dollars growing more slowly than revenue in dollar terms because of high AI server mix, a dynamic also visible at HPE and Supermicro (Source: fool.com). Component and GPU allocation constraints, not unique to Dell, also mean order-to-delivery lead times can stretch even for a vendor with Dell's purchasing scale.

Supermicro

Capabilities

Supermicro (Super Micro Computer, Inc.) built its AI server franchise on "Server Building Block Solutions," a modular architecture letting customers configure form factor, processor, memory, GPU, storage, networking, power, and cooling independently (Source: ir.supermicro.com). The company offers NVIDIA HGX B300, B200, and GB200 NVL72 solutions and was among the first companies, alongside QCT, to bring NVIDIA's MGX modular reference architecture to market in 2023 (Source: supermicro.com (Source: nvidianews.nvidia.com). Supermicro's signature technical differentiator is direct liquid cooling (DLC): the company states it has deployed more than **100,000 NVIDIA GPUs** with its DLC solution and delivered over **2,000 liquid-cooled racks** since June 2024, with a design that can dissipate up to **1,600 W** per cold plate for next-generation GPUs (Source: supermicro.com (Source: supermicro.com). Datacenter Dynamics reports that Supermicro claims to be the biggest global supplier of direct liquid cooling, though independent market-share verification for that specific claim was not available at the time of writing (Source: datacenterdynamics.com).

Adoption

Supermicro reported fiscal third-quarter 2026 (ended March 31, 2026) net sales of **$10.2 billion**, up **123%** year over year, with AI GPU-related platforms contributing more than **80%** of total revenue (Source: ir.supermicro.com (Source: fool.com). Non-GAAP gross margin recovered to **10.1%** after component shortages and customer site readiness delays pressured the prior quarter, and CEO Charles Liang pointed to new US manufacturing capacity in Silicon Valley as a driver of continued growth (Source: ir.supermicro.com). In IDC's Q1 2026 tracker, Supermicro retained second place in overall worldwide server revenue share at **7.6%**, growing **128.9%** year over year (Source: electronicsweekly.com). Supermicro's best-known AI deployment is xAI's Colossus cluster in Memphis, Tennessee, where its liquid-cooled 4U servers, each housing eight NVIDIA H100 GPUs, formed a substantial portion of the original **100,000-GPU** buildout completed in just **122 days** (Source: supermicro.com (Source: learn-more.supermicro.com). In March 2025, Supermicro also formed a distribution partnership with Eviden to sell its GB200 NVL72 AI SuperCluster offering across Europe, India, the Middle East, and South America, expanding its geographic reach beyond direct sales (Source: datacenterdynamics.com).

Strengths and Limitations

Supermicro's core strength is speed to market on liquid cooling and rack configuration flexibility: the company claims a two-to-four-week lead time from order to shipment for DLC racks, versus what it describes as four months to a year historically for the wider industry (Source: theregister.com). Its limitations are governance and margin related: the company severed ties with a co-founder named in a federal indictment in March 2026, and its non-GAAP gross margin of roughly **10%** trails Dell's and HPE's blended enterprise margins because Supermicro's revenue mix skews more heavily toward hardware-only, lower-margin AI GPU platforms and less toward services (Source: cnbc.com (Source: ir.supermicro.com). Supermicro also lacks the enterprise services and financing depth of Dell and HPE, which matters for buyers that need long-term support contracts rather than hardware alone.

Hewlett Packard Enterprise (HPE)

Capabilities

HPE's flagship AI server is the **Cray XD670**, a direct-liquid-cooled system optimized for LLM training, natural language processing, and multimodal training, which delivered six number-one results in the MLPerf Inference v5.1 benchmark suite, including in computer vision and LLM chat question-and-answer and text generation categories (Source: hpe.com). HPE also sells "**NVIDIA GB200 NVL72 by HPE**," a fully integrated rack-scale system combining NVIDIA compute, networking, and software with HPE's own liquid-cooling and services expertise, and has announced "NVIDIA Vera Rubin NVL72 by HPE" for frontier models exceeding one trillion parameters (Source: hpe.com). HPE positions its offering around "HPE AI Factory," a full-stack bundle including GreenLake consumption-based management, and CEO Antonio Neri noted the company now considers itself the largest OEM partner of Broadcom in networking, reinforcing a broader compute-plus-networking-plus-storage sales motion rather than compute alone (Source: investors.hpe.com).

Adoption

In its fiscal second quarter of 2026 (ended April 30, 2026), HPE reported total revenue of **$10.7 billion**, up **40%** year over year, a record for the company (Source: investors.hpe.com). Within that total, the Cloud & AI segment generated **$7.7 billion**, up **22.9%** year over year, with the server sub-segment contributing **$5.5 billion**, up **32.7%** (Source: hpe.com (Source: hpe.com). On the earnings call, CFO Marie Myers and CEO Antonio Neri described a "very large backlog in servers" and said the "pipeline remains multiples of the current backlog, which is record-breaking at the company level," while raising full-year Cloud & AI revenue growth guidance to the low-20% range from a prior mid-to-high single-digit outlook (Source: investors.hpe.com (Source: investors.hpe.com). IDC's Q1 2026 tracker placed HPE fifth in overall worldwide server revenue share at **3.0%**, up **17.2%** year over year (Source: electronicsweekly.com). Named HPE AI infrastructure deployments include a collaboration with Japanese telecom operator KDDI to open the Osaka Sakai Data Center with a GB200 NVL72 platform by early 2026, and a sovereign AI factory built for the University of Utah using Cray XD670 servers and NVIDIA Hopper GPUs, expected to more than triple the university's computing capacity (Source: hpe.com (Source: hpe.com).

Strengths and Limitations

HPE's principal strength is its multi-decade liquid-cooling and supercomputing pedigree from the Cray acquisition, combined with a networking portfolio strengthened by the Juniper Networks acquisition; Neri specifically credited "the strength of our combined networking portfolio" for the quarter's results (Source: hpe.com). HPE also explicitly prioritizes enterprise and sovereign AI customers over lower-margin service-provider deals, with CFO Myers noting that "enterprise and sovereign typically being a more profitable sort of part of the mix compared to, say, your classic service provider or model builder" (Source: investors.hpe.com). The limitation of this strategy is scale: HPE's **3.0%** overall server revenue share trails Dell and Supermicro by a wide margin, meaning HPE is less likely to win the largest hyperscaler-direct GPU deals and instead competes chiefly for enterprise, government, and sovereign AI factory contracts where its services model commands a premium (Source: electronicsweekly.com).

Foxconn, Lenovo, and the Broader ODM and Component Ecosystem

Capabilities

Foxconn, officially Hon Hai Precision Industry Co., is the world's largest electronics manufacturer and functions as NVIDIA's primary rack assembly partner, alongside Supermicro and Chenbro, for compute tray and chassis production (Source: blogs.nvidia.com). Bloomberg describes Hon Hai as "Nvidia Corp.'s server assembly partner," and the company is building its own showcase installation, the Hon Hai Kaohsiung Super Computing Center in Taiwan, featuring **64 racks** and **4,608** Blackwell Tensor Core GPUs for an expected **90 exaflops** of AI performance (Source: fortune.com (Source: blogs.nvidia.com). Lenovo, another MGX ecosystem partner, offers the liquid-cooled **ThinkSystem SC777 V4 Neptune**, built on the NVIDIA GB200 platform, and has shipped the **GB300 NVL72 (Type 7DJV)**, which integrates **72 Blackwell Ultra GPUs** and **36 Grace processors** into a single unified rack (Source: lenovopress.lenovo.com (Source: pubs.lenovo.com. Lenovo's Neptune liquid-cooling chassis, the ThinkSystem N1380, uses 100% direct water cooling and can reduce data center power consumption by up to **40%**, enabling 100kW-plus racks without specialized air conditioning (Source: news.lenovo.com). Below the branded OEM tier, NVIDIA's original **MGX** launch in May 2023 named **ASRock Rack**, **ASUS**, **GIGABYTE**, **Pegatron**, **QCT**, and **Supermicro** as first adopters, and both ASUS and GIGABYTE now ship their own GB300 NVL72 rack-scale products, the ASUS **XA GB721-E2** and GIGABYTE **GIGAPOD**, each integrating **72 Blackwell Ultra GPUs** and **36 Grace CPUs** (Source: nvidianews.nvidia.com (Source: dlcdnet.asus.com (Source: gigabyte.com). Two other Taiwanese ODMs, **Wiwynn** and **Quanta Cloud Technology (QCT)**, along with Wistron, round out the hyperscaler-direct supply base.

Adoption

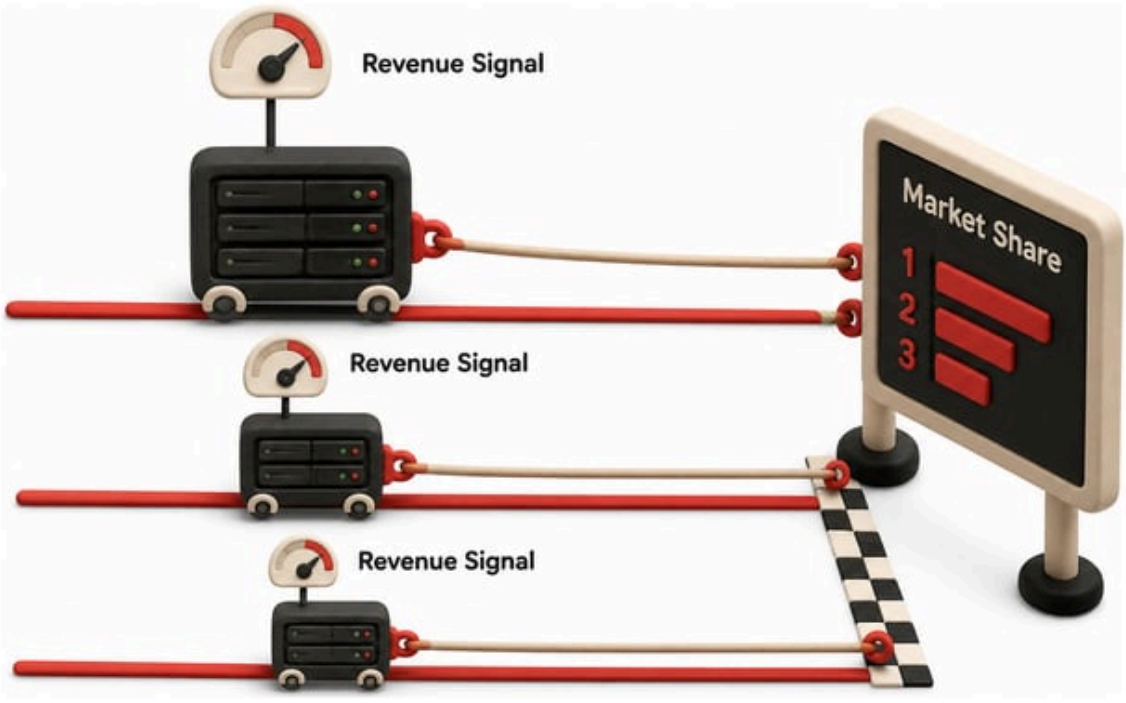
Foxconn's scale in AI servers is exceptional even by industry standards. Chairman Young Liu has said AI rack shipments are on track to double in 2026, with the company claiming roughly **40%** of global AI rack assembly, and Foxconn reported a **40%** year-over-year jump in quarterly sales with June 2026 revenue alone reaching approximately **NT$1.33 trillion** (roughly **$45 billion**) (Source: thenextweb.com (Source: thenextweb.com). In its own first-quarter 2026 results, Hon Hai reported record revenue of **NT$2.12 trillion**, up **29%** year over year, with the Cloud and Networking product segment, which includes AI servers, now accounting for nearly **50%** of total revenue, and it forecasts full-year AI rack shipments to more than double in 2026 (Source: honhai.com (Source: honhai.com (Source: honhai.com). Wiwynn reported first-quarter 2026 consolidated revenue of **NT$276.5 billion**, up **62.0%** year over year, a record for the period (Source: wiwynn.com). Quanta reported record first-quarter 2026 revenue of **NT$809.22 billion** (approximately **$25.68 billion**), up **66.6%** year over year, with servers accounting for **80%** of quarterly revenue, and executives said major cloud service providers now have order visibility through 2028 (Source: taipeitimes.com (Source: taipeitimes.com). In IDC's tracker, Lenovo moved to third place in overall worldwide server revenue share with **4.6%**, growing **36.5%** year over year (Source: electronicsweekly.com).

Strengths and Limitations

Foxconn's strength is unmatched manufacturing scale and vertical integration, spanning chassis, power, and increasingly components such as optical modules and connectors, which lets it absorb enormous hyperscaler order volumes that branded OEMs cannot match on unit economics (Source: honhai.com). Its limitation, shared by Quanta and Wiwynn, is that ODM-direct business generally lacks the branded warranty, enterprise financing, and dedicated account services that Dell, HPE, and to a lesser extent Supermicro provide, which is why ODM Direct market share compressed from **64.1%** to **50.2%** as more buyers outside the largest hyperscalers opted for branded OEM support (Source: electronicsweekly.com). Lenovo's strength is its Neptune liquid-cooling heritage from decades of supercomputing deployments, but its **4.6%** overall server share leaves it behind Dell and Supermicro in AI-specific volume (Source: electronicsweekly.com). Smaller MGX partners such as ASUS, GIGABYTE, and ASRock Rack compete chiefly on price and specialization for smaller deployments and channel/reseller markets rather than the largest gigawatt-scale contracts.

Feature Comparison

Table 1 below summarizes how the leading branded OEMs and principal ODMs compare on their flagship NVIDIA rack-scale platforms, cooling approach, disclosed financial scale, and market position as of mid-2026.



VENDOR	FLAGSHIP NVIDIA RACK PLATFORM	COOLING	RECENT QUARTERLY AI/SERVER REVENUE SIGNAL	WORLDWIDE SERVER REVENUE SHARE (Q1 2026, IDC)	PRIMARY BUYER PROFILE
Dell Technologies	PowerEdge XE9712 / XE9812 (GB200, GB300, Vera Rubin NVL72)	Direct liquid cooling, ORv3 rack infrastructure	$16.1B AI server revenue recognized in Q1 FY27; $43B backlog entering FY27 (Source: investors.delltechnologies.com)	16.5% (Source: electronicsweekly.com)	Hyperscalers, large enterprises, sovereign AI
Supermicro	GB200 NVL72 AI SuperCluster; HGX B300/B200	Direct liquid cooling (DLC), pioneer since 2024	$10.2B net sales in fiscal Q3 2026, >80% AI GPU-related (Source: ir.supermicro.com)	7.6% (Source: electronicsweekly.com)	Neoclouds, CSPs, price-sensitive large deployments
HPE	NVIDIA GB200 NVL72 by HPE; Cray XD670	Direct liquid cooling plus hybrid air/liquid options	$5.5B server revenue in Cloud & AI segment, Q2 FY26, up 32.7% (Source: hpe.com)	3.0% (Source: electronicsweekly.com)	Enterprises, sovereign/government, telecom
Lenovo	ThinkSystem SC777 V4 Neptune; GB300 NVL72 (Type 7DJV)	Neptune direct water cooling, up to 40% power reduction (Source: news.lenovo.com)	Included in "Rest of Market" by IDC; 4.6% overall server share (Source: electronicsweekly.com)	4.6% (Source: electronicsweekly.com)	HPC labs, enterprises, research institutions
Foxconn (Hon Hai)	GB200/GB300 NVL72 rack assembly for hyperscalers	Liquid cooling per hyperscaler spec; own 800 VDC Kaohsiung-1 facility (Source: blogs.nvidia.com)	~40% of global AI rack assembly, self-reported (Source: thenextweb.com)	Included in ODM Direct, 50.2% of market (Source: electronicsweekly.com)	Meta, Microsoft, Google, Oracle, hyperscaler-direct
Quanta / QCT	MGX-based S74G series; hyperscaler custom racks	Liquid cooling per customer spec	NT$809.22B (~$25.68B) Q1 2026 revenue, servers 80% of total (Source: taipeitimes.com)	Included in ODM Direct	Hyperscaler-direct, CSPs

Table 1 shows that despite very different corporate structures, from Dell's diversified enterprise IT conglomerate to Foxconn's contract-manufacturing scale, every vendor in this comparison has converged on the same underlying NVIDIA rack architecture and the same liquid-cooling imperative, because Blackwell and Blackwell Ultra GPUs generate too much heat for air cooling alone at rack-scale density. The meaningful differentiation for a buyer is therefore less about raw compute access, since NVIDIA allocates supply across all qualified partners, and more about balance sheet backlog depth, services attachment, and regional support footprint. Dell's **$43 billion** backlog and Foxconn's roughly **40%** share of physical rack assembly represent two different kinds of market power: financial commitment capacity versus manufacturing throughput.

Performance and Benchmarks

Independent, apples-to-apples performance benchmarking across OEM implementations of the same NVIDIA silicon is limited because the underlying GPU, CPU, and NVLink fabric are largely identical across vendors; performance differences stem primarily from cooling efficiency, network fabric choices, and software stack tuning rather than from a different chip. On the MLCommons MLPerf Inference v5.1 suite, HPE's Cray XD670 achieved six number-one results, including in computer vision and LLM chat question-and-answer and text-generation tasks, which HPE cites as evidence its system extracts

more usable throughput from a given generation of NVIDIA silicon than some competing implementations (Source: hpe.com). At the platform level, NVIDIA's own published specifications state that the GB200 NVL72 delivers **30x** faster real-time trillion-parameter LLM inference and **4x** faster LLM training versus the prior H100 generation, alongside **25x** better energy efficiency, figures that apply uniformly to any OEM's compliant GB200 NVL72 implementation because the performance envelope is defined by NVIDIA's reference design rather than by the integrator (Source: nvidia.com).

At the rack level, CoreWeave's own technical documentation for its GB200 NVL72-powered cloud instances confirms **130 TB/s** of total NVLink bandwidth and **13 TB** of high-bandwidth GPU memory per rack, matching NVIDIA's reference specification regardless of whether the underlying hardware came from Dell, Supermicro, or another qualified OEM (Source: docs.coreweave.com). The successor GB300 NVL72 platform, which Dell, HPE, Lenovo, ASUS, and GIGABYTE all now offer, delivers up to a **50x** overall increase in AI factory output performance compared to Hopper-generation platforms, according to NVIDIA's published specifications, again a chip-level improvement rather than an OEM-specific one (Source: nvidia.com). Where OEMs do differentiate measurably is cost-efficiency claims for next-generation platforms: Dell states its PowerRack systems built on NVIDIA's forthcoming Vera Rubin NVL72 platform deliver up to **10x** lower cost per token than Grace Blackwell NVL72 systems for large-scale agentic AI inference, a vendor claim not yet independently verified by third-party benchmarking as of this writing (Source: dell.com). Buyers evaluating vendor performance claims should treat MLPerf submissions, which are audited by MLCommons, as the most rigorous available third-party comparison point, while treating vendor-issued cost-per-token or TCO figures as directional marketing claims pending independent replication.

Data Analysis and Evidence

The overall server market, encompassing both AI-accelerated and traditional systems, reached **$122.6 billion** in worldwide vendor revenue in the first quarter of 2026 according to IDC's Worldwide Quarterly Server Tracker, up **30.4%** year over year (Source: electronicsweekly.com). Within that total, non-x86 servers, a category dominated by NVIDIA GPU and custom ASIC systems, reached **$58.7 billion**, up **107.6%** year over year and now representing **47.9%** of total market revenue, closing in on the traditional x86 segment, which actually declined **2.9%** year over year to **$63.9 billion** as component supply constraints, particularly in DRAM and NAND flash, limited shipment volumes (Source: electronicsweekly.com (Source: electronicsweekly.com). GPU-accelerated servers specifically generated **$68.9 billion**, or **56.2%** of total market revenue, up **24.8%** year over year, while other accelerated (non-GPU, largely ASIC) servers surged **122.1%** to **$17.7 billion** (Source: electronicsweekly.com).

TrendForce's separate research, published in January 2026, forecasts global AI server shipments will grow more than **28%** year over year in 2026, with total global server shipments including AI servers accelerating to **12.8%** growth, driven by the combined capital expenditures of the top five North American cloud service providers (Google, AWS, Meta, Microsoft, and Oracle), which TrendForce expects to increase **40%** year over year in 2026 (Source: trendforce.com (Source: trendforce.com). TrendForce also projects GPU-based systems will remain the leading AI server category at **69.7%** of shipments in 2026, while ASIC-based systems, reflecting custom silicon from Google, Meta, and other hyperscalers, are expected to rise to **27.8%** of shipments, the highest share since 2023 (Source: trendforce.com (Source: trendforce.com). This is a meaningful discrepancy worth noting honestly: TrendForce's forecast of **28%** AI server shipment growth versus IDC's measured **30.4%** overall server market revenue growth reflect different methodologies (unit shipment forecast versus vendor revenue actuals) and different scope (AI servers only versus the total server market), so the two figures are not directly comparable but both point in the same directional trend of continued rapid expansion.

Table 2 below consolidates the disclosed AI-related quarterly financial signals across the major branded OEMs, drawn directly from company filings and press releases, to give buyers a sense of relative financial momentum and disclosed backlog depth heading into the second half of 2026.

COMPANY	MOST RECENT QUARTER (2026)	AI/SERVER-RELATED REVENUE	YEAR-OVER-YEAR GROWTH	DISCLOSED BACKLOG / ORDERS SIGNAL
Dell Technologies	Q1 FY27 (ended ~May 2026)	$16.1B AI server revenue recognized; $29B total ISG	181% (ISG) (Source: fool.com)	$24.4B new AI orders booked; $51.3B backlog reported (Source: fool.com)
Supermicro	Fiscal Q3 2026 (ended Mar. 2026)	$10.2B total net sales, >80% AI GPU-related (Source: ir.supermicro.com)	123% (Source: ir.supermicro.com)	Management cited record backlog, deferred revenue tied to site readiness (Source: fool.com)
HPE	Fiscal Q2 2026 (ended Apr. 2026)	$7.7B Cloud & AI segment; $5.5B server sub-segment (Source: hpe.com)	22.9% (segment); 32.7% (server) (Source: hpe.com)	"Pipeline remains multiples of the current backlog, which is record-breaking" (Source: investors.hpe.com)
NVIDIA (for reference)	Fiscal Q1 2027 (ended Apr. 2026)	$75.2B Data Center revenue (Source: investor.nvidia.com)	92% (Source: investor.nvidia.com)	Partner data centers over 10MW nearly doubled to more than 80 sites (Source: fool.com)
Hon Hai (Foxconn)	Q1 2026	Cloud and Networking ~50% of NT$2.12 trillion revenue (Source: honhai.com)	29% (total revenue) (Source: honhai.com)	AI rack shipments to more than double for full year 2026 (Source: honhai.com)

Interpreting *Table 2*, Dell's disclosed **$51.3 billion** backlog reported in its fiscal first quarter of 2027 call represents the single largest forward revenue commitment among branded OEMs covered here, though HPE and Supermicro both describe backlog as record-setting without disclosing comparably precise dollar figures in their public materials, an important limitation for any buyer trying to benchmark delivery timelines purely from public disclosures (Source: fool.com). NVIDIA's own upstream growth rate, **92%** year-over-year Data Center revenue growth, exceeds every downstream OEM's disclosed AI segment growth rate, underscoring that the constraint on the entire ecosystem remains GPU allocation rather than assembly capacity: NVIDIA reported that the number of partner data centers exceeding 10 megawatts of capacity nearly doubled in one year to more than **80 sites** globally (Source: fool.com). Rising per-rack prices reinforce this allocation dynamic: analysts at Morgan Stanley and Bernstein estimated NVIDIA's forthcoming Vera Rubin NVL72 rack could cost between **$7.8 million** and **$9.1 million** per unit by mid-2026, up sharply from Blackwell-era pricing, driven substantially by HBM4 memory costs projected to triple from roughly **$16.6 per gigabyte** to **$53 per gigabyte** by 2027 (Source: gate.com (Source: wccftech.com). For a full 1-gigawatt AI data center deployment, Bernstein estimated total infrastructure investment at **$47.3 billion**, up **17%** from Blackwell-era costs of **$40.5 billion**, even as Vera Rubin's projected **3.5x** compute efficiency gain is expected to improve return on investment per unit of compute capacity (Source: gate.com).

Case Studies and Real-World Examples

xAI Colossus: Dell and Supermicro Build a 200,000-GPU Cluster in 122 Days

Elon Musk's xAI supercomputer, Colossus, located in Memphis, Tennessee, is the clearest public example of two competing OEMs building parallel infrastructure for the same customer. Both Dell Technologies and Supermicro were confirmed as server suppliers by Musk and Dell chairman and CEO Michael Dell in June 2024, with Michael Dell posting on X that Dell was "building a Dell AI factory with @nvidia to power @grok for @xai @elonmusk," while Musk confirmed Supermicro's involvement separately (Source: datacenterdynamics.com). The initial buildout used Supermicro's liquid-cooled 4U servers, each containing eight NVIDIA H100 GPUs, arranged into racks of 512 GPUs, reaching **100,000 GPUs** in just **122 days** from groundbreaking, a construction pace Supermicro's technology partner ServeTheHome called notable "not just for its size but also for the speed at which it was built" (Source: supermicro.com (Source: x.ai). By February 2025, xAI had doubled Colossus to **200,000 GPUs**, and the cluster's own published statistics claim **170 PB/s** of aggregate memory bandwidth and more than **0.5 exabytes** of storage for training data and checkpoints (Source: x.ai (Source: x.ai). As of May 2026, Wikipedia's tracking of the facility notes that Anthropic agreed to rent all compute capacity at the Colossus 1 data center, illustrating how multi-OEM AI factories increasingly serve as wholesale compute capacity for third parties beyond their original commissioning customer (Source: en.wikipedia.org.

CoreWeave and Dell: First-to-Ship Status Across Three GPU Generations

CoreWeave, the AI-focused cloud provider, has become Dell's most publicly documented reference account for staying ahead of each NVIDIA rack generation. Dell first shipped thousands of PowerEdge XE9860 servers with NVIDIA H100 GPUs to CoreWeave in December 2023, then became the first vendor to ship PowerEdge XE9712 racks with GB200 NVL72 in December 2024, then the first to ship GB300 NVL72 in July 2025, and most recently the first to ship systems built on NVIDIA's next-generation Vera Rubin platform in June 2026, delivering PowerEdge XE9812-based PowerRack systems (Source: investors.delltechnologies.com (Source: investors.delltechnologies.com (Source: dell.com (Source: dell.com). CoreWeave separately became the first cloud provider to make GB200 NVL72-based instances generally available in February 2025, describing its buildout as scalable to **110,000 GPUs**, using NVIDIA Quantum-2 InfiniBand at **400 Gbps** per GPU (Source: sdxcentral.com). This case demonstrates how a single AI-native cloud buyer can systematically extract "first-to-ship" positioning from an OEM partner across multiple silicon generations, a pattern of continuous co-engineering that smaller enterprise buyers are unlikely to replicate.

Foxconn's Hon Hai Kaohsiung Super Computing Center: Taiwan's Largest AI Supercomputer

Foxconn's own showcase deployment, the Hon Hai Kaohsiung Super Computing Center, revealed in 2024 at Hon Hai Tech Day, is built around NVIDIA's Blackwell architecture and the GB200 NVL72 platform, comprising a total of **64 racks** and **4,608 Tensor Core GPUs**, with an expected performance exceeding **90 exaflops** of AI performance, which NVIDIA's own blog describes as making it "easily the fastest in Taiwan" (Source: blogs.nvidia.com (Source: blogs.nvidia.com). Foxconn plans to use the system for cancer research, LLM development, and smart-city applications, and Foxconn Vice President James Wu described it as "one of the most powerful in the world, representing a significant leap forward in AI computing and efficiency" (Source: blogs.nvidia.com). Full deployment was targeted for 2026, illustrating how an ODM that spends most of its time assembling other companies' branded racks is simultaneously building sovereign-scale AI capacity domestically, partly to demonstrate its own engineering capability to hyperscaler customers evaluating ODM-direct purchasing (Source: blogs.nvidia.com).

KDDI and HPE: A Sovereign AI Data Center for the Japanese Telecom Market

Japanese telecommunications operator KDDI Corporation and HPE announced in June 2025 that they would jointly open the Osaka Sakai Data Center by early 2026, deploying a rack-scale system built on the NVIDIA GB200 NVL72 platform and constructed by HPE, using hybrid air- and liquid-cooling technology to reduce the facility's environmental footprint (Source: hpe.com (Source: hpe.com). KDDI intends to offer the resulting compute capacity through WAKONX, its AI-era business platform, to support startups and enterprises training large language models and generative AI systems in Japan (Source: hpe.com). This case demonstrates the OEM value proposition for regional telecom operators that need a trusted systems integrator with liquid-cooling expertise rather than a pure hardware transaction, since HPE President and CEO Antonio Neri framed the deal around "delivering powerful computing capabilities" rather than GPU count alone (Source: hpe.com).

Microsoft Fairwater: A Hyperscaler-Built AI Factory at Gigawatt Scale

Microsoft's Fairwater data center campus in Mount Pleasant, Wisconsin, illustrates the largest hyperscaler category of buyer, one that increasingly designs its own infrastructure while sourcing components from the broader NVIDIA OEM and ODM ecosystem rather than purchasing turnkey racks from a single branded OEM. CEO Satya Nadella described Fairwater on social media as "the world's most powerful AI data center," which "will bring together hundreds of thousands of GB200s into a single seamless cluster," and the site went live ahead of schedule in mid-2026 after roughly two years of construction (Source: wccftech.com (Source: datacenterdynamics.com). Microsoft has described the project as representing "tens of billions of dollars of investments and hundreds of thousands of cutting-edge AI chips," connected to Microsoft's broader global cloud of more than 400 datacenters in 70 regions (Source: blogs.microsoft.com). This case illustrates the ceiling of the OEM comparison: the very largest hyperscalers increasingly bypass fully branded OEM racks in favor of custom designs built with ODM partners such as Foxconn and Quanta, meaning the OEM competitive dynamics detailed elsewhere in this report apply most directly to enterprises, sovereign buyers, mid-sized clouds, and neoclouds rather than to the top four or five hyperscalers, which now represent their own distinct procurement category.

Implications and Future Directions

The near-term trajectory of the NVIDIA AI server OEM market is being reshaped by three forces converging simultaneously. First, NVIDIA itself is capturing more of the value chain by shipping increasingly complete rack systems rather than components for OEM integration, a shift Tom's Hardware describes as NVIDIA "moving closer to shipping entire full-scale systems," which compresses OEM assembly margins even as unit prices rise toward **$8.8 million** or more per Vera Rubin rack (Source: tomshardware.com). Second, the industry-wide transition to 800-volt direct current (VDC) power architecture for gigawatt-scale AI factories is drawing in an even broader ecosystem of partners beyond the traditional server OEMs, with more than **20**

silicon, power system, and cooling partners, including Analog Devices, Infineon, Schneider Electric, Siemens, and Vertiv, participating alongside the core server vendors in NVIDIA's Kyber rack architecture transition planned for around 2027 (Source: blogs.nvidia.com (Source: blogs.nvidia.com). Third, sovereign AI initiatives are becoming a structurally distinct and insulated demand layer: IDC noted that government-directed programs to build nationally controlled AI compute infrastructure now span more than **40 countries**, a category "largely insulated from commercial budget cycles," which favors OEMs like HPE and Dell with established government and public-sector sales channels over pure ODM-direct suppliers (Source: electronicsweekly.com).

Memory pricing represents a growing risk to the entire OEM ecosystem's economics. Bernstein's June 2026 analysis projected HBM4 memory prices could triple from approximately **$16.6 per gigabyte** to **$53 per gigabyte** by 2027 as Vera Rubin ramps into volume production, which would push a single 1-gigawatt AI data center's total infrastructure cost to an estimated **$47.3 billion**, a **17%** increase over Blackwell-generation costs (Source: wccftech.com (Source: gate.com). IDC separately flagged DRAM and NAND flash supply constraints as the principal cap on near-term non-accelerated server growth, a dynamic that could spill over into AI server component availability as memory manufacturers prioritize HBM allocation for AI accelerators over conventional DRAM (Source: electronicsweekly.com). For buyers, this suggests that locking in supply commitments and backlog priority earlier, rather than optimizing purely for unit price, will likely remain the dominant purchasing strategy through 2027, since IDC expects supply normalization to progress only gradually through 2027 as new fabrication capacity comes online (Source: electronicsweekly.com).

Longer term, the growing share of ASIC-based AI servers, forecast by TrendForce to reach **27.8%** of AI server shipments in 2026, the highest since 2023, represents a structural alternative to the NVIDIA OEM ecosystem entirely, as hyperscalers like Google expand in-house Tensor Processing Unit (TPU) silicon and increasingly sell that capacity to external customers such as Anthropic (Source: trendforce.com (Source: trendforce.com). This trend does not directly threaten the branded OEM and ODM vendors profiled in this report in the near term, since Google, Meta, and similar hyperscalers design and largely self-assemble their own ASIC infrastructure outside the traditional OEM channel, but it does cap the addressable market growth rate for NVIDIA-based OEM rack sales over the multi-year horizon.

Frequently Asked Questions (FAQs)

Who builds NVIDIA AI server racks? No single company builds all NVIDIA AI racks. Branded OEMs Dell, Supermicro, HPE, and Lenovo sell fully supported systems to enterprises and mid-sized cloud providers, while Taiwanese original design manufacturers Foxconn (Hon Hai), Quanta Cloud Technology, and Wiwynn assemble the majority of racks purchased directly by the largest hyperscalers such as Meta, Microsoft, Google, and Oracle (Source: blogs.nvidia.com (Source: electronicsweekly.com).

What is the best NVIDIA GB200 NVL72 OEM? There is no single "best" vendor because all qualified OEMs ship functionally identical NVIDIA reference architecture. Dell offers the largest disclosed backlog and fastest "first-to-ship" track record with CoreWeave, Supermicro offers the fastest liquid-cooling delivery lead times and lowest sticker price positioning, and HPE offers the deepest enterprise services and sovereign AI integration experience through its Cray heritage (Source: investors.delltechnologies.com (Source: theregister.com (Source: hpe.com).

What is the difference between Supermicro and Dell AI servers? Supermicro sells hardware-centric, highly configurable Server Building Block Solutions with industry-leading liquid-cooling deployment speed, generating **80%** of its revenue from AI GPU platforms, while Dell bundles rack-scale hardware with enterprise financing, professional services, and a broader AI Data Platform software stack, and posts a larger disclosed order backlog (Source: ir.supermicro.com (Source: siliconangle.com).

What NVIDIA AI infrastructure does HPE offer? HPE offers "NVIDIA GB200 NVL72 by HPE" and the announced "NVIDIA Vera Rubin NVL72 by HPE" rack-scale systems, plus its own Cray XD670 AI training server, which posted six number-one results in the MLPerf Inference v5.1 benchmark suite (Source: hpe.com (Source: hpe.com).

Does Foxconn build NVIDIA GB200 NVL72 racks? Yes. Foxconn (Hon Hai Precision Industry) is described by NVIDIA as a chassis and rack assembly partner for GB200 NVL72 systems alongside Supermicro and Chenbro, and Bloomberg describes Hon Hai directly as "Nvidia Corp.'s server assembly partner," with the company itself claiming roughly **40%** of global AI rack assembly volume (Source: blogs.nvidia.com (Source: fortune.com (Source: thenextweb.com).

What is NVIDIA's market share among AI server OEMs? This question is often confused with server vendor market share; NVIDIA is the GPU supplier, not an OEM. Among branded server OEMs selling NVIDIA-based systems, IDC's Q1 2026 tracker shows Dell at **16.5%**, Supermicro at **7.6%**, Lenovo at **4.6%**, IEIT Systems at **3.3%**, and HPE at **3.0%** of total worldwide server revenue, with unbranded "ODM Direct" vendors collectively holding **50.2%** (Source: electronicsweekly.com).

Conclusion

As of July 2026, the market for NVIDIA AI server OEMs has consolidated around a clear structure: a handful of branded system integrators, Dell, Supermicro, HPE, and Lenovo, compete for enterprise, sovereign, and mid-market cloud business by wrapping identical NVIDIA silicon in differentiated liquid cooling, services, and financing, while Taiwanese original design manufacturers Foxconn, Quanta, and Wiwynn continue to assemble the majority

of racks purchased directly by the largest hyperscalers. Dell currently leads on disclosed financial scale, with a **16.5%** worldwide server revenue share and a backlog exceeding **$51 billion** entering the second half of its fiscal year, while Supermicro leads on liquid-cooling deployment speed and Foxconn leads on raw manufacturing throughput, claiming roughly **40%** of global AI rack assembly. HPE and Lenovo trail on volume but compete effectively for services-intensive sovereign and enterprise deals where their liquid-cooling pedigree and integration depth carry a premium.

For a buyer evaluating this market, the practical decision rarely hinges on which vendor has access to better NVIDIA technology, since MGX and the GB200/GB300/Vera Rubin NVL72 reference architectures are open to all qualified partners and NVIDIA allocates supply across the ecosystem rather than favoring one OEM. Instead, the decision hinges on delivery timeline certainty given industry-wide component constraints, the buyer's need for integration services versus willingness to self-integrate, regional support and financing requirements, and each vendor's demonstrated backlog and manufacturing capacity relative to the buyer's own deployment timeline. With AI server shipments forecast to grow more than **28%** in 2026 and HBM memory costs threatening to push per-rack prices toward **$9 million** or higher for the next-generation Vera Rubin platform, the OEM landscape profiled in this report is likely to keep shifting quickly, and any comparison should be revisited against each vendor's most recent quarterly disclosures before a final purchasing decision is made.

Tags: nvidia ai server oem, supermicro vs dell, hpe nvidia ai server, gb200 nvl72, foxconn nvidia rack, ai server manufacturers, nvidia server market share, gb300 nvl72, ai infrastructure oem

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will GPUSmith or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.