

# NVIDIA GB200 Supply Chain: The Global Ecosystem Explained

12/22/2025 • 40 min read

nvidia gb200

supply chain

semiconductor manufacturing

ai hardware

blackwell architecture

cowos packaging

hbm3e

tsmc



[Revised April 18, 2026] Updated to reflect the GB200 NVL72 production ramp through Q1 2026, the GB300 (Blackwell Ultra) rack transition, TSMC CoWoS capacity expansion, and the ongoing shift toward panel-level fan-out packaging.

## Executive Summary

The NVIDIA GB200 (Grace/Blackwell) high-end AI accelerator represents one of the most complex and demanding hardware products ever assembled. Delivering terascale-and-beyond compute for next-generation AI workloads, a single GB200-based system (e.g. the GB200 NVL72 rack) integrates tens of GPUs, dozens of CPUs, and specialized interconnect, all with power draw measured in hundreds of kilowatts. As a result, **building a GB200 system requires an extremely large, global supply chain encompassing dozens to hundreds of distinct suppliers**. These include major silicon foundries, memory makers, advanced packaging houses, substrate fabricators, connector and cable manufacturers, board assemblers, coolant and power-electronics vendors, and numerous other component suppliers. Our analysis shows that **over 50 unique subcomponent categories** must be sourced for a GB200 rack (<sup>[1]</sup> [newsletter.semianalysis.com](https://newsletter.semianalysis.com)), with multiple potential vendors in each category. Major players span from TSMC (for GPU fabrication and CoWoS packaging) and SK Hynix/Micron (for HBM3e memory) to companies like Amphenol (for NVLink connectors), Foxconn (as a system integrator), and specialized OSAT firms (Amkor, ASE, Powertech, etc.). Moreover, recent reports highlight that even Tier-1 materials (e.g. substrates and fan-out laminates) are under strain, forcing Nvidia and its partners to adopt new packaging standards (such as panel-level fan-out) and to develop U.S. packaging capacity (<sup>[2]</sup> [wccftech.com](https://www.wccftech.com)) (<sup>[3]</sup> [www.datacenterdynamics.com](https://www.datacenterdynamics.com)).

This report provides a **comprehensive examination of the GB200 supply chain**, covering the chip architecture, fabrication and packaging process, memory and substrate suppliers, board and system integration, cooling and power subsystems, networking and interconnect components, OEM and hyperscale integration, and prevailing bottlenecks. We draw on industry analyses, technical disclosures, and market reports to quantify and name many of the companies involved. For example, a detailed SemiAnalysis breakdown explicitly identifies “over 50 different subcomponents of the GB200 rack” with distinct supply channels (<sup>[1]</sup> [newsletter.semianalysis.com](https://newsletter.semianalysis.com)). We also analyze expert cost models (e.g. Epoch AI’s BOM model) and official datasheets to illustrate how major cost centers (such as HBM memory and packaging) dominate the bill of materials (<sup>[4]</sup> [epoch.ai](https://epoch.ai)) (<sup>[5]</sup> [epoch.ai](https://epoch.ai)). Real-world case studies of OEMs and cloud providers (e.g. Foxconn’s server division, ZT Systems, AWS UltraServers) highlight the roles of system integrators and hyperscale customers in sourcing and assembling GB200 systems (<sup>[6]</sup> [www.theregister.com](https://www.theregister.com)) (<sup>[7]</sup> [ztsystems.com](https://ztsystems.com)) (<sup>[8]</sup> [aws.amazon.com](https://aws.amazon.com)). Throughout, we cite numerous sources on supply-chain constraints (digitimes, Wccftech, TheRegister, etc.) to show how CoWoS packaging and HBM supply remain limiting factors, and how the industry is responding.

In sum, **the GB200 supply chain is vast and multinational**. Beyond NVIDIA’s own IP and TSMC’s fabs, it includes Taiwanese substrate firms (Ibiden, Unimicron, Nan Ya PCB), U.S. advanced-packaging partners (Amkor), memory suppliers (SK Hynix, Micron), global electronic component companies (Murata, Texas Instruments, Infineon, etc.), and many ODM/OEMs (Foxconn/Hon Hai, Quanta, Supermicro, Dell, HPE, etc.). Each rack-scale system effectively constitutes an “AI supercomputer” whose construction depends on coordination among hundreds of vendors. We systematically document these layers below, providing tables of key components vs suppliers, data on production volumes and bottlenecks, and expert commentary from industry sources. The result is an in-depth report on the **massive ecosystem of suppliers and vendors** behind NVIDIA’s GB200.

## Introduction and Background

The rapid rise of [large-scale AI](#) has intensified demand for ever-more-powerful compute engines. NVIDIA, as a market leader in [AI accelerators](#), has continually pushed GPU performance with successive architectures (e.g. Volta, Pascal, Ampere, Hopper) and now Blackwell (the architecture behind “Grace Blackwell” GB200). The GB200 design couples NVIDIA’s new Blackwell Tensor Core GPUs (for AI compute) with the Grace CPU into an integrated “superchip” for data

center deployment (<sup>[9]</sup> [developer.nvidia.com](#)) (<sup>[10]</sup> [www.nvidia.com](#)). In a first demonstration, NVIDIA announced the GB200 NVL72 rack in early 2024, an entire cabinet containing 36 GB200 compute cards (totaling 72 GPUs and 36 CPUs) interconnected by NVLink into a single massive system (<sup>[11]</sup> [www.nvidia.com](#)) (<sup>[9]</sup> [developer.nvidia.com](#)). The NVL72 achieves unprecedented scale: 360 PFLOPS of FP8 throughput and over 13 TB of HBM3e memory in one liquid-cooled rack (<sup>[12]</sup> [aws.amazon.com](#)) (<sup>[10]</sup> [www.nvidia.com](#)). Such scale comes at a cost: each GB200 GPU is rated at roughly 2.4 kW, and the full rack draws on the order of 100–150 kW of power (<sup>[6]</sup> [www.theregister.com](#)). Building these systems demands not only large amounts of semiconductor silicon and memory, but also specialized cooling, power delivery, and interconnect infrastructure on an industrial scale.

**Silicon and Architecture.** The core of the GB200 is a *multi-chip package* containing two Blackwell GPUs and one Grace CPU linked by 900 GB/s NVLink C2C HDL interconnect (<sup>[9]</sup> [developer.nvidia.com](#)). This “Grace Blackwell Superchip” package (called the GB200 module) is in turn assembled onto a server board. Each compute tray in a NVL72 rack holds two of these modules (2 Grace CPUs + 4 GPUs) (<sup>[13]</sup> [developer.nvidia.com](#)). In total, the NVL72 rack houses 36 such modules (72 GPUs, 36 72-core CPUs) in 18 1U servers, all bridged by NVLink switch systems to form a single GPU domain (<sup>[6]</sup> [www.theregister.com](#)) (<sup>[8]</sup> [aws.amazon.com](#)). The GPUs themselves are built on TSMC’s latest node (3nm/5nm), while the Grace CPU is a fully custom Arm-based chip (fabbed on TSMC’s 5nm process). Crucially, the Blackwell GPU employs eight 24GB stacks of the new **HBM3e memory** (total 192 GB per GPU) (<sup>[5]</sup> [epoch.ai](#)) (<sup>[14]</sup> [www.anandtech.com](#)). This is a dramatic increase from the previous generation (**H100 GPU** used roughly 80–95 GB of HBM) and makes memory a dominant part of the design.

**Supply Chain Complexity.** The GB200 is effectively not a single processor but a rack-scale supercomputer in a box. Its supply chain encompasses every layer of the electronics ecosystem from raw silicon to housings. On one end, NVIDIA provides the chip design and some proprietary ASICs (the GPUs and NVSwitch chips), but outsources wafer fabrication to TSMC. On the other, hyperscale server OEMs (Foxconn, Supermicro, Dell, etc.) assemble the boards, power systems, and racks for end customers. In between are dozens of specialty vendors: memory manufacturers (SK Hynix, Micron), packaging houses (Amkor, ASE, Powertech), substrate companies (Unimicron, Ibiden, Nanya, etc.), board fabricators, passive component manufacturers (Murata, TDK, Yageo, etc.), connector and cable makers (Amphenol, TE Connectivity, US Conec, etc.), cooling companies (CoolIT, Enago, etc.), and many more. Stakeholders range from multinational chip foundries to local laminated-glass substrate suppliers. Summing all tiers, **the number of distinct suppliers is easily in the hundreds**. One industry analysis explicitly notes that assembling a GB200 rack involves “over 50 different subcomponents” with multiple vendors for each (<sup>[1]</sup> [newsletter.semianalysis.com](#)).

**Historical Context.** To appreciate the scale, it is instructive to consider GPU evolution. Early NVIDIA GPUs in the GTX/Tesla series had a single monolithic silicon block with off-chip GDDR/small HBM stacks. By the time of Ampere/Hopper (2020–2022), NVIDIA was already using 7nm GPUs with HBM2e memory (e.g. H100 with 80 GB of HBM). Each new generation added more memory (H100 had five stacks ⇒ 80 GB; Blackwell has 8×24 GB = 192 GB) and higher-performance packaging (CoWoS). The shift to GB200 has amplified this trend: it uses the most advanced packaging (CoWoS-L with multi-die stacking, now moving to panel-level FO) and unprecedented memory. As a result, suppliers who catered to H100 production (HBM makers, substrate fabs, OSAT labs) now face even far greater demand. Well-known bottlenecks from the Hopper era – limited HBM supply and CoWoS packaging capacity – have continued to limit Blackwell rollout (<sup>[2]</sup> [wccftch.com](#)) (<sup>[15]</sup> [www.theregister.com](#)), driving NVIDIA to explore alternative solutions (e.g. panel-level fan-out and onshoring of packaging facilities).

**Current State.** As of early 2026, Blackwell (GB200) chips are in full mass production, and NVIDIA has begun the transition to the **GB300 NVL72** (Blackwell Ultra) platform, announced at GTC 2025 and ramping through the first half of 2026 with 288 GB of HBM3e per GPU and higher FP4 throughput than GB200. TSMC’s 3nm fab in Arizona has produced the first U.S. wafer (heralding onshoring of AI chipmaking) (<sup>[16]</sup> [blogs.nvidia.com](#)) (<sup>[17]</sup> [www.datacenterdynamics.com](#)), although full 2-chip/CPU packages still rely on Taiwanese OSAT lines (<sup>[3]</sup> [www.datacenterdynamics.com](#)) (<sup>[18]</sup> [www.theregister.com](#)). NVIDIA has begun shipping integrated NVL72 racks (e.g. DGX systems) in small volume and major cloud players (AWS, Microsoft, Google, Meta) are deploying Blackwell nodes. Analysts expect surging demand: TrendForce projects Blackwell (B200/GB200) to represent over 80% of NVIDIA’s high-end GPU shipments by 2025 (<sup>[19]</sup> [www.trendforce.com](#)), and forecast shipments to ramp from a few hundred thousand in 2024 to on the order of 1.5–2

million units in 2025 (<sup>[20]</sup> [wccftech.com](#)) (<sup>[19]</sup> [www.trendforce.com](#)). Key OEMs (Foxconn, Supermicro, Dell, HPE, GIGABYTE, Inspur, etc.) have all announced or are preparing Blackwell-based servers. This market enthusiasm has led to a “domino effect” in the supply chain: Taiwanese server manufacturing giants (Hon Hai/Foxconn, Accton, Chenbro, Auras) are poised to reap growth by supplying major subsystems (servers, network switches, chassis, liquid-cooling systems) for the Blackwell wave ([cloudnews.tech](#)) ([cloudnews.tech](#)).

**Structure of this Report.** We explore the GB200 supply chain in depth. The **next section** details the GB200 hardware architecture (Chips, NVSwitch, NVL72 rack). We then examine **chip fabrication and packaging**, highlighting the roles of TSMC and OSAT partners. Following that, sections cover **memory and substrate suppliers** (HBM3e, coWoS substrates), **system integration and board assembly** (PCBs, OEMs), **power and cooling subsystems**, and **interconnect and networking components**. Throughout, we integrate data on production volumes, cost breakdowns, and supply constraints. We include **tables** summarizing major component categories vs example suppliers, and an estimated **BOM cost breakdown** for the GB200 GPU. **Case studies** (e.g. Foxconn’s DGX gear, ZT Systems’ ACX200, AWS UltraServers) illustrate real-world examples of how the ecosystem is put together. Finally, we discuss the implications for the industry, including ongoing supply-chain bottlenecks, new packaging innovations, and the future of AI hardware manufacturing.

## NVIDIA GB200 Architecture and System Overview

The GPU architecture at the heart of GB200 is NVIDIA’s **Blackwell** tensor-core GPU, paired with the **Grace** CPU in a combined module. Each *GB200 module* (aka “Grace Blackwell Superchip”) consists of two Blackwell GPU dies and one Grace CPU die, interconnected by NVIDIA’s proprietary NVLink-C2C interconnect (900 GB/s bidirectional) (<sup>[9]</sup> [developer.nvidia.com](#)). Blackwell is a third-generation tensor-core design (successor to Hopper), fabricated on TSMC’s most advanced nodes. Each Blackwell GPU wraps eight 24GB stacks of HBM3e memory (total 192 GB per GPU) (<sup>[5]</sup> [epoch.ai](#)) (<sup>[14]</sup> [www.anandtech.com](#)), making the memory subsystem extremely large.

**NVIDIAScale System:** In practice, NVIDIA deploys GB200 in very large multi-GPU servers. The flagship *NVL72* system uses 36 GB200 modules (72 GPUs, 36 CPUs) in a single rack. NVIDIA’s own datasheet describes NVL72 as a “single massive 72-GPU rack” that can overcome communication bottlenecks” using NVLink and dedicated liquid cooling (<sup>[10]</sup> [www.nvidia.com](#)). The architecture not only stitches GPUs on a board but also links entire boards via NVLink *switch fabrics*: NVL72 employs two NVLink switch trays per rack to form one unified NVLink domain (<sup>[13]</sup> [developer.nvidia.com](#)) (<sup>[7]</sup> [ztsystems.com](#)). Each NVL72 server (often called a DGX or similar) contains two such boards in 1U, and 18 servers (36 modules) fill the rack (<sup>[6]</sup> [www.theregister.com](#)).

Key performance figures highlight the scale: each NVL72 rack delivers ~360 petaflops of FP8 compute and carries 13.4 TB of HBM3e memory (<sup>[8]</sup> [aws.amazon.com](#)). Amazon, for example, explicitly cites “up to 72 NVIDIA Blackwell GPUs interconnected using fifth-generation NVLink — all functioning as a single compute unit” for its EC2 **P6e-GB200 UltraServer** offering (<sup>[8]</sup> [aws.amazon.com](#)). Even the 1U server variants (e.g. “P6-B200” instances) pack 4–8 GPUs with up to 96 GB or 192 GB of HBM, enabling massive models at hyperscalers. In all cases, these systems require advanced infrastructure: direct-to-chip liquid cooling loops, extremely high-voltage power delivery, and bespoke networking.

Summarizing the components of the full GB200 system (e.g., one NVL72 rack) illustrates the breadth of suppliers needed:

- **Compute Modules:** 36 GB200 modules per rack, each containing 2 Blackwell GPUs + 1 Grace CPU (<sup>[9]</sup> [developer.nvidia.com](#)) (<sup>[6]</sup> [www.theregister.com](#)). (NVIDIA designs GPUs; fabrication by TSMC.)
- **Memory:**  $36 \times 2 \times 192 \text{ GB} = 13.8 \text{ TB}$  HBM3e total. (Memory chips from SK Hynix/Micron.)

- **NVLink Backplane and Cable System:** 72 GPU connections plus switch fabrics. (Connectors/cables from Amphenol, Samtec, US Conec, etc. <sup>[21]</sup> newsletter.semianalysis.com) <sup>[22]</sup> newsletter.semianalysis.com.)
- **PCIe Interconnect:** Up to 8 PCIe Gen6 slots for external NICs (e.g. 800G Ethernet or HDR InfiniBand). (NICs from Broadcom/Mellanox, connectors from Samtec.)
- **Switch ASICs:** 2× NVIDIA NVSwitch per rack (to link racks in NVL72x2), or 4× per two-rack system <sup>[23]</sup> newsletter.semianalysis.com). (NVSwitch chips by NVIDIA, CS.)
- **Power Delivery:** Rack busbars (48 VDC) driving 12V VRMs on blades <sup>[24]</sup> newsletter.semianalysis.com). (Busbar and high-current connectors from companies like Alpha & Omega Semiconductor, TE Connectivity (RapidLock), etc.)
- **Cooling System:** Custom cold-plate assemblies, pumps, quick-disconnects, fitting lines. (Chillers/CDUs by CoolIT, Asetek, Submer; coolant by 3M or HPE Synap.)
- **Chassis and Rack:** 18×1U server chassis, rack, top-of-rack switch space, etc. (Chassis by Foxconn, Supermicro, Chenbro, etc.; rack/PDU frame by Vertiv, Schneider.)
- **Auxiliary Control:** Baseboard management controllers (ASPEED AST2500 BMCs, sensors) for each server.

The system-level integration itself was historically the domain of NVIDIA and its ODM partners. ZT Systems, for example, announced an **ACX200** rack integrating GB200 modules into a liquid-cooled server with NVLink switch trays <sup>[7]</sup> ztsystems.com). Similarly, Foxconn's server division claimed to be first in line to ship GB200 systems by late 2024 <sup>[25]</sup> www.theregister.com) <sup>[26]</sup> www.theregister.com). These OEMs manage procurement from many part vendors and do the final assembly.

## Chip Fabrication and Packaging

At the core of the GB200 supply chain is the semiconductor manufacturing process. NVIDIA designs the Blackwell GPU and Grace CPU, but **outsources fabrication to TSMC**. This begins at the wafer level: the first NVIDIA Blackwell wafer was produced on TSMC's cutting-edge Arizona 3nm fab in October 2025 <sup>[17]</sup> www.datacenterdynamics.com) <sup>[16]</sup> blogs.nvidia.com). This milestone "signals that Blackwell has reached volume production" in the onshore fab <sup>[17]</sup> www.datacenterdynamics.com). However, crucially, the wafer **gross die fabrication** is only the first step. Because Blackwell uses a multi-chip package, each wafer's individual dies (two GPUs + one CPU) must be *assembled together* with HBM memory in advanced packaging processes. Currently, **NVIDIA relies on TSMC's CoWoS (Chip-on-Wafer-on-Substrate) advanced packaging**, which to date exists only in TSMC's Taiwanese sites. As Ming-Chi Kuo and others note, the Arizona-fab wafer still "needs to be shipped to Taiwan for TSMC's CoWoS advanced packaging, only then [is] production of the Blackwell chip [complete]" <sup>[3]</sup> www.datacenterdynamics.com).

Indeed, all of TSMC's current AI-chips packaging facilities are in Taiwan <sup>[27]</sup> www.theregister.com), so NVIDIA must still send wafers to OSAT partners there. A journalist from *The Register* emphasizes that "Nvidia remains reliant on Taiwanese packaging plants to turn those wafers into its most powerful GPUs" <sup>[18]</sup> www.theregister.com). In practice, the GPU dies and memory stacks are aligned on a large organic substrate (or interposer) using CoWoS-L (2.5D fan-out) methodology <sup>[28]</sup> www.theregister.com) <sup>[29]</sup> wccftech.com). The substrate itself—essentially a high-density PCB interconnect—is supplied by specialized manufacturers (see next section). Multiple bonding, underfill, and RDL processes integrate the components into a single package.

This advanced packaging is extremely complex and capacity-limited. CoWoS lines must handle very large-die GPUs with fine pitch interconnects. According to supply-chain analyses, the packaging yield rate is a critical concern: failed wafers waste both the expensive logic dies and HBM stacks <sup>[30]</sup> epoch.ai). NVIDIA's recent delays have been largely attributed to CoWoS bottlenecks. Foxconn and DigiTimes both reported that "*advanced packaging tech used to stitch together the compute dies and HBM3e modules*" presented challenges, exacerbated by "CoWoS capacity [that] remains extremely limited" through 2025 <sup>[15]</sup> www.theregister.com) <sup>[2]</sup> wccftech.com). Thus, expanding packaging capacity is a top priority:

TSMC has agreed to build dedicated CoWoS lines with partners like Amkor, and various firms are investing in onshore OSAT facilities (see *Future Directives* below).

**Major partners:** The key players in packaging include TSMC (packaging operations) and leading OSAT houses:

- **Amkor Technology** (U.S.) – Signed an agreement with TSMC to provide “turnkey advanced packaging and test services” in a new U.S. OSAT fab (slated by 2027–28) (<sup>[31]</sup> [www.datacenterdynamics.com](http://www.datacenterdynamics.com)) (<sup>[27]</sup> [www.theregister.com](http://www.theregister.com)). Amkor is expected to handle US packaging of future NVIDIA chips, including Blackwell and beyond.
- **ASE Group** (Taiwan) – A top OSAT firm (with Powertech/PTI in affiliate) that currently performs CoWoS packaging for various premium chips (including NVIDIA GPUs).
- **Powertech Technology (PTI)** (Taiwan) – Another major substrate/packaging supplier; recent reports mention Powertech’s role in next-gen panel-level fan-out (PFLO) for NVIDIA BLACKWELL (<sup>[32]</sup> [wccftech.com](http://wccftech.com)).
- **Samsung Electronics** (South Korea) – While Samsung is a leader in HBM DRAM, it has not been reported as a supplier of HBM3e for NVIDIA. (NVIDIA’s HBM3E comes primarily from SK Hynix and Micron.)
- **Intel/Moog** – (Intel’s EMIB or Foveros alternatives; not currently used by NVIDIA.)

NVIDIA also contributes specialized chiplets: the **NVIDIA NVSwitch** ASIC used in racks is itself an NVIDIA-designed chip, but likely fabricated/co-packaged with their GPUs; it uses the same CoWoS infrastructure.

In summary, **the chip production supply chain** includes: TSMC 3nm/5nm fab → wafer dicing → advanced packaging by OSAT. The primary suppliers in this chain are TSMC, Amkor, ASE (PTI), and now initiatives like Powertech/InnoLux for new PFLO packaging (<sup>[32]</sup> [wccftech.com](http://wccftech.com)). The number of wafer starts for a full production run is enormous: one article estimates NVIDIA shipping 0.42 million Blackwell GPUs in H2 2024 and 1.5–2 million in 2025 (<sup>[33]</sup> [wccftech.com](http://wccftech.com)), implying on the order of millions of wafers processed and packaged.

## Memory and Substrate Suppliers

A major part of the GB200 BOM is **HBM3e memory**. Each Blackwell GPU contains 8 stacks × 24 GB HBM3e, for a total of 192 GB (<sup>[5]</sup> [epoch.ai](http://epoch.ai)) (<sup>[14]</sup> [www.anandtech.com](http://www.anandtech.com)). HBM3e is the third generation high-bandwidth DRAM, manufactured by only a few suppliers:

- **SK Hynix (South Korea):** Leading HBM vendor; has 16–24 GB HBM3e products. SK Hynix is expected to supply a significant volume for NVIDIA’s AI GPUs (as it did for Hopper).
- **Micron Technology (USA):** Offers 24 GB HBM3e “Gen2” stacks and has explicitly identified NVIDIA as a key customer. Micron’s CEO stated that NVIDIA is “one of its primary customers” for HBM3e (<sup>[34]</sup> [www.anandtech.com](http://www.anandtech.com)). In fact, Micron’s first HBM3e device (8-Hi 24 GB) is already qualified for NVIDIA’s H200 series, and launch in early 2024 (<sup>[14]</sup> [www.anandtech.com](http://www.anandtech.com)) (<sup>[34]</sup> [www.anandtech.com](http://www.anandtech.com)).
- **Samsung Electronics (South Korea):** Also a major HBM player; Samsung has HBM2/HBM2e/3 lines but has not been publicly confirmed to supply HBM3e to NVIDIA.

The competitive situation: As of 2023–2024, Micron was first to ship HBM3e, but SK Hynix and Samsung were already preparing their products. There have been concerns about HBM supply for AI: one report notes that in 2024, Micron’s entire HBM3e capacity was sold to customers (mostly NVIDIA) and even fully booked into 2025 (<sup>[35]</sup> [www.anandtech.com](http://www.anandtech.com)) (<sup>[36]</sup> [www.anandtech.com](http://www.anandtech.com)). Thus, the HBM DRAM supply chain is very constrained, reflecting the handful of players in a niche market. NVIDIA must coordinate orders directly with SK Hynix and Micron (and perhaps Samsung) to secure the billions of dollars of memory needed.

Beyond chips, a key material is the **substrate** or interposer onto which the dies are mounted. NVIDIA uses *organic substrates* in CoWoS packaging. High-end substrates (with many layers of copper) are produced primarily by Taiwanese firms. For example, Yole Intelligence notes that **Unimicron, Ibiden, Nan Ya PCB** and similar companies dominate the IC

substrate market (93% share) (<sup>[37]</sup> [www.edge-ai-vision.com](http://www.edge-ai-vision.com)). These companies fabricate the glass-reinforced laminates and build-up layers needed for chip modules. In years prior, substrate shortages have become choke points; Taiwan-based substrate makers have ramped capacity for AI chips. A Taiwan news report specifically warns that NVIDIA's GB200 is facing delays because "key materials used for chip substrates are in short supply due to insufficient production capacity" (<sup>[38]</sup> [www.digitimes.com](http://www.digitimes.com)). Although [12] (Digitimes) is behind a paywall, its headline and snippet confirm substrate/CoWoS constraints. In response, both substrate makers and TSMC have announced expansions: for instance, in late 2025 TSMC invested heavily in new CoWoS lines and substrate fabs (Erudite Asia on 15 Oct 2025).

**Example:** The [epoch.ai](http://epoch.ai) cost model implicitly includes substrate cost in its "packaging" component. Packaging (inc. substrate cost) plus memory together represent ~66% of a Blackwell GPU's cost (<sup>[4]</sup> [epoch.ai](http://epoch.ai)). With memory roughly \$15/GB (NVIDIA's B200 uses 192 GB → ~\$2.9k memory cost) (<sup>[5]</sup> [epoch.ai](http://epoch.ai)), the remainder (~\$2k–3k) goes to packaging/substrate and yield losses. This underscores that the substrate and packaging materials (from Ibiden/Unimicron/Nan Ya, using materials like Ajinomoto copper/glass, adhesives, etc.) are a major expense, driving intense demand for these suppliers.

Below is a **summary table** of major GB200 components and representative suppliers:

Component Category	Function / Description	Key Suppliers / Vendors
GPU Compute Dies	NVIDIA Blackwell Tensor-Core GPUs	NVIDIA (design); TSMC (3nm/5nm foundry)
CPU Die	NVIDIA Grace 72-core ARM CPU	NVIDIA (design); TSMC (5nm)
HBM3e DRAM Stacks	High-Bandwidth Memory (8x24GB per GPU)	SK Hynix, Micron (major HBM3e suppliers)
Advanced Packaging CoWoS	Multi-chip integration (GPU+GPU+CPU+HBM)	TSMC (CoWoS tech), Amkor, ASE (PTI), Powertech
IC Substrates / Interposers	Organic/glass build-up substrates for CoWoS/FO	Unimicron, Ibiden, Nan Ya, Nanya, Sumco (PCB)
PCBs / Server Boards	Motherboard hosting GB200 modules	Foxconn (Hon Hai), Quanta, Wistron, Inventec, etc.
Connectors & Cables (NVLink)	High-speed NVLink interconnect (GPU ↔ switch, QSFPs, etc.)	Amphenol (Paladin), US Conec (DensiLink), TE Connectivity, Samtec (SkewClear, FireFly)
NVSwitch ASICs	NVIDIA's switch chips for multi-GPU fabrics	NVIDIA (design); TSMC (manufacturing)
Power Delivery (VRMs)	Voltage regulation modules on-board	TI, Infineon, Richtek (VRM controllers), Alphabet FET, TE Connectivity (RapidLock connectors)
Busbars & Harness	Rack-level 48VDC busbar, cables to 12V in servers	Alpha & Omega (copper busbar), TE Connectivity
Cooling Systems	Direct-to-chip liquid cold plates, pumps, quick-disconnect fittings	CoolIT Systems, 3M (coolants, thermal interface), Asetek, Submer, APTEK, Alphacool; Conxall (CPC quick-fits)
Chassis & Rack Cabinets	Server enclosures, racks, power distribution units	Foxconn, Supermicro, Chenbro, Auras (liquid cooling modules), Vertiv, Schneider
Network Adapters (NICs)	Ethernet/InfiniBand connectivity (up to 800GbE)	Broadcom/Mellanox (ConnectX-8 NICs, switches); NVIDIA (Mellanox)
Baseboard Mgmt & Controllers	BMC microcontrollers, sensors	ASPEED (AST2500 BMC), Renesas (MCUs), NXP, Texas Instruments
Capacitors/Inductors/Resistors	Power smoothing & filtering	Murata, TDK, Taiyo Yuden (inductors); AVX, Rubycon, Samsung Electro-Mechanics (caps)
Firmware/Software	System BIOS, NVLink on-chip firmware	NVIDIA, ASPEED (BMC firmware)

Table 1: Major subcomponents of NVIDIA's GB200 system and exemplar suppliers. Each category may involve many vendors.

This non-exhaustive table highlights that virtually every step of the GB200's assembly chain involves companies beyond NVIDIA. Memory stacks alone involve two major DRAM makers; substrates involve multiple Taiwanese PCB firms; board-level components (passives, connectors, chips) draw from global electronics conglomerates; and large ODMs tie it all together. In the next sections we dive deeper into these areas.

## Packaging and Advanced Interconnect

NVIDIA's multi-die package requires a sophisticated integration strategy. The GB200 GPU uses **CoWoS-L (Chip-on-Wafer-on-Substrate, Large)** packaging. This means a large organic substrate carries the silicon dies and HBM stacks. TSMC pioneered CoWoS for GPUs, and NVIDIA is one of its biggest customers. In CoWoS, after logic die and HBM are bonded to a glass carrier, they are embedded in an organic substrate made by Unimicron or Ibiden. The die-to-substrate

connections use copper pillars and TSVs. After stacking, the module is encapsulated, tested, and mounted on the server board.

**NVLink Fabric:** Inter-chip and inter-board connectivity is another packaging aspect. Each GB200 GPU connects to the system via a high-pin-count connector. In NVL72, each GPU plugs into an Amphenol **Ultrapass Paladin** connector (72 differential pairs, 224 Gb/s bandwidth) on the backplane (<sup>[21]</sup> newsletter.semianalysis.com). A corresponding 72-pair Paladin receptacle on the server backplane mates with the GPU. Internally, GPUs connect to NVSwitch chips via **SkewClear EXD Gen2** twinax cables, and between NVSwitch trays NVIDIA uses **OverPass** and **DensiLink** flyover cables (<sup>[22]</sup> newsletter.semianalysis.com). These connectors/cables are extremely high-speed and expensive; SemiAnalysis notes that “much of the cost is not from the cables but from the connectors [which] need to prevent crosstalk” (<sup>[21]</sup> newsletter.semianalysis.com). NVIDIA’s choices here (Amphenol Paladin, DensiLink, SkewClear, etc.) drive orders for those vendors. (Samtec’s FireFly and PACMIC also compete in this space.) Future NVL36×2 setups require additional twinax 1.6 Tbps cables, further increasing cable demand (<sup>[23]</sup> newsletter.semianalysis.com).

**Networking and IO:** Aside from NVLink, GB200 servers include other interconnects. Many designs have **Ethernet or Infiniband NICs**. NVIDIA acquired Mellanox (Broadcom) years ago; GB200 systems often use Broadcom’s 800GBE switches and ConnectX-8 NICs for external network. These are mounted on mezzanines or riser cards. SemiAnalysis points out that high-speed switch cables (e.g. twinax or DAC) to interconnect NVSwitch trays also contribute to BOM cost (<sup>[23]</sup> newsletter.semianalysis.com). Again, connector vendors like Amphenol (for QSFP cages) and cable vendors (like Amphenol/VersaCore or Samtec) benefit.

**Advance to Panel-Level FO:** Due to CoWoS limits, NVIDIA is shifting to **Panel-Level Fan-Out (PFLO)** packaging for Blackwell next year (<sup>[29]</sup> wccftech.com). In PFLO, multiple ICs are embedded in a larger laminate panel, allowing scaling beyond wafer size. This approach uses materials like glass substrates or laminates instead of silicon interposers. WCCFtech reports that suppliers for PFLO are currently scarce, naming **Powertech (PTI)** and **Innolux** (a large TFT LCD manufacturer) as early contenders to package NVIDIA’s chips (<sup>[32]</sup> wccftech.com). If adopted, this would involve new supply chains: Powertech (with experience in substrates) and Innolux (leveraging glass panel capabilities) would become NVIDIA packaging vendors.

In summary, the **advanced packaging and interconnect** stack involves:

- **Die-to-die integration:** TSMC/Amkor/ASE under CoWoS-L or PFLO processes (substrates by Unimicron/Ibiden/Nanya; test services by Amkor/ASE).
- **GPU-to-backplane connectors:** Amphenol Paladin (Ultrapass) 72-pair connectors, TE (Amphenol) coax cables (SkewClear), US Conec DensiLink cables.
- **NVLink switch fabric:** NVIDIA NVSwitch chips (ASICs) plus Amphenol or Celestica-made switch boards.
- **Infiniband/Ethernet ports:** Broadcom/Mellanox NIC chips with QSFP-DD cages (Samtec/Amphenol).
- **Riser connectors and midplane:** Samtec high-speed board-to-board connectors for NVSwitch trays, midplane harnesses, etc.

This layer alone accounts for dozens of supplier relationships. For example, SemiAnalysis estimates that the NVLink backplane assembly by itself costs several thousand dollars per board because of premium connectors (<sup>[21]</sup> newsletter.semianalysis.com). In practice, NVIDIA likely qualifies 2–3 connector/cable vendors for each interface (e.g. Amphenol plus a backup like TE Connectivity). Thus, even this single subdomain brings in multiple vendors.

## Power Delivery and VRMs

Feeding each GB200 module’s ~2700 W to the board requires robust power infrastructure. The NVL72 architecture uses **48 V DC busbars** at the rack level. Busbars (often aluminum or copper conductors) run vertically, bringing 48 V from the room UPS/PDU to each 1U server. In each server, a **power distribution board (PDB)** steps down 48 V to 12 V for the

GPUs/SOC, via large VRM arrays. According to supply-chain reports, each GB200 card receives power through four 12V DC and four *ground* high-current connectors (RapidLock style) around the GPUs (<sup>[24]</sup> [newsletter.semianalysis.com](#)). These TE Connectivity “RapidLock” connectors (or similar) carry ~675 A per cable.

Companies involved here include:

- **Power MOSFETs and controllers:** Texas Instruments, Infineon, UPI (uPI Semiconductor) supply the multi-phase VRM chips and FETs on the server boards. (For example, NVIDIA's past servers have used TI's DrMOS power stages.)
- **Busbar hardware:** Taiwanese companies like APower and Alpha & Omega produce heavy copper busbars and warning distribution assemblies. The connectors and distribution PCBs often come from contract manufacturers.
- **Passives:** High-current inductors (coupled air-chokes by TDK) and capacitors (polymer or ceramic by Murata/NCC) are needed in large quantity on each VRM phase.
- **RapidLock connectors:** TE Connectivity's high-current cable assemblies (or equivalents from Huber+Suhner, Kyocera, etc.) link the busbar to the board.
- **Cable harnesses:** Custom power cables (48V harnesses, PCIe power cables) are sourced from cable suppliers, often made in-house by system assemblers like Foxconn or Quanta.

SemiAnalysis notes that the GB200 card's four 12V connectors (and their huge gauge wires) bundle to the PDB. Each PDB converts 48V → 12V via likely 16+ phase VRMs. Thus, roughly a dozen VRM controllers and 80+ FETs sit per board. Suppliers like TI benefit substantially: a server with 72 GPUs might include hundreds of DrMOS stages from TI.

**Power Infrastructure at Rack Scale:** In addition to the cards, the power distribution units (PDUs) and power shelves factor in. The rack must handle ~100–150 kW. Companies like **Eaton**, **Vertiv (Liebert)**, **Schneider Electric** provide the rack PDU and chillers. Busbar trunking might be supplied by systems integrators in data centers.

Overall, the power delivery sub-system adds **tens of suppliers** for everything from semiconductors to metalwork. While none of these companies are as famous as NVIDIA or TSMC, they are essential: without them, the GB200 modules cannot be powered or controlled.

## OEMs, System Integrators, and End-Users

Once the individual components (chips, memory, boards) exist, final assembly and integration involve large OEMs (Original Equipment Manufacturers) and CSPs (Cloud Service Providers). These organizations source GB200 modules and related parts to build complete servers or clusters.

- **Foxconn (Hon Hai) and Pegatron:** Foxconn's server business (often under Foxconn Industrial Internet) has claimed to be the first to ship GB200 servers (<sup>[26]</sup> [www.theregister.com](#)). Analysts expect Foxconn to build a substantial share of Blackwell racks; one brokerage projects Foxconn could ship 17,000–30,000 racks (40–70% market share) by 2026 ([cloudnews.tech](#)). Foxconn and Taiwanese ODMs like Quanta and Wistron receive GB200 boards or modules from NVIDIA/OEMs and assemble them into branded servers (e.g. Foxconn's “Star Gate” AI servers). They also manage customization (e.g. adding custom cooling loops or integrated cabling).
- **Supermicro, Dell, HPE, Lenovo:** Western OEMs likewise incorporate GB200. Dell announced at NVIDIA GTC 2024 new PowerEdge servers (“B200” nodes) with NVIDIA Blackwell GPUs. HPE offers ProLiant and Apollo systems with Blackwell acceleration (as per HPE community posts). These companies collaborate with NVIDIA on system design; e.g., they decide on chassis layouts, which card slot schemes to use, etc. While they largely function as system integrators, they also have purchasing influence: NVIDIA builds baseline GB200 boards, then these OEMs may buy or license them for their server designs.
- **ZT Systems, Penguin Solutions, Cirrascale, etc.:** Smaller companies that specialize in HPC appliances have also launched Blackwell-based platforms. For instance, ZT Systems (a Microway subsidiary) introduced the **ACX200** rack, a turnkey liquid-cooled platform using GB200 modules and Grace CPUs (<sup>[7]</sup> [ztsystems.com](#)). The AWS blog similarly brands its ultraservers but acknowledges that “the P6e-GB200 UltraServers represent [AWS's] most powerful GPU offering to date” with 72 GPUs per server (<sup>[8]</sup> [aws.amazon.com](#)). These examples illustrate that integrators configure the final product around NVIDIA's raw modules.

- **Cloud Providers (Hyperscalers):** AWS (Amazon), Microsoft Azure, Google Cloud, Meta, and others are major end-users. They partner with ODMs or deploy Nvidia's DGX-branded systems. For example, AWS's UltraClusters integrate hundreds or thousands of GB200 GPUs. The AWS blog emphasizes that their new P6e-GB200 machines have 72 GPUs linked by NVLink and can be liquid-cooled or air-cooled <sup>(18)</sup> [aws.amazon.com](https://aws.amazon.com) <sup>(39)</sup> [aws.amazon.com](https://aws.amazon.com)). These cloud customers influence the supply chain too: they place massive orders that ripple down through NVIDIA, TSMC, and OSAT, affecting supplier demand planning.

**Implications:** The diversity of OEMs means multiple supply channels exist. NVIDIA can sell GB200 modules to Foxconn for one customer, while directly supplying Dell for another. Some OEMs like Dell/HPE might design their own PCBs and simply purchase Blackwell GPUs from NVIDIA. Others (like ZT Systems) might buy GB200 cards and mount them on their own motherboard (as indicated in the ZT press release). In any case, practically **all major server vendors worldwide** are involved. This broad ecosystem further multiplies the number of vendors: for each component in our table above, the customer market is global.

## Case Study: Foxconn and DGX GB200

In mid-2024, *The Register* reported on Foxconn's plans for GB200: Foxconn's AI server unit claimed it would begin shipping "small volumes" of GB200 systems in Q4 2024, ramping in Q1 2025 <sup>(25)</sup> [www.theregister.com](https://www.theregister.com)). According to Foxconn, their system ("DGX NVL72") packs 36 GB200 superchips (72 GPUs + 36 CPUs) into 18 1U servers, delivering 1.44 exaFLOPS (FP4) and 13.5 TB HBM3e total <sup>(6)</sup> [www.theregister.com](https://www.theregister.com)). The article noted that Foxconn would be "the first supplier" to ship GB200 accelerators. This implies that NVIDIA (who co-brands some systems) was relying on Foxconn's manufacturing capabilities to fulfill initial orders.

Foxconn's involvement exemplifies how OEM supplier networks operate:

- Foxconn likely sources GB200 modules from NVIDIA as fully-packaged cards, along with NVIDIA-GPU-based switch trays and other components.
- On the Foxconn side, many parts are Foxconn's own or sourced by them: for example, Foxconn might build the custom fan-cooled liquid-cycles (3M coolant, CoolIT cold plates), assemble the chassis (Foxconn has many chassis production lines), provide the backplane assembly (Amphenol connectors, curated cable looms), and integrate everything into a turnkey DGX server.
- Internal Foxconn supply: Foxconn Electronics might even produce some PCBs or cable assemblies. For parts it does not make, Foxconn coordinates with vendors (e.g. calling Texas Instruments for VRMs, Murata for caps, etc.).

The Register piece also highlights supply issues: it says NVIDIA and TSMC encountered "challenges with the advanced packaging tech" for Blackwell, and that CoWoS capacity remained limited <sup>(15)</sup> [www.theregister.com](https://www.theregister.com)), causing NVIDIA to prioritize GB200 over smaller B100/B200 parts. Foxconn, however, was optimistic about ramping. This underscores that multiple companies (NVIDIA HQ, TSMC fabs, Foxconn/ODM factories) each play a role in launching a product. If any one stock has issues (e.g. packaging yields at TSMC's OSATs), it reverberates across the chain.

## Cost Breakdown and Economics

Understanding *how much* each part contributes to the cost helps reveal supply dependencies. An independent analysis by Epoch AI estimates the variable manufacturing cost of a single **NVIDIA B200 GPU module** (roughly one of the two GPUs in a GB200 superchip) at \$5,700–7,300 <sup>(4)</sup> [epoch.ai](https://epoch.ai)). Crucially, **HBM memory and advanced packaging together make up roughly two-thirds of that cost**. In its model, the authors assume 192 GB of HBM3e priced ~\$15/GB, yielding about \$2,880 of memory content <sup>(5)</sup> [epoch.ai](https://epoch.ai)). The rest (~\$1–3k more) goes to packaging and substrate. The compute die silicon (logic) is a smaller fraction (~30–40%). This breakdown highlights why memory and packaging suppliers capture a major portion of value.

Using their data, we can outline an approximate **BOM cost distribution** for one GPU die (B200):

Component	Estimated Cost (USD)	% of Module	Sources/Notes
Logic dies (2 GPUs)	~\$2,000–2,500	~30–40%	(Based on 3nm wafer cost assumptions <sup>(4)</sup> epoch.ai)
HBM3e memory (192GB)	~\$2,688–3,264	~40–50%	192 GB @ \$14–17/GB (triangular dist.; center \$15) <sup>(5)</sup> epoch.ai
Advanced packaging	~\$1,500–2,000	~20–30%	Est. by difference (EB PMI) <sup>(4)</sup> epoch.ai <sup>(5)</sup> epoch.ai
Auxiliary (PCB, VRM, etc.)	~\$500–1,000	~5–15%	Board-level parts not included above, yields etc.

Table 2: Approximate cost breakdown for a single NVIDIA B200 GPU module (192 GB HBM) <sup>(4)</sup> epoch.ai <sup>(5)</sup> epoch.ai.

(These figures exclude R&D, marketing, and system-level costs.) Notably, even optimistically the GPU compute dies contribute a minority of the cost. The HBM stacks (supplied by SK Hynix or Micron) and the CoWoS-style packaging (which includes substrate from Unimicron/Ibiden and OSAT labor) are the lion’s share. This aligns with industry claims that NVIDIA’s die-level margins are huge (over 80% in selling price) because materials cost is relatively low <sup>(40)</sup> epoch.ai, but at the rack/system level margins shrink.

The high memory content means **HBM suppliers earn a fortune** from Blackwell. For example, at \$15/GB, each GPU’s memory costs ~\$3k, so 72 GPUs in NVL72 consume ~\$216k just in DRAM. Packaging spend is also massive: if we assume \$1.5k per GPU for CoWoS packaging, that’s another \$108k per rack on substrates and OSAT costs. These simple estimates underscore why companies like Micron and SK Hynix focus intensely on AI markets (Micron mentioned hundreds of millions in revenue from HBM in FY2024 <sup>(36)</sup> www.anandtech.com).

## Supply-Chain Bottlenecks and Constraints

Despite the advanced planning, **supply constraints** have repeatedly emerged. Two key bottlenecks have been HBM memory and CoWoS packaging capacity:

- HBM Memory Shortage:** In March 2024, NVIDIA’s major memory partner Micron announced that its entire HBM3E output for 2024 was pre-sold (mostly to NVIDIA) and that it expected to “capture a sizable chunk” of the HBM market <sup>(36)</sup> www.anandtech.com). Micron’s CEO stated HBM3E was “sold out for calendar 2024, and the overwhelming majority of 2025 supply has already been allocated” <sup>(36)</sup> www.anandtech.com). This implies extremely tight supply for other players (like AMD or Chinese AI chip makers). Meanwhile, SK Hynix and Samsung will ramp their own HBM3e production, but bridging that gap could take years. The result is that HBM manufacturers hold bargaining power and NVIDIA likely had to prioritize who gets early shipments (Cloud Service Providers over enterprises).
- CoWoS Packaging Capacity:** TSMC’s CoWoS lines have been oversubscribed. After the H100 launch in 2022, even CEO C.C. Wei warned of an AI chip “shortage through 2025” due to CoWoS limits. By 2024, NVIDIA began exploring alternatives. In mid-2024 WCCFtech reported that NVIDIA had “decided to resolve CoWoS issues” by developing a new “PFLO” (panel-level fan-out) packaging for later Blackwell units <sup>(2)</sup> wccftech.com). Indeed, by late 2024/25 NVIDIA is reportedly shifting new Blackwell (GB200) units to this PFLO method <sup>(29)</sup> wccftech.com), in partnership with substrate firms.

Taiwan-based commentary echoes these issues. One semiconductor newsletter notes that individual component “supply chains get reworked” for each server OEM; specialized evaluation of over 50 subcomponents shows many had to ramp suddenly for GB200 <sup>(1)</sup> newsletter.semianalysis.com). Digitimes reported that GB200 shipments were delayed due to substrate/material shortfalls <sup>(38)</sup> www.digitimes.com) (the snippet indicates “key materials for chip substrates are in short supply”). All these signals point to the fact that even Tier-1 vendors cannot produce GB200 at will; they are constrained by the limited production capabilities of their upstream partners.

To overcome these, NVIDIA and the industry are taking steps:

- Onshoring:** Building the Arizona fab and planning US packaging lines (Amkor’s OSAT by 2027) reduce reliance on Taiwanese facilities <sup>(3)</sup> www.datacenterdynamics.com) <sup>(27)</sup> www.theregister.com).
- New packaging tech:** Transition to panel-level FO (PFLO) packaging with suppliers like Powertech and Innolux <sup>(32)</sup> wccftech.com) aims to relieve CoWoS load.

- **Prioritization:** NVIDIA explicitly prioritized GB200 (high-end racks) over lower-tier B100/B200 chips to maximize the use of scarce resources (<sup>[15]</sup> [www.theregister.com](http://www.theregister.com)) (<sup>[41]</sup> [www.trendforce.com](http://www.trendforce.com)).

## Case Study: AWS and DGX Cloud Deployments

Hyperscale cloud providers illustrate the system-scale integration of GB200 and its supply chain. **Amazon Web Services (AWS)** announced new EC2 instances leveraging NVIDIA's GB200. The P6e-GB200 UltraServers offer up to 72 GB200 GPUs in one system, with 360 PFLOPS of FP8 acceleration (<sup>[8]</sup> [aws.amazon.com](http://aws.amazon.com)). AWS highlights that they aggregated five generations of Nitro networking and liquid cooling to achieve these specs. Notably, **AWS's deployment underscores two supply-chain facets:**

1. **NVLink Integration:** AWS's description of "72 NVIDIA Blackwell GPUs interconnected using 5th-generation NVLink" (<sup>[8]</sup> [aws.amazon.com](http://aws.amazon.com)) confirms how AWS is adopting NVIDIA's supplied NVL72 architecture at scale. To implement this, AWS needed tens of thousands of NVLink switch chips and associated high-speed cables/connectors. (International Datacenters and network equipment from Accton or Arista supply AWS with 800G Ethernet switching, but the core GPU domain is NVIDIA/Amphenol technology.)
2. **Cooling Solutions:** AWS explicitly distinguishes their two offerings: P6-B200 instances ("air-cooled") versus P6e-GB200 Ultra ("liquid-cooled") (<sup>[39]</sup> [aws.amazon.com](http://aws.amazon.com)). They note that liquid cooling is needed to achieve the highest density. For suppliers, AWS has demanded scalable cooling subsystems. Companies like CoolIT or Chillydne (which provided cooling for AWS's earlier P4/G4 clusters) likely supply the rack-level chillers or cold-plate systems. The mention of "novel mechanical cooling solutions" implies that innovative coolant circulation equipment (e.g. Lytron cold plates, Bell-Free pumps) has been integrated. At large hyperscalers, 3M (owner of Novec fluids) and Fresenius (chemical coolants) also often play roles.

All told, AWS's roll-out demonstrates how the GB200 supply chain scales: Amazon had to order GPUs and boards from NVIDIA, network equipment from Accton, cables and switches from Nvidia's NVLink suppliers, chassis from Supermicro or Foxconn brand, and cooling from specialized vendors. Similarly, **NVIDIA DGX Cloud** (an internal DGX-based PaaS) announced hosting samples on DGX Cloud with GB200, leveraging the same ecosystem.

## Quantitative Estimates: Shipments and Scalability

To put numbers on the supply chain, analysts have made bold forecasts. WCCFtech reported (citing Taiwan Economic Daily) that NVIDIA expects to **ship ~420,000 Blackwell GB200 GPUs** in the second half of 2024, rising to 1.5–2 million units in 2025 (<sup>[33]</sup> [wccftech.com](http://wccftech.com)). These figures dwarf the ~100k H100 GPUs shipped in all of 2022. If true, by the end of 2025 NVIDIA will have deployed roughly 2 million Blackwell GPUs into the field. Converting devices to servers: since each 1U server holds at most 2 GPUs, that implies on the order of ~1 million GB200-accelerated servers, or ~20,000 full NVL72 racks (72 GPUs each yields ~360k GPUs per 100 racks). Sources similarly project **tens of thousands of racks** by 2026 – one analysis even suggests up to 60,000 racks by 2026 ([cloudnews.tech](http://cloudnews.tech)) (consistent with 4.3 million GPUs under such a scenario).

Driving this scale requires enormous ramping of suppliers. For example:

- **Memory:** 2 million GPUs × 192 GB = 384,000 TB of HBM3e. At ~\$15/GB, that's \$5.76 billion in DRAM.
- **Copper and Substrate:** Tens of thousands of 700+mm<sup>2</sup> substrates needed, each made of specialized materials (copper foil, epoxy, glass cloth from suppliers like Kolon or Ibiden).
- **Electronics Components:** Millions of high-end capacitors/inductors.
- **Cabling:** If each rack uses ~1000 feet of specialized cable, the total cable length is enormous.
- **Power Supplies:** Gigawatts of DC-to-DC conversion capacity installed.

No single supplier can handle such volumes alone. Even TSMC had to scale: it reportedly booked over half of its CoWoS capacity for 2026-27 for NVIDIA's chips (<sup>[38]</sup> [www.digitimes.com](http://www.digitimes.com)). Likewise, Micron has been expanding HBM fabs. OSAT players (Amkor, Powertech, PTI) are building new fab lines with backing from TSMC, driven by NVIDIA's orders. Meanwhile, PCB fabricators (Unimicron, Ibiden) are adding layers and copper volume capacity for AI substrates.

In effect, Blackwell has become an anchor tenant for the high-end semiconductor ecosystem. As one industry commentator put it, "NVIDIA's Blackwell: a milestone for US chips but Taiwan's packaging power" (<sup>[42]</sup> [www.linkedin.com](http://www.linkedin.com)). The entire chain – from raw silicon through final server assembly – is now targeted on this product.

## Discussion: Implications and Future Directions

The supply chain realities of the GB200 have broad implications:

- **Concentration of Suppliers:** The high-end AI chip supply chain is heavily concentrated in a few regions and firms. TSMC (Taiwan) and now Samsung/Intel (South Korea/USA) control the most advanced wafer production. Only a handful of OSATs can do CoWoS. Major DRAM makers (SKH, Micron) dominate HBM. This concentration means geopolitical and capacity risks. For example, a policy decision affecting Taiwan or export controls on HBM could ripple through NVIDIA's deliveries. This has led to initiatives like U.S. onshoring of fabs and packaging (through CHIPS Act funding) to mitigate risk (<sup>[17]</sup> [www.datacenterdynamics.com](http://www.datacenterdynamics.com)) (<sup>[27]</sup> [www.theregister.com](http://www.theregister.com)).
- **Cost Structure and Profit Margins:** As noted, the BOM for GB200 is heavily weighted toward memory and packaging costs (<sup>[4]</sup> [epoch.ai](http://epoch.ai)). NVIDIA's selling price per chip (\$30k–40k reported (<sup>[40]</sup> [epoch.ai](http://epoch.ai))) yields very large gross margins on silicon itself, but system prices (\$2–3M per rack) ensure NVIDIA's partners also see healthy revenue. However, rapid scaling may compress margins for some vendors if competition increases (e.g. if Samsung enters HBM3e at scale). The capital intensity is huge, so suppliers need steady orders to justify new fabs and lines.
- **Innovations in Packaging:** The GB200 has spurred packaging R&D. Panel-level FO and other techniques (embedded die laminates, fan-out wafer level) are being fast-tracked. If PFLO proves superior (as NVIDIA believes), it could displace CoWoS for future products. Already, specialized companies are emerging. For example, *Cohu* and *KLA* have accredited equipment for panel-level packaging aimed at NVIDIA's roadmap (<sup>[32]</sup> [wccftch.com](http://wccftch.com)). Over the next 2–3 years, new vendors in the packaging segment (beyond the traditional OSATs) may emerge, drawn by NVIDIA's demand.
- **Environmental and Infrastructure:** A side effect of GB200's power density is increased interest in liquid cooling. Analysts predict liquid cooling penetration in AI datacenters will jump to 30% by 2025 on NVIDIA's order (<sup>[43]</sup> [www.linkedin.com](http://www.linkedin.com)). This means vendors of cooling infrastructure (cooling distribution units, rack-level chillers, etc.) will see booming demand. Companies like Asetek, CoolIT, Altec, and Metz are likely to expand. Conversely, ordinary air-cooled designs (with fewer GPUs per rack) may coexist, as indicated by NVIDIA shipping a trimmed-down B200A variant for customers unsuited to liquid cooling (<sup>[44]</sup> [www.trendforce.com](http://www.trendforce.com)).
- **Supply Chain Coordination:** Meeting Blackwell demand requires unprecedented coordination. NVIDIA and government are pushing supply chain mapping, strategic materials stockpiling, and flexible manufacturing. For instance, TrendForce noted companies are securing wafer and substrate orders years in advance. NVIDIA's detailed briefings to investors mention dozens of upstream issues being tracked (from quick-disconnect leakage to PCB trace design) (<sup>[45]</sup> [newsletter.semianalysis.com](http://newsletter.semianalysis.com)). This level of granularity suggests that NVIDIA's own supply team must manage hundreds of vendor relationships.

**Future Outlook:** Through 2026, the GB200 is already giving way to the **GB300 NVL72** (Blackwell Ultra) rack — which features 288 GB of HBM3e per GPU (up from 192 GB), higher FP4 throughput, and is ramping in volume this year — with NVIDIA's next-generation **Rubin** platform on the roadmap for 2026–2027. The industry is already talking about "beyond trillion parameter models" requiring even denser interconnect. If NVIDIA pushes to, say, 100 GPUs per rack, supply chain demands will intensify further. Meanwhile, competitors (AMD, Intel, specialized chips) will drive complementary chains (e.g. memory for AMD's MI300, Intel's 4th-Gen Gaudi). But for now, NVIDIA's Blackwell is a "demand shock" that has financiers and supply chain analysts projecting strong growth for substrate makers, packaging companies, and integrators in 2026–2027 ([cloudnews.tech](http://cloudnews.tech)) ([cloudnews.tech](http://cloudnews.tech)).

## Conclusion



- [11] <https://www.nvidia.com/en-us/data-center/gb200-nvl72#:~:Unloc...>
- [12] <https://aws.amazon.com/blogs/machine-learning/aws-ai-infrastructure-with-nvidia-blackwell-two-powerful-compute-solutions-for-the-next-frontier-of-ai/#:~:AWS%2...>
- [13] <https://developer.nvidia.com/blog/nvidia-gb200-nvl72-delivers-trillion-parameter-llm-training-and-real-time-inference/#:~:for%2...>
- [14] <https://www.anandtech.com/show/21319/micron-sells-out-entire-hbm3e-supply-for-2024-most-of-2025#:~:Micro...>
- [15] [https://www.theregister.com/2024/08/14/nvidia\\_foxconn\\_blackwell/#:~:The%2...](https://www.theregister.com/2024/08/14/nvidia_foxconn_blackwell/#:~:The%2...)
- [16] <https://blogs.nvidia.com/blog/tsmc-blackwell-manufacturing/#:~:NVIDI...>
- [17] <https://www.datacenterdynamics.com/en/news/first-us-made-nvidia-blackwell-wafer-manufactured-at-tsmcs-arizona-fab/#:~:The%2...>
- [18] [https://www.theregister.com/2025/10/20/nvidia\\_arizona\\_blackwell/#:~:But%2...](https://www.theregister.com/2025/10/20/nvidia_arizona_blackwell/#:~:But%2...)
- [19] <https://www.trendforce.com/presscenter/news/20240807-12244.html#:~:Trend...>
- [20] <https://wccftech.com/nvidia-ship-half-a-million-blackwell-gb200-ai-chips-this-year-2-million-in-2025/#:~:It%20...>
- [21] <https://newsletter.semianalysis.com/p/gb200-hardware-architecture-and-component#:~:other...>
- [22] <https://newsletter.semianalysis.com/p/gb200-hardware-architecture-and-component#:~:Each%...>
- [23] <https://newsletter.semianalysis.com/p/gb200-hardware-architecture-and-component#:~:With%...>
- [24] <https://newsletter.semianalysis.com/p/gb200-hardware-architecture-and-component#:~:In%20...>
- [25] [https://www.theregister.com/2024/08/14/nvidia\\_foxconn\\_blackwell/#:~:Nvidi...](https://www.theregister.com/2024/08/14/nvidia_foxconn_blackwell/#:~:Nvidi...)
- [26] [https://www.theregister.com/2024/08/14/nvidia\\_foxconn\\_blackwell/#:~:The%2...](https://www.theregister.com/2024/08/14/nvidia_foxconn_blackwell/#:~:The%2...)
- [27] [https://www.theregister.com/2025/10/20/nvidia\\_arizona\\_blackwell/#:~:Up%20...](https://www.theregister.com/2025/10/20/nvidia_arizona_blackwell/#:~:Up%20...)
- [28] [https://www.theregister.com/2025/10/20/nvidia\\_arizona\\_blackwell/#:~:Moder...](https://www.theregister.com/2025/10/20/nvidia_arizona_blackwell/#:~:Moder...)
- [29] <https://wccftech.com/nvidia-ship-half-a-million-blackwell-gb200-ai-chips-this-year-2-million-in-2025/#:~:Taiwa...>
- [30] <https://epoch.ai/data-insights/b200-cost-breakdown#:~:,down...>
- [31] <https://www.datacenterdynamics.com/en/news/first-us-made-nvidia-blackwell-wafer-manufactured-at-tsmcs-arizona-fab/#:~:TSMC%...>
- [32] <https://wccftech.com/nvidia-ship-half-a-million-blackwell-gb200-ai-chips-this-year-2-million-in-2025/#:~:It%20...>
- [33] <https://wccftech.com/nvidia-ship-half-a-million-blackwell-gb200-ai-chips-this-year-2-million-in-2025/#:~:It%20...>
- [34] <https://www.anandtech.com/show/21078/micron-to-ship-hbm3e-to-nvidia-in-early-2024#:~:Micro...>
- [35] <https://www.anandtech.com/show/21319/micron-sells-out-entire-hbm3e-supply-for-2024-most-of-2025#:~:Shilo...>
- [36] <https://www.anandtech.com/show/21319/micron-sells-out-entire-hbm3e-supply-for-2024-most-of-2025#:~:,shar...>
- [37] <https://www.edge-ai-vision.com/2023/11/advanced-ic-substrates-taiwanese-companies-lead-the-market/#:~:%2A%2...>
- [38] <https://www.digitimes.com/news/a20250327PD220/nvidia-capacity-materials-high-end-ai-server.html#:~:Nvidi...>
- [39] <https://aws.amazon.com/blogs/machine-learning/aws-ai-infrastructure-with-nvidia-blackwell-two-powerful-compute-solutions-for-the-next-frontier-of-ai/#:~:Where...>
- [40] <https://epoch.ai/data-insights/b200-cost-breakdown#:~:Repor...>
- [41] <https://www.trendforce.com/presscenter/news/20240807-12244.html#:~:Trend...>
- [42] [https://www.linkedin.com/posts/anny-yu-5a40a8175\\_nvidia-tsmc-blackwell-activity-7385640543088939009-YBil#:~:Power...](https://www.linkedin.com/posts/anny-yu-5a40a8175_nvidia-tsmc-blackwell-activity-7385640543088939009-YBil#:~:Power...)

[ 43 ] [https://www.linkedin.com/posts/ai-chips-datacenters\\_ai-artificialintelligence-ainews-activity-7365854238637309952-9z84#:~:Liqui...](https://www.linkedin.com/posts/ai-chips-datacenters_ai-artificialintelligence-ainews-activity-7365854238637309952-9z84#:~:Liqui...)

[ 44 ] <https://www.trendforce.com/presscenter/news/20240807-12244.html#:~:Trend...>

[ 45 ] <https://newsletter.semianalysis.com/p/nvidias-blackwell-reworked-shipment#:~:This%...>

[ 46 ] <https://www.linkedin.com/pulse/questions-glass-substrate-aken-cheung-uriac#:~:The%2...>

[ 47 ] <https://newsletter.semianalysis.com/p/nvidias-blackwell-reworked-shipment#:~:out%2...>

## IntuitionLabs - Industry Leadership & Services

**North America's #1 AI Software Development Firm for Pharmaceutical & Biotech:** IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

**Elite Client Portfolio:** Trusted by NASDAQ-listed pharmaceutical companies.

**Regulatory Excellence:** Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

**Founder Excellence:** Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

**Custom AI Software Development:** Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

**Private AI Infrastructure:** Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

**Document Processing Systems:** Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

**Custom CRM Development:** Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

**AI Chatbot Development:** Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

**Custom ERP Development:** Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

**Big Data & Analytics:** Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

**Dashboard & Visualization:** Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

**AI Consulting & Training:** Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at <https://intuitionlabs.ai/contact> for a consultation.

---

## 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 IntuitionLabs.ai 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.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by [Adrien Laurent](#), a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

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

© 2025 IntuitionLabs.ai. All rights reserved.