# IT Hardware

## GB200 rack the spotlight for AI server in 2H24-2025F

### Key message

1. We expect total AI GPU shipments for Nvidia will rise to 4.0mn units in 2024F, and 4.7-4.8mn units in 2025F, based on a Hopper GPU allocation of 85% in 2024F and 25% in 2025F, and Blackwell GPU allocations of 10% in 2024F and 75% in 2025F.
2. Blackwell production shipments will start in 2Q24F, ramp up in 3Q24F, and customers should have the GPUs in data centers in 4Q24F. We expect Nvidia to see a 10% sales weighting for Blackwell in 2024F and 70-80% in 2025F. CSPs, enterprise, consumer IoT and governments are clients for AI GPUs & the main sales growth drivers.
3. GB200 rack demand is from top CSPs and server brands, and we expect 40-45k racks to ship in 2025F, with the top-4 US CSP comprising 80-85% of demand, and Hon Hai (2317 TT) & Quanta (2382 TT) are the key assemblers. Thermal solution, power supply, DAC cable & high-end optical fiber connector sales will also benefit from this trend.

### Event

Per Nvidia's (US) guidance that the Blackwell platform is in production in 2Q24, with a ramp up in 3Q24F, customers should have Blackwell GPUs installed in data centers in 4Q24F. GPU transition hiccups will be limited due to strong AI demand, so we expect AI server supply chain to continue to benefit from growing sales and profits in 2024-25F.

### Impact

**Nvidia's Blackwell GPUs in the pipeline to boost 2025F AI server sales.** On increased demand for Nvidia's AI GPUs, TSMC (2330 TT) has expanded CoWoS capacity from 13k wafers per month (kwpm) in 4Q23 to 40kwpm in 4Q24F and 55kwpm in 4Q25F expected. Total AI GPU shipments for Nvidia will rise to 4.0mn units in 2024F, versus our previous expectation of 3.55mn units, and up to 4.7-4.8mn units in 2025F, based on a Hopper GPU allocation of 85% in 2024F and 25% in 2025F, and Blackwell GPU allocations of 10% in 2024F and 75% in 2025F. The market expects GB200 AI server models to be in high demand on strong client interest. We think Blackwell GPUs should come on stream in September, with some volume in the market in 4Q24F. B100 and B200 GPUs, and the GB200, should launch in the market around the same time, with significant volume available in 2025F. Combined with AMD's (US) MI300X/350, Intel's (US) Gaudi 2 and 3, and ASIC shipments (mainly from Google's (US) TPU and Amazon Web Services' (AWS; US) Trainium), we expect total training AI GPU shipments will grow to 5.5mn units in 2024F and 7.9mn units in 2025F, translating to AI training server shipments of 715k units in 2024F and 1.25mn in 2025F. Training AI servers will comprise 6% of total server shipments in 2024F and 10% in 2025F. The total AI server shipment weighting will be even higher if inference servers are taken into consideration. On AI servers' high ASP, we forecast an AI server revenue weighting of 60-70% of global server revenue in 2024-25F.

**AI servers carry higher content value for thermal, power supply & rack assembly sectors.** During a May 22 earnings call, Nvidia cited Blackwell production shipments will start in 2Q24F, ramp in 3Q24F, and customers should have the GPUs in data centers in 4Q24F. We expect Nvidia to see a 10% GPU shipments weighting for Blackwell in 2024F and 70-80% in 2025F. CSP, enterprise, consumer IoT and sovereign are clients for the firm's AI GPUs, and the main sales growth drivers. We expect GB200 rack demand among top CSPs and server brands to result in 40-45k racks shipping in 2025F. Microsoft (US) plans to receive 35% of GB200 racks in 2025F, while AWS, Meta (US), Google, Supermicro (US), Dell (US), Oracle (US) are also customers. We expect these CSPs will keep revising up capex in 2025F from upward-revised high level in 2024F. Hon Hai (2317 TT) group's FII (CN) is the major GB200 rack supplier to Microsoft and Oracle, and Quanta Computer (2382 TT) is for AWS, Meta and Google. We expect Hon Hai's rack assembly market share for GB200 servers to be approximately 40%, with Quanta holding around 30%. The remaining 30% will go to Wiwynn (6669 TT), Wistron (3231 TT), ZT Systems (US) and Supermicro. The high ASP of GB200 server racks, of between US$1.8mn-3.5mn for NVL36/ 72 racks, should contribute ODMs' AI server sales weighting to over 50% of their total server revenue in 2025F. Among component sectors, thermal sector will see the greatest content value rise on liquid cooling adoption. With cold plate modules, coolant distribution units (CDU), manifolds, racks, RDHx, fan door and chassis required, we expect the thermal content value will rise to US$40-80k for each GB200 rack, versus an air cooling content value of US$2-3k per rack. On the power supply side, with power consumption per GB200 chip reaching 2.7kW, a NVL72 rack could require power of 120kW, or even higher. We expect the content value for each NVL72 rack will be at least US$18-20k, versus US$6-8K for H100 designs. We are also optimistic about the increase in the number and specifications of NVLinks when GPUs migrate from Hopper-series to Blackwell-series, benefiting DAC cable and high-end optical fiber connector suppliers.

### Rating

We expect key benefacaries of the AI server trend will include Hon Hai, Quanta Computer, Wiwynn, Wistron, Asia Vital Components (AVC; 3017 TT), Auras Technology (3324 TT), Kaori Heat Treatment (8996 TT), Delta Electronics (2308 TT), Chenbro Micom (8210 TT), King Slide Works (2059 TT), Browave (3163 TT), and Jess-Link Products Co. (6197 TT).

### Risks

Weak demand; over-ordering of AI servers by CSPs.

## Key assumption chart

**Figure 1: AI GPU & AI server shipments forecast in 2024-25F**

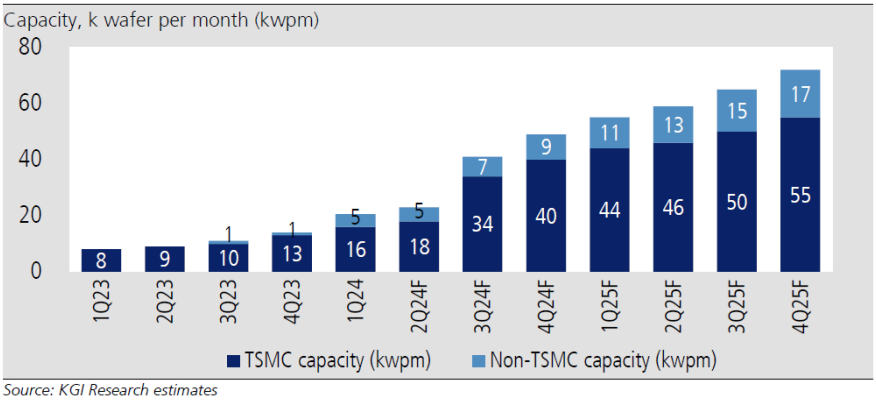
| Nvidia's training GPU units (k units) | 2024F | 2025F |
|---|---|---|
| Total Nvidia GPU | 4,005 | 4,760 |
| H100/ A100 shipments | 3,605 | 1,190 |
| Blackwell shipments (B100/B200/GB200) | 401 | 3,570 |
| B100/B200 | 200 | 1,474 |
| GB200 | 200 | 2,096 |
| **Training GPU weighting (%)** | **2024F** | **2025F** |
| H100/ A100 shipments | 90 | 25 |
| Blackwell shipments (B100/B200/GB200) | 10 | 75 |
| B100/B200 | 50 | 41 |
| GB200 | 50 | 59 |
| **Nvidia's AI training server shipments (k units)** | **2024F** | **2025F** |
| H100 /A100 (8 GPU per server) | 451 | 149 |
| B100/B200 (8 GPU per server) | 25 | 184 |
| GB200 (4 GPU per server) | 50 | 524 |
| **Nvidia's total AI training server** | **526** | **857** |
| **AI training GPU shipments (k units)** | **2024F** | **2025F** |
| Nvidia (H/B-series) | 4,005 | 4,760 |
| AMD (MI300X/350) | 420 | 768 |
| Intel (Gaudi 2/3) | 232 | 450 |
| Subtotal | 4,657 | 5,978 |
| ASIC (Google TPU/AWS Trainium) | 864 | 1,888 |
| Total training GPU | 5,521 | 7,866 |
| **AI training server shipments (k units)** | **2024F** | **2025F** |
| Nvidia (H/B-series) | 526 | 857 |
| AMD (MI300X/350) | 53 | 96 |
| Intel (Gaudi 2/3) | 29 | 56 |
| Subtotal | 607 | 1,009 |
| ASIC (Google TPU/AWS Trainium) | 108 | 236 |
| **Total AI training server shipments** | **715** | **1,245** |
| **Server shipments (k units)** | **2024F** | **2025F** |
| AI training server | 715 | 1,245 |
| Others (General server + AI inferencing) | 11,106 | 11,759 |
| Total server | 11,822 | 13,004 |
| **YoY (%)** | **2024F** | **2025F** |
| AI training server | 188 | 74 |
| Others (General server + AI inferencing) | 1 | 6 |
| Total server | 5 | 10 |
| **Weighting (%)** | **2024F** | **2025F** |
| AI training server | 6 | 10 |
| Others (General server + AI inferencing) | 94 | 90 |
| Total server | 100 | 100 |

*Source: KGI Research estimates*

### CoWoS capacity expansion to support AI server GPU demand surge

- On increased demand for Nvidia's (US) AI GPUs, TSMC (2330 TT, NT$875, NR) continues to expand CoWoS capacity, along with other suppliers like Amkor (US) and ASE Technology (3711 TT, NT$161.5, NR). We expect TSMC's CoWoS capacity will grow from 13k wafers per month (kwpm) in 4Q23 to 40kwpm in 4Q24F and 55kwpm in 4Q25F, and from 1kwpm, and 9kwpm to 17kwpm from others over the same period (Figure 2).
- We therefore expect total AI GPU shipments for Nvidia will grow to 4.0mn units in 2024F, versus our previous expectation of 3.55mn units, and 4.7-4.8mn units in 2025F, based on a Hopper GPU allocation of 85% in 2024F and 25% in 2025F, and Blackwell GPU allocations of 10% in 2024F and 75% in 2025F (Figure 3).
- Following Nvidia's showcase of Blackwell GPUs and GB200 superchips during GTC 2024, the market expects GB200 AI server models, under Nvidia's DGX architecture, to achieve high sales volumes on strong client interest. We think Blackwell GPUs may come on stream in September, with some volume in the market in 4Q24F. B100 and B200 GPUs (upgrades from the H100 and H200 under the x86 CPU platform) and the GB200 (Grace CPU and B200 GPU under the ARM platform) should launch in the market around the same time, with significant volume available in 2025F.
- In 2025F, we assume 75% of Nvidia's AI GPU sales will be of the Blackwell-series, with 25% of the Hopper-series. Considering CSPs' AI server demand of between 42-43k racks, we expect meeting GB200 demand will require around 2.1mn units of GPUs, comprising 50-60% of the total Blackwell GPU supply, while B100 & B200 will comprise 40-50% of supply in 2025F (Figure 3).
- The allocation between Hopper and Blackwell, and the B100, B200 and GB200 are still under discussion, dependant upon Nvidia's market strategy. We will have further review of the volume of Blackwell GPUs and GB200 systems, as they are subject to CoWoS production and yield rate, memory supply (HBM3e) and NVLink switch production.

**Figure 2: CoWoS capacity expansion to support AI server demand increase**




*Source: KGI Research estimates*

**Figure 3: Nvidia's AI training server shipments will grow to 526k units in 2024F and 857k units in 2025F**

| Nvidia's training GPU units (k units) | 2024F | 2025F |
|---|---|---|
| Total Nvidia GPU | 4,005 | 4,760 |
| H100/ A100 shipments | 3,605 | 1,190 |
| Blackwell shipments (B100/B200/GB200) | 401 | 3,570 |
| B100/B200 | 200 | 1,474 |
| GB200 | 200 | 2,096 |
| **Training GPU weighting (%)** | **2024F** | **2025F** |
| H100/ A100 shipments | 90 | 25 |
| Blackwell shipments (B100/B200/GB200) | 10 | 75 |
| B100/B200 | 50 | 41 |
| GB200 | 50 | 59 |
| **Nvidia's AI training server shipments (k units)** | **2024F** | **2025F** |
| H100 /A100 (8 GPU per server) | 451 | 149 |
| B100/B200 (8 GPU per server) | 25 | 184 |
| GB200 (4 GPU per server) | 50 | 524 |
| **Nvidia's total AI training server** | **526** | **857** |

*Source: KGI Research estimates*

### AI server demand keeps growing in 2024-25F; system integration becomes more important with GB200 designs in 2025F

- Based on industry CoWoS capacity expansions, we see AI server GPU shipment growth in 2024-25F just meeting AI server demand from CSPs and enterprises. Based on 4mn units of AI GPUs from Nvidia in 2024F, with 90% being A100, H100, and H200 designs and 10% of B100, B200, and GB200 designs in 2024F, we expect AI training server shipments of Nvidia GPUs will reach 526k units, as most AI servers are configured in eight-GPU designs (Figure 3).
- During a May 22 earnings call, Nvidia cited Blackwell production shipments will start in 2Q24F, ramp in 3Q24F, and customers should have the GPUs in data centers in 4Q24F. We expect Nvidia to see a 10% sales weighting for Blackwell in 2024F and 70-80% in 2025F. CSP, enterprise, consumer IoT and sovereign are clients for the firm's AI GPUs, and the main sales growth drivers.
- In 2025F, as most (70-80%) AI training servers will be upgraded to B100, B200 and GB200 designs, we currently assume 2.1mn GPU units for GB200 designs, around 1.5mn units for B100 and B200 designs, and 1-1.2mn for H100 and H200 designs (Figure 4). This is based on CSPs' GB200 orders, and the allocation will be clearer after 3Q24F.
- As for GB200 rack demand among top CSPs and server brands, we expect 40-45k racks to ship in 2025F, according to the supply chain. Microsoft (US) plans to receive 15k GB200 racks in 2025F, AWS (US) 10k racks, Meta (US) 5k racks, Google (US) 6k racks, Supermicro (US) 5k racks, and Dell (US), Oracle (US) and others combined at 1-2k racks.
- Based on current GB200 designs, only Microsoft will adopt the NVL72 design (1U servers), while AWS will design 2 racks of NVL36 (2U) as a set, and Meta and Google will use the NVL36 design. The NVL36 is a 2U server design, and will adopted by CSPs on heat dissipation considerations in 2025F.
- The GB200 NVL72 system, which includes 36 Grace CPUs and 72 Blackwell GPUs, and the GB200 NVL36, which includes 18 Grace CPUs and 36 GPUs, are both very likely to use liquid cooled designs.
- Microsoft's GB200 AI server assemblers will be Hon Hai (2317 TT, NT$173, OP) and FII (CN), and Quanta (2382 TT, NT$284, OP) for AWS, Meta and Google. We therefore expect Hon Hai's rack assembly market share for GB200 servers will be approximately

40%, while Quanta holding around 30%. The remaining 30% will go to Wiwynn (6669 TT, NT$2,740, OP), Wistron (3231 TT, NT$114, OP), ZT (US) and Supermicro.

- With expected Nvidia's total AI GPU shipments of 4.0mn units in 2024F and 4.76mn units in 2025F, Nvidia's AI training server GPU shipments will be 526k units in 2024F and 857k units in 2025F.
- Besides Nvidia's high-end AI training GPUs, AMD (US) should produce 420k units of GPUs in 2024F and 768k units in 2025F. Intel's (US) Gaudi 2 and Gaudi 3 will see 232k units produced in 2024F and 450k units in 2025F. Thus, global AI GPU shipments should grow to 4.66mn units in 2024F and 5.98mn units in 2025F (Figure 5).
- Combined with ASIC designs, including Google's TPU and AWS' Trainium, we expect total training GPU shipments to be 5.52mn units in 2024F, and 7.87mn units in 2025F. Global training AI server shipments will be 715k units in 2024F and 1.25mn units in 2025F. This forecast is revised up from our previous one, to reflect a higher ASIC volume than we had expected (Figure 6).
- With the expected AI training server shipments, we anticipate training servers to comprise 6% of total server shipments in 2024F, and up to 10% in 2025F. Strong growth in AI training server demand should be the major driver of global server demand during this period.

**Figure 4: Top-4 CSPs will account for 85% of GB200 demand in 2025F**

| Units | Racks | Weighting (%) | Major ODM |
|---|---|---|---|
| MSFT | 15,000 | 35 | Hon Hai |
| AWS | 10,000 | 24 | Quanta |
| Google | 6,000 | 14 | Celestica / Quanta |
| Meta | 5,000 | 12 | Quanta |
| Others | 6,500 | 15 | |
| Total | 42,500 | 100 | |

*Source: KGI Research estimates*

**Figure 5: AI server shipments rising in 2024-25F; we revise up 2024-25F demand**

| AI training GPU shipments (k units) | 2024F | 2025F |
|---|---|---|
| Nvidia (H/B-series) | 4,005 | 4,760 |
| AMD (MI300X/350) | 420 | 768 |
| Intel (Gaudi 2/3) | 232 | 450 |
| Subtotal | 4,657 | 5,978 |
| ASIC (Google TPU/AWS Trainium) | 864 | 1,888 |
| Total training GPU | 5,521 | 7,866 |
| **AI training server shipments (k units)** | **2024F** | **2025F** |
| Nvidia (H/B-series) | 526 | 857 |
| AMD (MI300X/350) | 53 | 96 |
| Intel (Gaudi 2/3) | 29 | 56 |
| Subtotal | 607 | 1,009 |
| ASIC (Google TPU/AWS Trainium) | 108 | 236 |
| **Total AI training server shipments** | **715** | **1,245** |

*Source: KGI Research estimates*

**Figure 6: AI training server weighting of total server shipments up from 2% in 2023 to 10% in 2025F**

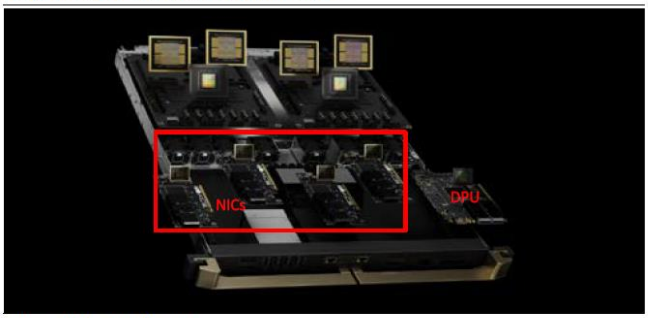
| Server shipments (k units) | 2022 | 2023 | 2024F | 2025F |
|---|---|---|---|---|
| AI training server | 124 | 249 | 715 | 1,245 |
| Others (General server + AI inferencing) | 13,703 | 11,010 | 11,106 | 11,759 |
| Total server | 13,827 | 11,259 | 11,822 | 13,004 |
| **YoY (%)** | **2022** | **2023F** | **2024F** | **2025F** |
| AI training server | | 101 | 188 | 74 |
| Others (General server + AI inferencing) | | (20) | 1 | 6 |
| Total server | 7 | (19) | 5 | 10 |
| **Weighting (%)** | **2022** | **2023F** | **2024F** | **2025F** |
| AI training server | 1 | 2 | 6 | 10 |
| Others (General server + AI inferencing) | 99 | 98 | 94 | 90 |
| Total server | 100 | 100 | 100 | 100 |

*Source: Gartner; KGI Research estimates*

### Nvidia showcased GB200 NVL72 / NVL36 platforms, a DGX design

- Nvidia's GB200 NVL72 is a rack-level solution with liquid-cooling design and is capable of more than 1.4 exaflops. Nvidia GB200 NVL72 is structured as 10 compute trays (1U height) atop nine switch trays (2U each) over eight additional compute trays (1U each), a structure referred to as "10+9+8". There will be a TOR Switch on the top of the rack, and two power trays (3U height) with one at the top, and the other at the bottom of the rack (Figure 7-9).
- Nvidia has also introduced the DGX SuperPOD, which consists of eight GB200 NVL72 racks with liquid cooling for optimal efficiency (Figure 10).
- The GB200 NVL36 is a customized solution and most CSP will adopt it. The GB200 NVL36 is structured as five compute trays (2U height) atop nine switch trays (2U each), over four additional compute trays (2U each), referred to as "5+9+4".
- The major difference between the GB200 NVL72 and GB200 NVL36 is that the latter only has 9 compute trays per rack, with each compute tray at a 2U height to allow for customized designs and better thermal efficiency.
- Each compute tray is composed of two GB200 Grace Blackwell Superchips (each with 1 Grace GPU and 2 Blackwell GPU per Superchip), an NVLink and four Smart NIC cards. That is, there will be 36 CPUs and 72 Blackwell GPUs connected together via NVLinks in a GB200 NVL72 rack, and 18 CPUs and 36 GPUs in a GB200 NVL36 rack (Figure 7).
- Each Switch tray will include two NVLink Switch chips and provide 14.4TB/s of aggregate bandwidth (Figure 8).
- Per our calculations, we estimate the ASP of a GB200 NVL72 to be US$3-3.5mn, and a GB200 NVL36 at US$1.8-2mn per rack, with the GPUs and CPUs accounting for the bulk of BOM cost, at between 75-80% (Figure 11).
- In addition, several key components will enjoy a higher content cost in GB200 NVL72/36 servers compared to H100 racks (4xH100 servers per rack), including thermal solutions, power supplies, and rack assembly. We will discuss spec upgrades and content value differences of key components in the following section.

**Figure 7: GB200 compute trays are composed of GB200 Superchips, smart NICs and DPU**

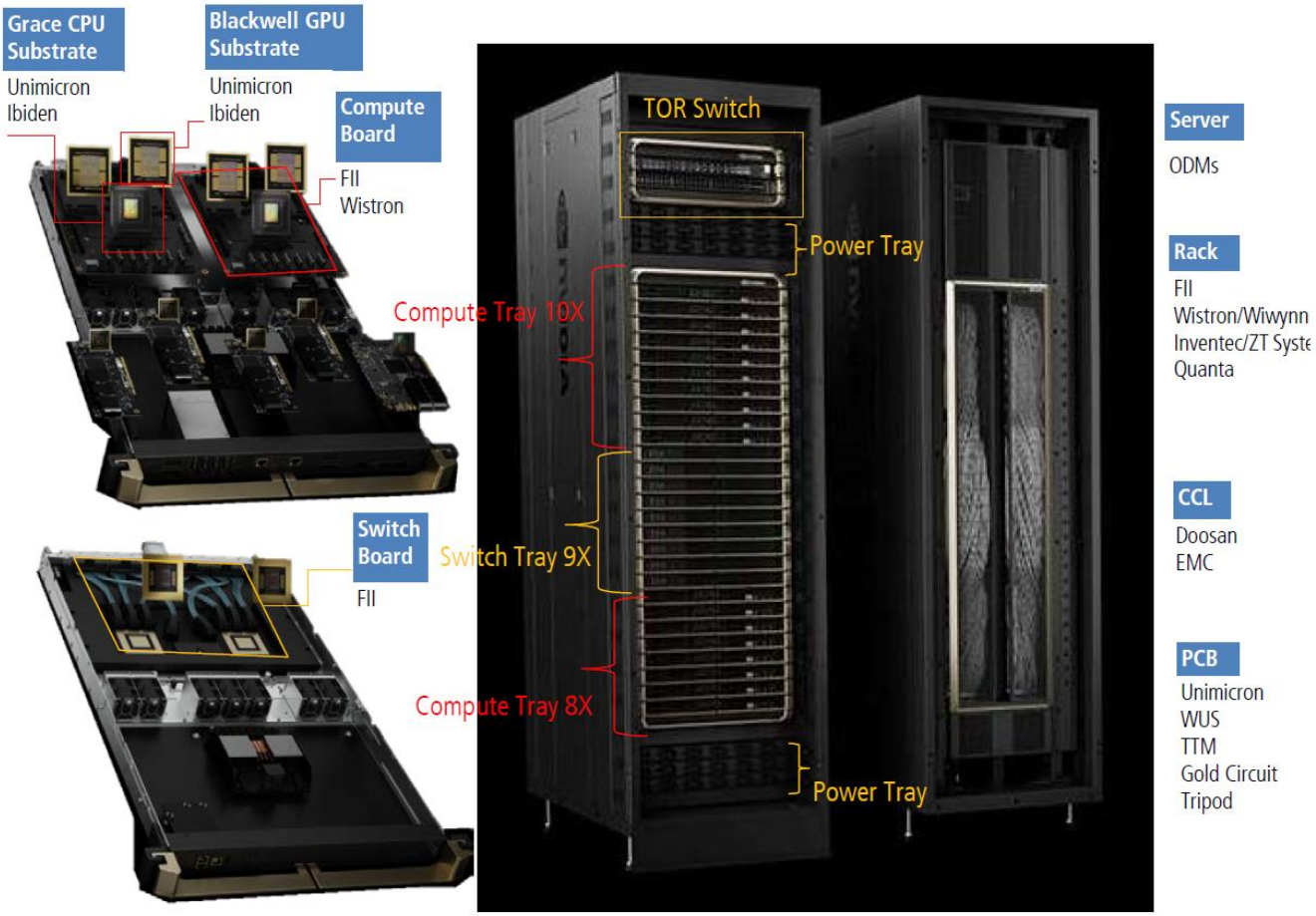



Source: Nvidia; KGI Research

**Figure 8: Each GB200 switch tray includes two Nvlink chips**

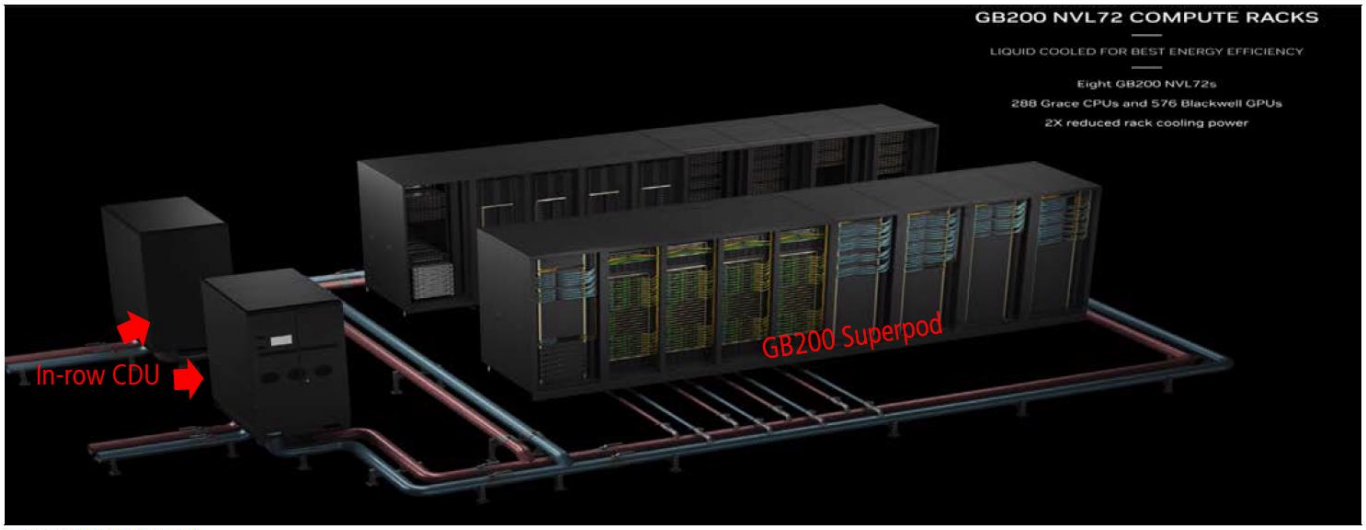

Source: Nvidia; KGI Research

**Figure 9: GB200 NVL72 breakdown & supply chain**




Source: Nvidia; KGI Research

**Figure 10: GB200 Superpod – Contains 8 NVL72 racks**




*Source: Nvidia; KGI Research*

**Figure 11: BOM cost analysis - NVL 72 & NVL 36 racks**

| | NVL36 | NVL72 |
|---|---|---|
| Rack shipment (units) | 26,775 | 15,725 |
| Weighting of total GB series (%) | 63 | 37 |
| **Assumption** | | |
| Compute tray | 9 | 18 |
| GPU per tray | 4 | 4 |
| CPU per tray | 2 | 2 |
| Smart NIC per tray | 4 | 4 |
| Swtich tray | 9 | 9 |
| **Total shipments (k units)** | | |
| GPU | 964 | 1,132 |
| CPU | 482 | 566 |
| **ASP** | | |
| GPU | 35,000 | 35,000 |
| CPU | 2,500 | 2,500 |
| Smart NIC | 2,000 | 2,000 |
| **Content value per rack (US$mn)** | **1.8-2** | **3-3.5** |
| **Content value mix per rack (%)** | | |
| GPU | 67.2 | 74.9 |
| Grace CPU | 2.4 | 2.7 |
| NIC | 3.8 | 4.3 |
| Compute tray | 73.5 | 81.9 |
| Switch tray | 16.0 | 8.9 |
| Thermal | 3.6 | 2.9 |
| Power, chassis and others | 2.1 | 1.5 |
| Assembly | 4.8 | 4.8 |
| **Total** | **100.0** | **100.0** |

*Source: KGI Research estimates*