Networking sector

NVLink to upgrade with GPU platforms

Key message

1. SerDes chips are at the center of Blackwell NVLink interconnect upgrades. We believe the SerDes spec upgrade will drive new product launches from Broadcom (US), Marvell (US), Synopsys (US), Texas Instruments (US), Cadence (US), Credo (US), and MediaTek (2454 TT, NT$1,385, NR).
2. A single GB200 NVL72 compute node can connect 72 GPUs in one-tier architecture, which means it has lower demand for two-tier networking for inter-node connections. GPU connections within a compute node are carried out via copper cables. Thus we believe upgrades to copper cables will be more significant than fiber-optic cable upgrades.
3. As the NVLink 5 of the next-generation Blackwell platform will reach 1.8TBps, and the NVLink 6 Switch used for the Rubin platform after Blackwell will reach 3.6TBps, we believe connector and splitter makers will benefit from faster intra-node connections, such as Browave (3163 TT, NT$110.5, NR), Jess-Link Products (6197 TT, NT$177.5, NR) and Bizlink Products (3665 TT, NT$304, OP)

Event

As a follow-up to our previous industry report, we further compare changes to NVLink interconnects, which provide chip-to-chip and rack-to-rack communication in Hopper & Blackwell AI servers.

Impact

SerDes at the center of Blackwell NVLink interconnect upgrades. According to Nvidia (US), each GPU in a Blackwell server may support up to eighteen NVLink interconnects, which is the same as for a Hopper server GPU. However, the data transfer rate of the latest NVLink 5, for use in Blackwell servers, is 100GBps, doubling the 50GBps for Hopper's NVLink 4. The total bandwidth of a Blackwell server will expand to 1.8TBps as a result, up from 900GBps for a Hopper server. The difference between the transfer rate of NVLink 4 and NVLink 5 interconnects stems primarily from the spec upgrade of serializer/ deserializer (SerDes) chips from 112Gbps to 224Gbps. We believe the SerDes spec upgrade will drive new product launches from Broadcom (US), Marvell (US), Synopsys (US), Texas Instruments (US), Cadence (US), Credo (US), and MediaTek (2454 TT, NT$1,385, NR).

Upgrade to copper cables more significant than fiber-optic cable upgrades. With Hopper GPU and Blackwell GPU both supporting up to eighteen NVLink interconnects, we believe the focus will be more on the cable spec upgrades triggered by the SerDes upgrade. As far as compute nodes are concerned, the content ratio of GPUs to NVSwitch chips in a GH200 node is 4:3, whereas in a GB200 NVL72 node the ratio is 4:1. When there is more than one GH200 node, that is more than eight GPUs to connect, a 1:2 two-level tapered fat-tree network architecture is required to connect two different nodes. A single GB200 NVL72 compute node can connect 72 GPUs in one-tier architecture, which means it has lower demand for two-tier networking for inter-node connections. GPU connections in single compute nodes are carried out via copper cables, whereas fiber-optic cables are needed for second-tier connections. Notably, the benefits of a higher fiber-optic transfer rate will be partially offset by reduced requirements for fiber-optic cables in Blackwell servers, and thus we conclude that intra-node upgrades will be more significant than inter-node upgrades.

Transceiver, connector & splitter makers to benefit from faster connections. A DGX H100 SuperPod is equipped with a ConnectX-7 network card and a Quantum-2 QM9700 switch system, along with a multitude of 2*400Gbps OSFP transceiver modules to enable 4 NVLink 4 connections. While the datasheet of a DGX B200 SuperPod (Blackwell platform) recommends the same specifications, we note that the next-generation 800Gbps Quantum-X800 switch series and ConnectX-8 SuperNIC will become available in 2025F. Against such a backdrop, we believe the specifications of the OSFP transceiver module will improve to 2x800Gbps, leading to network transceivers, connectors, and splitters being upgraded accordingly.

Stocks for Action

As the NVLink 5 of the next-generation Blackwell platform will reach 1.8TBps, and the NVLink 6 used in the generation (Rubin platform) after Blackwell will reach 3.6TBps, we believe that the number of NVLink interconnects and Serdes specifications will continue to evolve. In addition to suppliers of 224Gbps SerDes chips, we believe network cable, connector, and splitter manufacturers in Taiwan, such as Browave (3163 TT, NT$110.5, NR), Jess-Link Products (6197 TT, NT$177.5, NR) and Bizlink (3665 TT, NT$304, OP) will benefit from migrations to the Blackwell server platform.

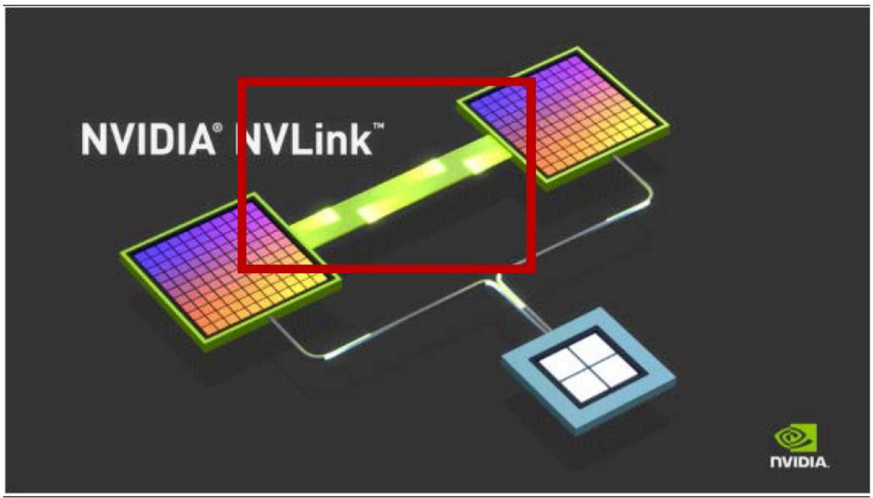
Risks

Slower-than-expected AI development; macroeconomic downturn.

Introduction to NVLink

Conceptually, NVLink is a technology developed by Nvidia for high-speed data transmission between GPUs in the same system. Data transmission via NVLink interconnect is more efficient, as it takes place without going through memory or the CPU. The technology helps optimize GPU and memory loading, consequently increasing data transmission speed requirements for high-performance computing.

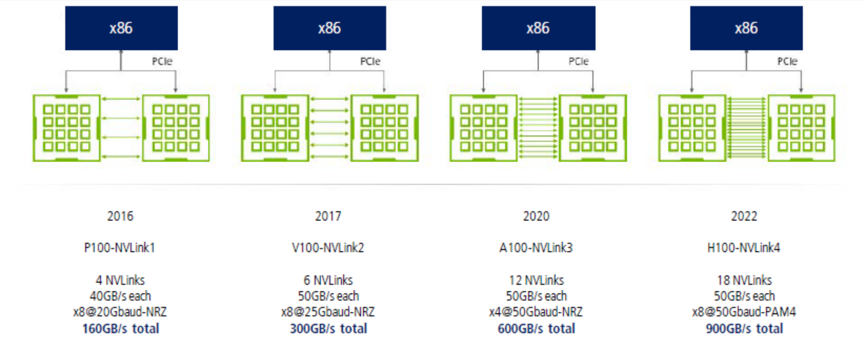
Figure 1: High-speed data transmission between two GPUs is primary application for NVLink interconnects



Source: Nvidia; KGI Research

The proliferation of AI and HPC applications means growing demand for high-performance systems, and thus an increasing number of computing designs have migrated from single-node to multi-node architecture. A multi-node compute unit contains multiple GPUs to perform large-scale computational tasks. To that end, it is important to ensure efficient GPU-to-GPU communication. Therefore, NVLink interconnect specs are being upgraded along with the increase in the number of GPUs that need to be connected and the demand for faster computing.

Figure 2: NVLink interconnect specs are advancing in tandem with GPUs



Source: Nvidia; KGI Research

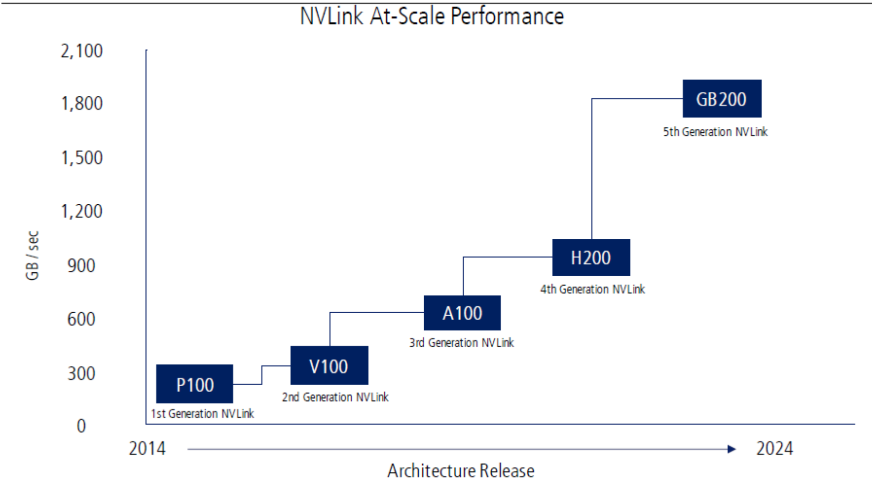
The recently announced NVLink 5 interconnect will substantially improve connections between large-sized multi-GPU computing systems. An Nvidia Blackwell Tensor core GPU can support up to 18 NVLink connections, each capable of transmitting data at 100GBps, for a total bandwidth of 1.8TBps, which is twice of the bandwidth of the NVLink 4 for the Hopper platform (previous generation). Connections between NVLink interconnects are bridged by NVSwitch chips, and with proper integration, allow chip-to-chip communication at full speed within a service rack or amongst multiple racks.

Figure 3: Evolution of NVLink generations

	NVLink 2	NVLink 3	NVLink 4	NVLink 5
NVLink bandwidth per GPU	300GB/s	600GB/s	900GB/s	1,800GB/s
Maximum Number of Links per GPU	6	12	18	18
Supported NVIDIA Architectures	Volta	Ampere	Hopper	Blackwell

Source: Nvidia; KGI Research

Figure 4: NVLink interconnect performance comparison

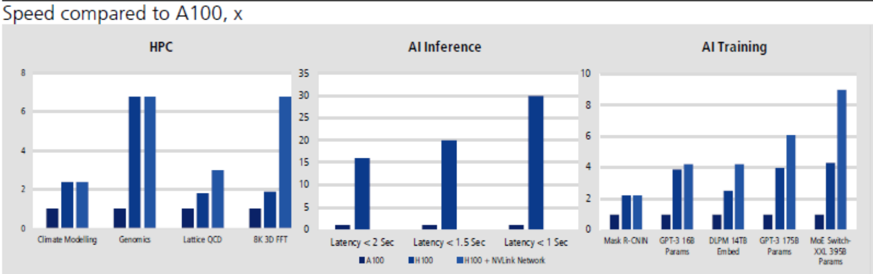


Source: Nvidia; KGI Research

According to Nvidia, a computing system with eight H100 Tensor core GPUs, NVLink 4 interconnects, and NVSwitch Gen 3 chips features 3.6TBps of bisection bandwidth and 450GBps of bandwidth for reduction operations, which are 1.5x and 3x increases over the prior generation (A100).

Figure 5: High-speed GPU-to-GPU data transmission is primary application of NVLink interconnects

Speed compared to A100, x



Source: Nvidia; KGI Research

NVLink speed conversion

Specific phrases, including "differential pairs" and "bidirectional", often appear in Nvidia's documents on NVLink interconnect specs. Take the following chart as an example. According to official data and Figure 4, NVLink 3 interconnects, designed for the A100 series, supports bandwidth of a maximum of 600GBps and 12 NVLink interconnects per GPU. Therefore, based on 1Byte = 8bits, we can convert 600GBps to 4,800 Gbps for even distribution to 12 NVLink interconnects at 400Gbps (or 50GBps) per NVLink interconnect. What a single pair adopts is a 50Gbps SerDes. A single direction link needs four different pairs to aggregate 50Gbps into 200Gbps. As data transmission is bidirectional, the other direction's transmission also needs four different pairs to form another 200Gbps (bit converted to Byte, equal to 25GBps). A total of eight 50Gbps pairs forming bidirectional 200Gbps is equal to the speed of the transmitting and receiving ends of a transceiver module. Since one A100 supports 12 NVLink interconnects, the total bandwidth of

600GBps of A100 is actually 300Gbps for each direction of transmission. In some Nvidia documents, we can also see that bidirectional SerDes is called Dual SerDes.

Figure 6: NVLink interconnect used to transmit data at high speed between GPUs

Third-Generation NVLink

The third-generation of NVIDIA's high-speed NVLink interconnect is implemented in the NVIDIA Ampere architecture-based A100 GPU and the new NVSwitch. NVLink is a lossless, high-bandwidth, low-latency shared memory interconnect, and includes resiliency features such as link-level error detection and packet replay mechanisms to guarantee successful transmission of data.

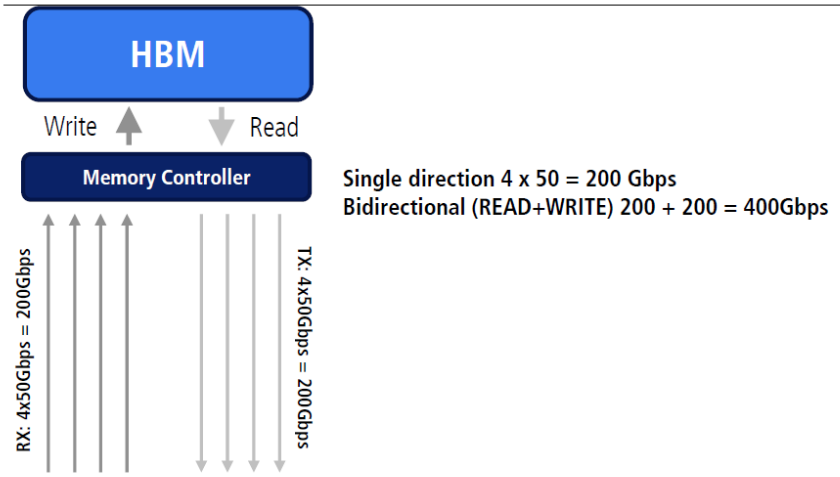
The new NVLink significantly enhances multi-GPU scalability, performance, and reliability with more links per GPU, much faster GPU-GPU communication bandwidth, and improved error-detection and recovery features. A100 GPUs can use NVLink links to access peer GPU memory at bandwidths much higher than achievable with PCI Express.

The new NVLink has a data rate of 50 Gbit/sec per signal pair, nearly doubling the 25.78 Gbits/sec rate in Tesla V100. Each link uses 4 differential signal pairs (4 lanes) in each direction compared to 8 signal pairs (8 lanes) in Volta. A single link provides 25 GB/second bandwidth in each direction similar to Volta GPUs, but uses only half the signals compared to Volta. The total number of NVLink links is increased to twelve in A100, versus six in Tesla V100, yielding a whopping 600 GB/sec total bandwidth for an entire A100 versus 300 GB/sec for Tesla V100.

The twelve NVLink links in each A100 allow a variety of configurations with high-speed connections to other GPUs and switches. To meet the growing computational demands of larger and more complex DNNs and HPC simulations, the new DGX A100 system (see Appendix A) includes eight A100 GPUs connected by the new NVLink-enabled NVSwitch. Multiple DGX A100 systems can be connected via a networking fabric like Mellanox InfiniBand and Mellanox Ethernet to scale out data centers, creating very powerful, even supercomputer-class systems. More powerful NVIDIA DGX POD™ and NVIDIA DGX SuperPOD™ systems will include multiple DGX A100 systems to provide much greater compute power with strong scaling.

Source: Nvidia; KGI Research

Figure 7: NVLink interconnect bandwidth for A100 GPUs

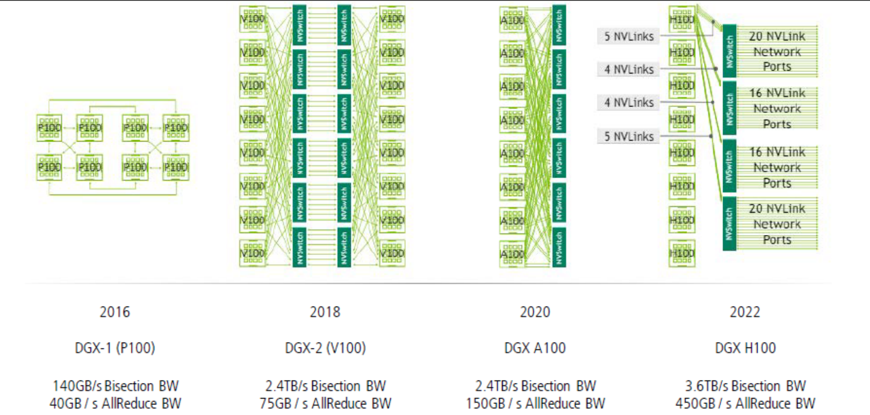


Source: Nvidia; KGI Research

Introduction to NVSwitch

Due to an increasing number of GPUs connected by NVLink interconnects, leading to demand for signal processing and bridging, among others, an NVSwitch framework was required. NVSwitch Gen 1 was launched along with V100 GPUs and NVLink 2, connecting GPUs at different nodes through NVSwitch technology.

Figure 8: GPUs can be connected between multiple nodes via NVSwitch



Source: Nvidia; KGI Research

Figure 9: NVLink Switch specs

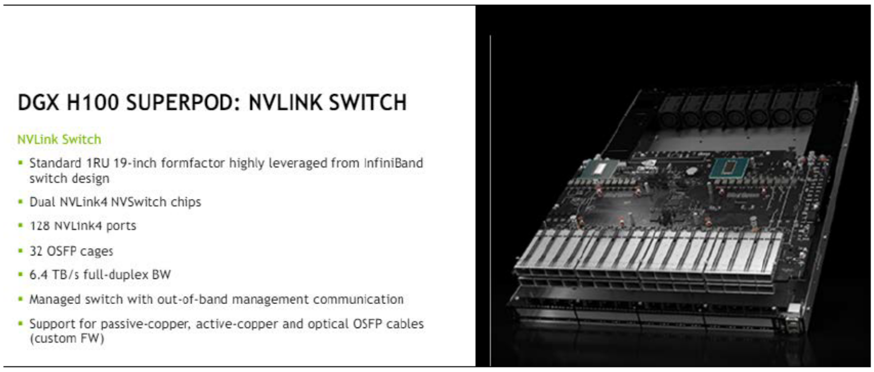
	Gen 1	Gen 2	Gen 3	Gen 4
Number of GPUs with direct connection within a NVLink domain	Up to 8	Up to 8	Up to 8	Up to 576
NVSwitch GPU-to-GPU bandwidth	300GB/s	600GB/s	900GB/s	1,800GB/s
Total aggregate bandwidth	2.4TB/s	4.8TB/s	7.2TB/s	1PB/s
Supported NVIDIA architectures	Volta	Ampere	Hopper	Blackwell

Source: Nvidia; KGI Research

Introduction to NVLink Switches

Compared with NVSwitch chips, which were mostly configured within a single node in accordance with NVLink demand, NVLink Switch is bundled with NVSwitch chips, and integrates the chips into a completely independent device when connected to multiple systems. Take the DGX H100 SUPERPOD system, which is bundled with NVLink Switches, as an example. It has two NVSwitch Gen 3 chips to support NVLink 4 interconnects. Each chip supports 64 NVLink 4 ports, for a total of 128 NVLink 4 ports. The interface that connects to outside devices has 32 OSFP cages, with each cage connecting a maximum of four NVLink interconnects.

Figure 10: NVLink Switch used in DGX H100 SUPERPOD



Source: Nvidia; KGI Research

Introduction to H100 series

In H100 servers, NVLink 4 interconnects are used, which provide bandwidth of 900GBps, or 50% higher than NVLink 3. One H100 GPU can support up to 18 NVLink interconnects. Therefore, the bandwidth of each NVLink interconnect is 50GBps. Since 1Byte = 8bits, the speed of each NVLink can be converted into to 400Gbps.

Each NVSwitch Gen 3 can provide 3.2TBps full-duplex bandwidth on 64 NVLink ports. As 1Byte = 8bits, 3.2TB converts to 25.6Tb. Based on 64 NVLink ports, a single port is 400Gbps.

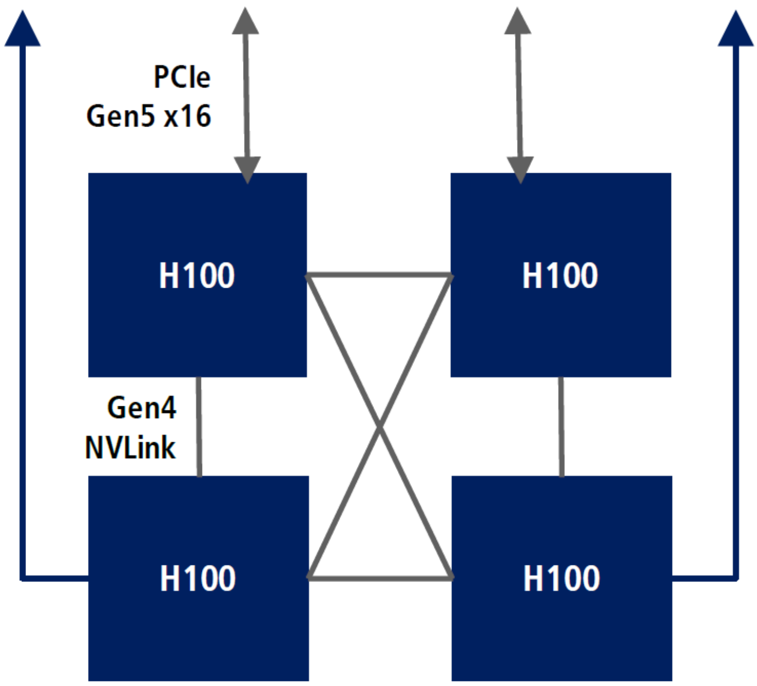
HGX H100 4-GPU – 1.5 NVLink interconnects per GPU

In the HGX H100 4-GPU framework, four H100 GPUs connect with each other directly through NVLink 4 interconnects. The minimum number of NVLink interconnects is calculated by choosing two of the four GPUs as follows:

$$C_2^4 = \frac{4!}{2! \cdot (4-2)!} = 6$$

That is, each HGX H100 4-GPU needs at least six NVLink interconnects (1.5 NVLink interconnects per GPU).

Figure 11: HGX H100 4-GPU framework

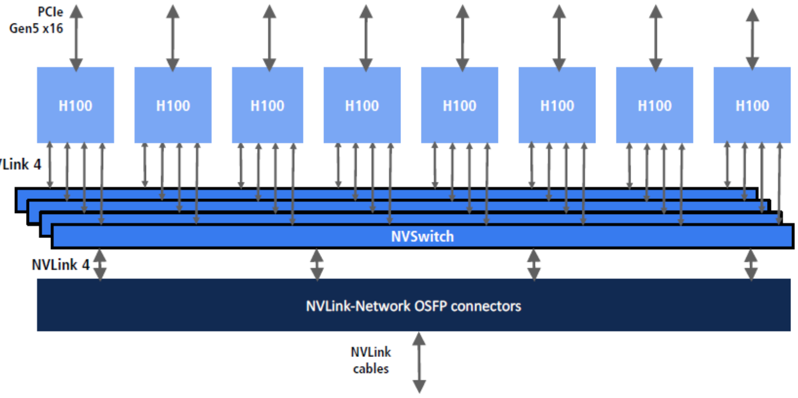


Source: Nvidia; KGI Research

HGX H100 8-GPU – 18 NVLink interconnects per GPU

In the HGX H100 8-GPU framework, there are eight H100 GPUs. Each GPU fully connects to four sets of NVSwitch Gen 3 via NVLink 4 interconnects.

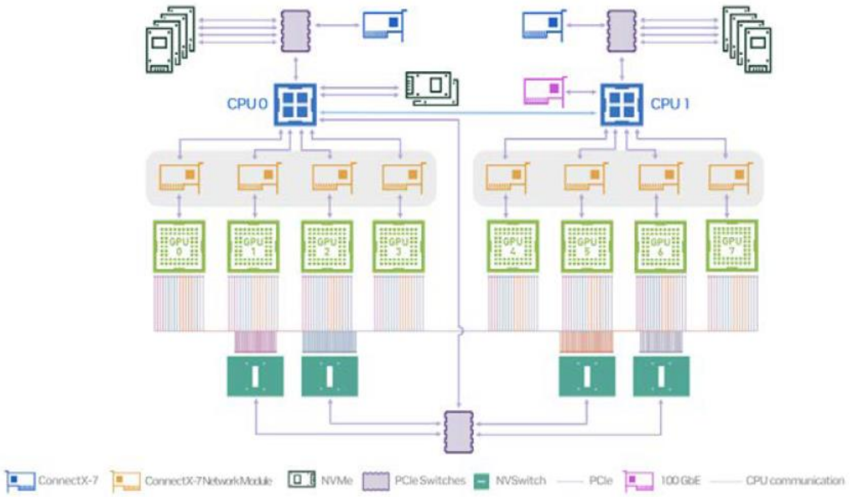
Figure 12: HGX H100 8-GPU framework



Source: Nvidia; KGI Research

However, in the HGX H100 8-GPU framework, the distribution of NVLink interconnects bundled with NVSwitch is not even. In order to protect the effectiveness of the connections with other HGX H100 8-GPU nodes, two NVSwitch chips provide five NVLink interconnects per GPU internally, meaning eight GPUs and two NVSwitch chips connect through a total of 80 NVLink interconnects. The other two NVSwitch chips provide four NVLink interconnects per GPU, meaning eight GPUs and two NVSwitch chips provide a total of 64 NVLink interconnects. Therefore, we estimate that in a single HGX H100 8-GPU system, there are 144 NVLink interconnects, or 18 NVLink interconnects per GPU, the maximum allowed per GPU by design.

Figure 13: HGX H100 8-GPU framework



Source: Nvidia; KGI Research

DGX H100 SuperPOD – 22.5 NVLink interconnects per GPU

DGX H100 SuperPOD is HGX H100 8-GPU based. It can expand to a maximum of 32 computing nodes or 256 GPUs. Its network topology has a 2:1 tapered fat-tree framework. With 256 GPUs, the maximum bandwidth is 57.6TBps, equal to 460.8Tbps.

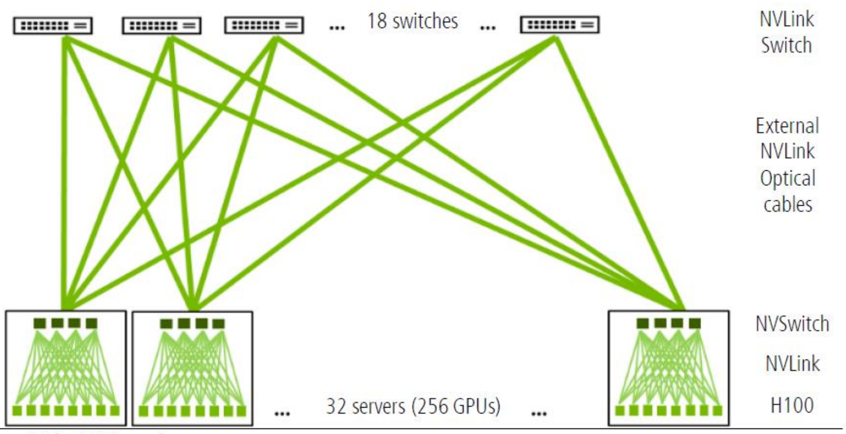
In order to allow full series connections for GPUs, the spine network framework of DGX H100 SuperPOD is bundled with 18 NVLink Switches for data exchange purposes. A standard NVLink Switch at 1U height has 32 OSFP sockets. Each NVLink Switch contains two NVSwitch Gen 3 chips, each supporting 64 NVLink 4 interconnects and providing a total of 128 NVLink 4 ports. The total bandwidth is 6.4TBps, equivalent to 51.2Tbps, with a corresponding 25.6Gbps per NVSwitch Gen 3.

Figure 14：A100 SuperPod & H100 SuperPod comparision

	A100 SuperPod			H100 SuperPod			Speedup	
	Dense PFLOP/s	Bisection【GB/s】	Reduce【GB/s】	Dense PFLOP/s	Bisection【GB/s】	Reduce【GB/s】	Bisection	Reduce
1 DGX / 8 GPUs	3	2,400	150	16	3,600	450	**1.5x**	**3x**
32 DGX / 256 GPUs	80	6,400	100	512	57,600	450	**9x**	**4.5x**

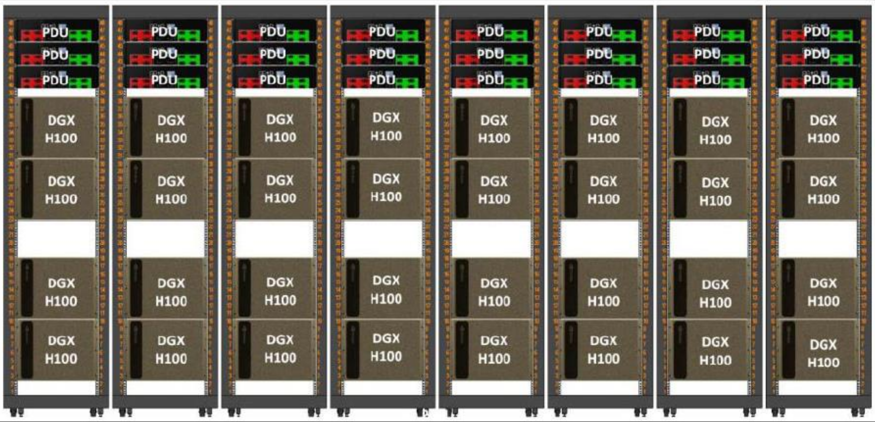
Source: Nvidia; KGI Research

Figure 15: DGX H100 SuperPOD supports maximum of 256 H100



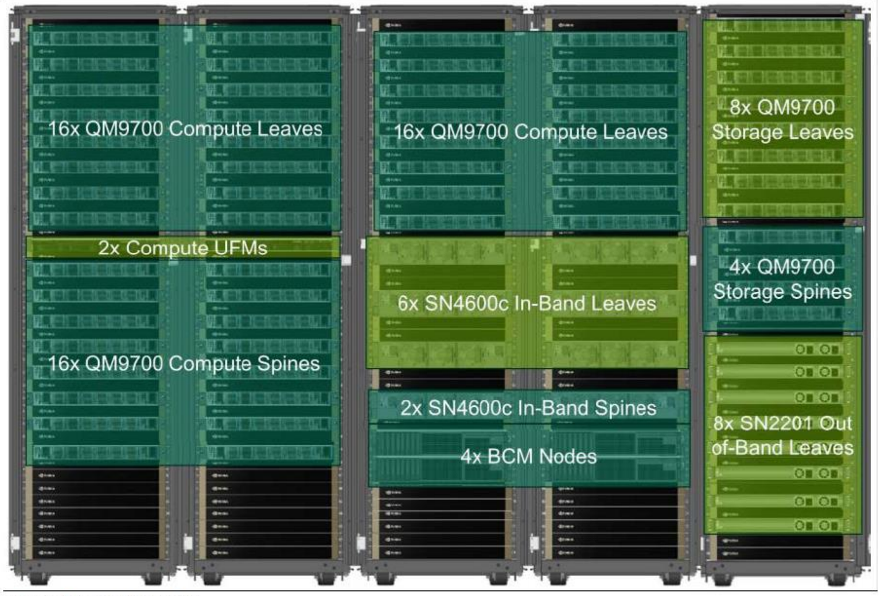
Source: Nvidia; KGI Research

Figure 16: DGX H100 SuperPOD single SU framework



Source: Nvidia; KGI Research

Figure 17: DGX H100 SuperPOD management rack framework

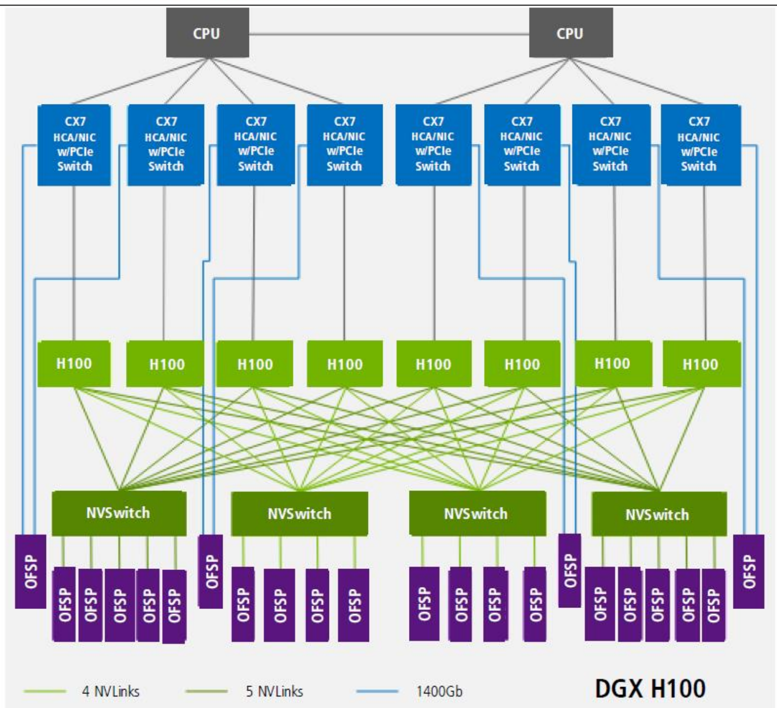


Source: Nvidia; KGI Research

According to the DGX H100 network deployment framework, the NVSwitch adopts OSFP ports for external connections, and each OSFP port is formed by two 400Gbps. Similar way also appears in the explanation for DGX H100's NVSwitch. The explanation also mentions that wires can stretch up to 20 meters from 5 meters.

With a maximum of 256 H100 GPUs, there are 18 NVLink interconnects per GPU in every DGX H100-8GPU system, for a total of 144 NVLink interconnects, as described above. In addition, in each NVLink Switch, a second layer of the framework where there are 18 spines, there are two NVLink Switch chips, and 32 DGX H100 systems will increase this number to 1,152 second-layer NVLink interconnects (18 x 2 x 32 = 1,152). This directly implies a maximum of 144 NVLink connection lines per node. As there are 32 nodes in DGX H100 SuperPOD 256, there are a maximum of 4,608 NVLink interconnects in such a system (144 x 32 = 4,608). Adding this to 1,152 interconnects equals 5,760 NVLink interconnects. 1,152 second-layer NVLink connections are 50GBps each, and the total is 1,152 x 50 = 57,600 GBps, which results in a rate calculation of 57,600GBps (57.6TBps) in Figure 14.

Figure 18: DGX H100 network configuration



Source: Nvidia; KGI Research

GH200

There are eight Grace Hopper GPUs in a GH200 system, each supporting 18 NVLink 4 interconnects, and the NVLink topology consists of three NVLink Switch trays with two NVLink Switch chips in each tray, for a total of six NVLink Switch chips. The ratio of GPUs to NVSwitch Chips is 8:6, or 4:3.

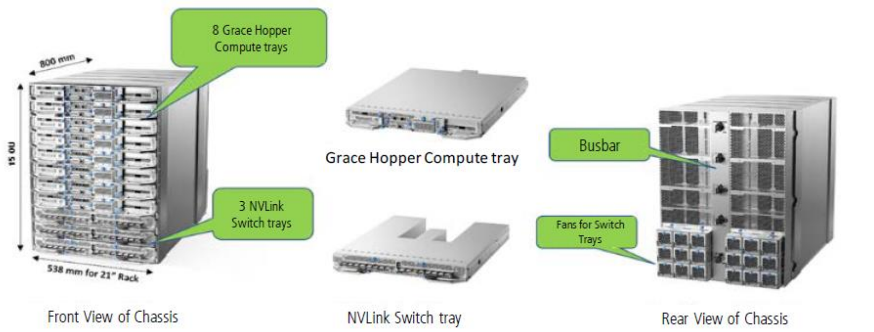
Each NVLink Switch provides 36 OSFP ports. It is noteworthy that that the 18 NVLink interconnects for each GPU are evenly distributed among the six NVLink Switch chips. There are eight GPUs, resulting in a total of 144 NVLink connections. Each NVLink Switch tray is divided into six NVLink interconnects, with passive cables and custom cable cartridges. There are 12 OSFP ports in the second layer of external NVLink Switch ports, and thus in every GH200 server we find a total of 36 OSFP ports in three NVLink Switch trays.

Figure 19: Grace Hopper specs

Compute	CPU/GPU	1x NVIDIA Grace Hopper Superchip with NVLink-C2C
	CPU/GPU	18x NVLink fourth-generation ports
	Networking	1x NVIDIA ConnectX-7 with OSFP: >NDR400 InfiniBand Compute Network 1x Dual port NVIDIA BlueField-3 with 2x QSFP112 or 1x Dual port NVDIA ConnectX-7 with 2x QSFP112: >200 GbE In-band Ethernet network >NDR200 IB storage network Out of Band Network: >1 GbE RJ45
	Storage	Data Drive: 2x 4 TB (U.2 NVMe SSDs) SW RAID 0 OS Drive: 2x 2 TB (M.2 NVMe SSDs) SW RAID 1
Switch	NVSwitch	2x Third-Generation NVSwitch ASIC supporting NVLink fourth-generation
	NVLink Ports	48x NVLink to Compute trays through passive cable cartridge. Inside Chassis > 6x NVLink per Compute tray > 12x OSFP (48x NVLink) to connect to second-level Switches

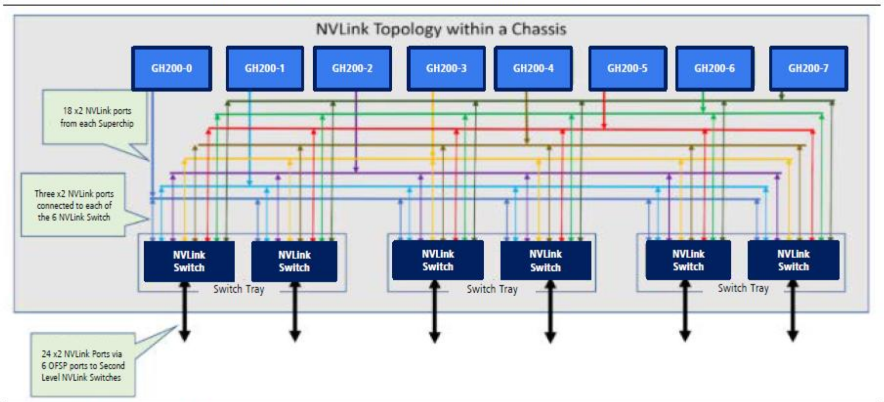
Source: Nvidia; KGI Research

Figure 20: 8-Grace Hopper Superchip chassis



Source: Nvidia; KGI Research

Figure 21: NVLink topology within 8-Grace Hopper Superchip chassis

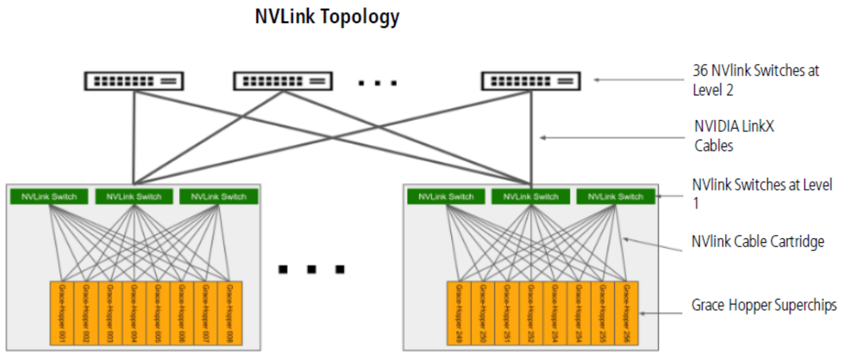


Source: Nvidia; KGI Research

DGX GH200 supercomputer

In the DGX GH200 supercomputer, in addition to the 32 GH200 systems, 32 NVLink Switches have been added to the spine layer, creating a fat-tree architecture. Hence, there are 32 GH200 systems for NVLink, connected by Nvidia LinkX cables. For OSFPs in level 2, there are 36 NVLink Switches, each with 32 OSFP ports, leading to a total of 1,152 level 2 to level 1 OSFP ports. There are 32 GH200 systems in level 1 to level 2, thus there are 36 ports as calculated in the previous paragraph, for a total of 1,152 OSFP ports, which corresponds exactly to the number of OSFPs in level 1 to level 2.

Figure 22: NVLink topology within 8-GraceHopper Superchip chassis



Source: Nvidia; KGI Research

GB200 NVL36/NVL72

In the GB200 NVL36 configuration, each rack hosts 18 single GB200 compute node, and each GB200 corresponds to two GPUs, for a total of 18 GB200s and 36 GPUs. In the GB200 NVL72 architecture, there are 18 dual GB200 compute nodes in a single rack, for a total of 36 GB200s and 72 GPUs. It can also be simply composed of two NVL36 nodes. In the NVL72 architecture (with 72 GPUs), the ratio of GPUs to NVSwitch Chips is 72:18, or 4:1, which is significantly less than the 4:3 ratio of the GH200.

Regarding network topology, the GB200 NVL72 does not adopt fat-tree architecture, but uses a single layer switch architecture instead. Therefore, 72 x 18 = 1,296 NVLink interconnects are needed. The speed for each NVLink interconnect is 100GBps, hence the total support is 1296 x 100GBps = 129,600GBps = 130TBps.

Figure 23: GB200 NVL72 specs

	GB200 NVL72	GB200 Grace Blackwell Superchip
Configuration	36 Grace CPU: 72 Blackwell GPUs	1 Grace CPU 2 Blackwell GPL
FP4 Tensor Core	1,440 PFLOPS	40 PFLOPS
FP8 / FP6 Tensor Core	720 PFLOPS	20 PFLOPS
INT8 Tensor Core	720 POPS	20 POPS
FP16 / BF16 Tensor Core	360 PFLOPS	10 PFLOPS
TF32 Tensor Core	180 PFLOPS	5 PFLOPS
FP64 Tensor Core	3,240 TFLOPS	90 TFLOPS
GPU Memory | Bandwidth	Up to 13.5 TB HBM3e | 576 TB/s	Up to 384 GB HBM3e | 16 TB/s
NVLink Bandwidth	130TB/s	3.6TB/s
CPU Core Count	2,592 Arm® Neoverse V2 cores	72 Arm Neoverse V2 cores
CPU Memory | Bandwidth	Up to 17 TB LPDDR5X | Up to 18.4 TB/s	Up to 480GB LPDDR5X | Up to 512 GB/s

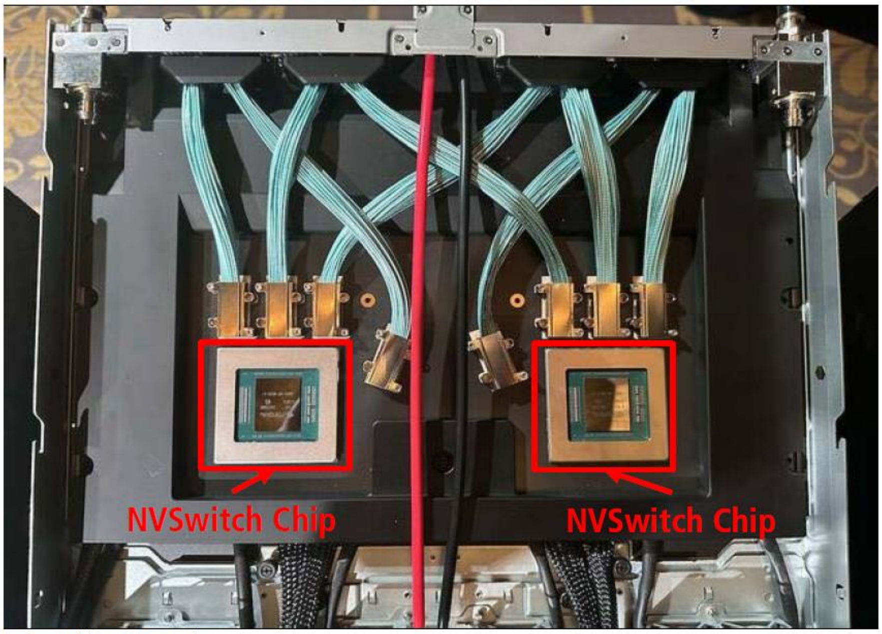
Source: Nvidia; KGI Research

On the other hand, each NVL72 rack will be equipped with nine NVSwitch trays, with each tray containing two NVSwitch chips and 144 100GBps NVLink interconnects, corresponding to nine trays x 144 NVLink interconnects per tray = 1,296 NVLink interconnects. This number corresponds to the number of NVLink interconnects needed by the GPUs in the previous calculation, so the nine NVSwitch will be fully connected to the 72 Blackwell GPUs, with a maximum of 18 NVLink interconnects needed by each GPU.

Figure 24: Exterior of DGX GB200 NVL72

Source: Nvidia; KGI Research

Figure 25: Every NVSwitch tray contains two NVSwitch chips



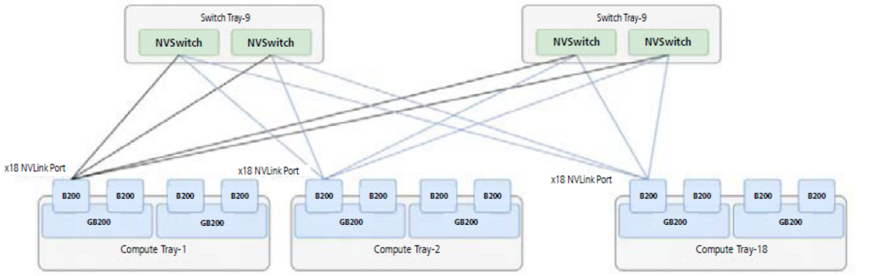
Source: Nvidia; KGI Research

We note that the NVLink 5, developed for the Blackwell platform, will use 72 224Gb SerDes. The speed of SerDes is doubled from NVLink 4, resulting in a total of 14.4Tbps when converting bytes to bits, which is in line with the total bandwidth of 72 Dual 224Gb SerDes.

Figure 26: NVLink Switch chip specs

Source: Nvidia; KGI Research

Figure 27: GB200 NVL72 network topology

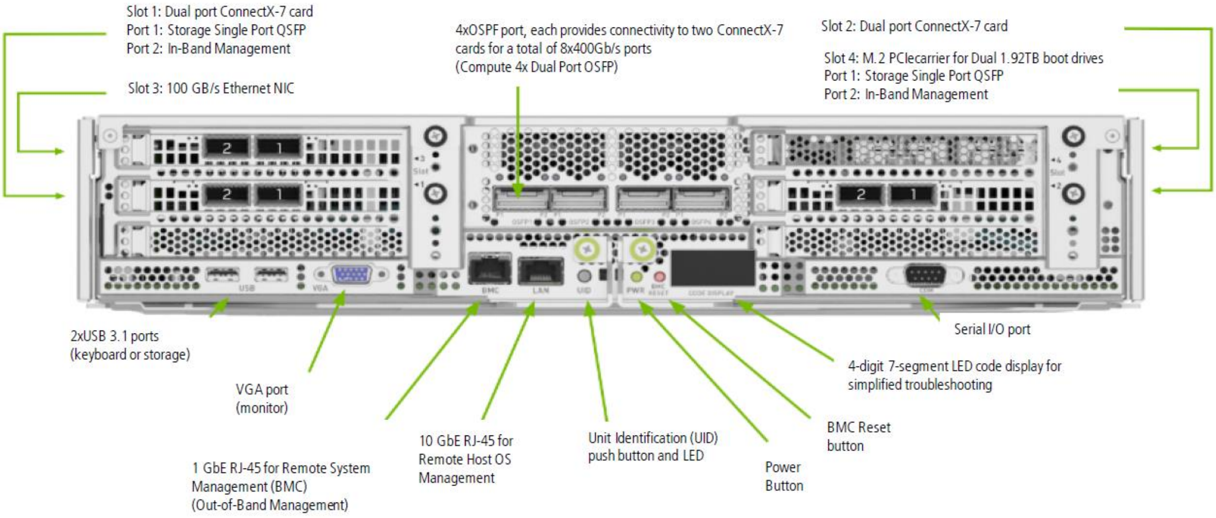


Source: naddod; KGI Research

NVLink and transceiver module conversion

From the previous DGX H100 Network Configuration figure, we see that the OSFP specification is used between the NVLink interconnects, while in the DGX H100 product figure, we see that it uses four OSFP ports, with each port corresponding to two ConnectX-7 cards, for a total of eight 400Gbps ports. The DGX H100 document mentions that each OSFP port supports eight channels of 100G PAM4 signaling. In other words, when using a single DGX H100, each DGX H100 has eight GPUs and a total of four 800Gbps OSFP ports. Assuming each of the connected 800Gbs corresponds to an 800Gbps transceiver module, we see each GPU approximately corresponds to one 800Gbps OSFP transceiver module.

Figure 28: DGX H100 network architecture



Source: Nvidia; KGI Research

Figure 29: DGX H100 NVSwitch

Figure 26. DGX A100 vs DGX H100 32-node, 256 GPU NVIDIA SuperPOD Comparison

Maximum cable length switch-to-switch is increased from 5 meters to 20 meters. OSFP (Octal Small Form Factor Pluggable) LinkX cables made by NVIDIA are now supported. They feature Quad-Port optical transceivers per OSFP, and 8-channels of 100G PAM4 signaling. The Quad-Port OSFP transceiver innovations enable a total of 128 NVLink ports in a single 1 RU, 32-cage NVLink Switch with each port transferring data at 25 GB/sec.

Source: Nvidia; KGI Research

Nvidia has also given a reference for cable requirements. In the H100 generation, we see the number of cables required for node-to-leaf and leaf-to-spine connections are usually the same as the number of GPUs. When the number of SU reaches 16 or more, that is, 4,096 CPUs or more, the network architecture will increase by one layer to serve as the connection between the spine and core. As a result, cable usage will also increase significantly.

Figure 30: Larger SuperPOD component count

SU Count	Node Count	GPU Count	InfiniBand Switch Count			Cable Counts		
			Leaf	Spine	Core	Compute and UFM		Spine-Leaf
1	31*	248	8	4	-	252		256
2	63	504	16	8	-	508		512
3	95	760	24	16	-	764		768
4	127	1016	32	16	-	1020		1024
SU Count	**Node Count**	**GPU Count**	**Leaf**	**Spine**	**Core**	**Node-Leaf**	**Leaf-Spine**	**Spine-Core**
4	128	1024	32	16	-	1024	1024	1024
8	256	2048	64	32	-	2048	2048	2048
16	512	4096	128	128	64	4096	4096	4096
32	1024	8192	256	256	128	8192	8192	8192
56	2048	16384	512	512	256	16384	16384	16384

* This is a 32 node per SU design, however a DGX system must be removed to accommodate for UFM connectivity.

Source: Nvidia; KGI Research

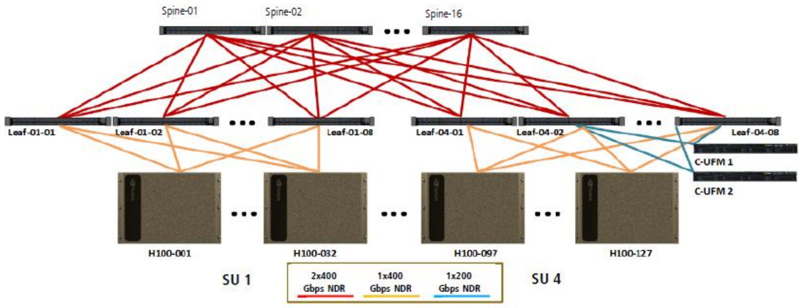
In terms of architecture, we see that two 400Gbps networks are used in the spine-to-leaf layer, while 400Gbps connections are used between the node and the leaf. Similarly, since each link has a head and a tail, we reckon an architecture with less than 16 SUs would have two 400Gbps transceiver modules and two 800Gbps OSFP transceiver modules per GPU. For architecture with more than 16 SUs, it is necessary to use 800Gbps OSFP transceiver modules at the spine-to-core level. Hence, two 400Gbps transceiver modules and four 800Gbps OSFP transceiver modules will correspond to each GPU.

Figure 31: Network planning at the compute level

Compute-InfiniBand Fabric

Figure shows the compute fabric layout for the full 127-node DGX SuperPOD. Each group of 32 nodes is rail-aligned. Traffic per rail of the DGX H100 systems is always one hop away from the other 31 nodes in a SU. Traffic between nodes, or between rails, traverses the spine layer.

Compute InfiniBand fabric for full 127 node DGX SuperPOD



Source: Nvidia; KGI Research

Official configuration offered by Nvidia complies with our calculations.

Figure 32: Estimate of components required for a four SU, 127-node DGX SuperPOD

Count	Component	Recommended Model
Racks		
38	Rack(Legrand)	NVIDPD 13
Nodes		
127	GPU Nodes	DGX H100 system
4	UFM appliance	NVIDIA Unified Fabric Manager Appliance 3.1
5	Management Server	Intel based x86 2xSocket, 24 core or greater, 384 GB RAM, OS (2x480GB M2 or SATA/SAS SSD in RAID 1), NVME 7.68 TB (raw), 4x HDR200 VPI Parts, TPM 2.0
Ethernet Network		
8	In-band management	NVIDIA SN4600C switch with Cumulus Linux
8	OOB management	NVIDIA SN2201C switch with Cumulus Linux
Compute infiniBand Fabric		
48	Fabric switches	NVIDIA Quantum QM9700 switch, 920-9B210-00FN-OMO
Storage infiniBand Fabric		
16	Fabric switches	NVIDIA Quantum QM9700 switch, 920-9B210-00FN-OMO
PDUs		
96	Rack PDUs	Raritan PX3-587812R-P1Q2R1A15D5
12	Rack PDUs	Raritan PX3-58747V-V2

Count	Component	Connection	Recommended Mode	Cable Type
In-Band Ethernet Cables				
254	100 Gbps	DGX H100 system	Varies	
32	100 Gbps QSFP to QSFP AOC	Management nodes	Varies	
6	100 Gbps	ISL Cables	Varies	
Varies	Ethernet (perf varies)	Storage	Varies	
Varies	Varies	Core DC	Varies	
OOB Ethernet Cables				
127	1 Gbps	DGX H100 systems	Cat5e	
64	1 Gbps	InfiniBand Switches	Cat5e	
11	1 Gbps	Management/UFM nodes	Cat5e	
8	1 Gbps	In-band Ethernet switches	Cat5e	
Varies	1 Gbps	Storage	Cat5e	
108	1 Gbps	PDUs	Cat5e	
16	100 Gbps	Two uplinks per OOB to in-band	Varies	
Compute InfiniBand Cabling				
2040	NDR Cables, 400 Gbps	DGX H100 systems to leaf, leaf to spine	980-9I57X-00N010	Fiber
2	NDR Cables, 200 Gbps	UFM to leaf ports	980-9I111-00H010	AOC
1536	Switch OSFP Transceivers	Leaf and spine transceivers	980-9IA2O-00NS00	
508	System OSFP Transceivers	Transceivers in the DGX H100 Systems	980-9I89P-00N000	
4	UFM System Transceivers	UFM to leaf connections	980-9I89R-00NS00	
Storage InfiniBand Cables				
494	NDR Cables, 400 Gbps	DGX H100 systems to leaf, leaf to spine	980-9I57X-00N010	Fiber
48	NDR Cables, 200 Gbps	Storage	980-9I11-00H010	AOC
4	UFM System Transceivers	UFM to leaf connections	980-9I51S-00NS00	
369	Switch Transceivers	Leaf and spine transceivers	980-9I510-00NS00	
254	DGX System Transceivers	QSFP112 transceivers	980-9I693-00NS00	
2	NDR Cables, 200 Gbps	UFM to leaf ports	980-9I557-00N030	Fiber
4	HDR 400 Gbps to 2x200 Gbps	Slurm management	980-9I117-00H030	AOC
Varies	Storage Cables, NDR200	Varies	980-9I117-00H030	AOC

Source: Nvidia; KGI Research

DGX GH200 Supercomputer

Since the link between a single GH200 internal GPU and the NVLink Switch has a special cable specification, it is not applicable to general transceiver modules. We thus only calculate the specification between the GH200 systems. From the previous calculation of 1,152 connections with OSFP specification at both ends, it can be converted to 2,304 units of 800Gbps transceiver modules, equivalent to nine 800Gbps transceiver modules per GPU.

DGX B200 SuperPod

For the DGX B200 SuperPod, the configuration is similar to the Hopper series. However, the general length has been changed as shown in the table below.

Figure 33: Estimate of conponent required for a 4 SU, 127-node DGX SuperPOD

Count	Component	Recommended Model
Racks		
38	Rack (Legrand)	NVIDPD13
Nodes		
127	GPU nodes	DGX B200 system
4	UFM appliance	NVIDIA Unified Fabric Manager Appliance 3.1
5	Management servers	Intel based x86 2 × Socket, 24 core or greater, 384 GB RAM, OS (2×480GB M.2 or SATA/SAS SSD in RAID 1), NVME 7.68 TB (raw), 4× HDR200 VPI Ports, TPM 2.0
Ethernet Network		
8	In-band management	NVIDIA SN4600C switch with Cumulus Linux, 64 QSFP28 ports, P2C 920-9N302-00F7-0C
8	OOB management	NVIDIA SN2201 switch with Cumulus Linux, 48 RJ45 ports, P2C, 920 9N110-00F1-0C0
Compute InfiniBand Fabric		
48	Fabric switches	NVIDIA Quantum QM9700 switch, 920-9B210-00FN-0M0
Storage InfiniBand Fabric		
16	Fabric switches	NVIDIA Quantum QM9700 switch, 920-9B210-00FN-0M0
PDUs		
96	Rack PDUs	Raritan PX3-5878I2R-P1Q2R1A15D5
12	Rack PDUs	Raritan PX3-5747V-V2

Count	Component	Connection	Recommended Mode	Cable type
In-Band Ethernet Cables				
254	200 Gbps QSFP56 to QSFP56 AOC	DGX B200 system	980-9I4A0-00H030	
8	100 Gbps QSFP28 to QSFP28 AOC	Management nodes	980-9IA30-00C030	
4	100 Gbps QSFP28 CWDM4 Single mode 2km Transceiver	Uplink to core DC	980-9I17Q-00CM00	
6	100 Gbps QSFP-QSFP DAC Passive Copper cable	ISL Cables	980-9I620-00C00	
8	100 Gbps QSFP28 to QSFP28 AOC	NFS Storage	980-9I13N-00C03	
24	100 Gbps QSFP28 to QSFP28 AOC	Leaf – Core cables	980-9I13N-00C03	
OOB Ethernet Cables				
127	1 Gbps	DGX B200 systems	Cat5e	
64	1 Gbps	InfiniBand Switches	Cat5e	
8	1 Gbps	Management/UFM nodes	Cat5e	
8	1 Gbps	In-band Ethernet switches	Cat5e	
2	1 Gbps	UFM Back-to-Back	Cat5e	
108	1 Gbps	PDUs	Cat5e	
4	QSFP to SFP+ Adapter	For the UFM connections	980-9I71G-00J000	
4	Ethernet Module SFP BaseT 1G	For the UFM connections	980-9I251-00IS00	
16	100 Gbps AOC QSFP28 to QSFP28 Cable	Two uplinks per OOB to in-band	980-9I13N-00C030	
Varies	1 Gbps	Storage	Cat5e	
Compute InfiniBand Cabling				
2044	NDR Cables, 400 Gbps	DGX B200 systems to leaf, leaf to spine, UFM to leaf ports	980-9I570-00N030	Fiber
1536	Switch 2x400G QSFP Finned-top Multimode Transceivers	Leaf and spine transceivers	980-9I510-00NS0	
508	System 2x400G QSFP Flat-top Multimode Transceivers	Transceivers in the DGX B200 Systems	980-9I51A-00NS00	
4	UFM System 4x400G QSFP Multimode Transceivers	UFM to leaf connections	980-9I510-00NS00	
Storage InfiniBand Cables				
498	NDR Cables, 400 Gbps	DGX B200 systems to leaf, leaf to spine UFM to leaf ports	980-9I570-00N030	Fiber
48	NDR AOC Cables, 2×200 Gbps QSFP56-QSFP56	Storage	980-9I117-00H030	AOC
4	UFM System 4x400G QSFP Multimode Transceivers	UFM to leaf connections	980-9I510-00NS00	
369	Switch 2x400G QSFP Finned-top Multimode Transceivers	Leaf and spine transceivers	980-9I510-00NS00	
254	DGX System 4x400G QSFP112 Multimode Transceivers	QSFP112 transceivers	980-9I693-00NS00	
4	HDR 400 Gbps to 2x200 Gbps AOC Cables	Slurm management	980-9I117-00H030	AOC
Varies	Storage Cables, 400 Gbps to 2x200 Gbps AOC Cables	Varies	980-9I117-00H030	AOC

Source: Nvidia; KGI Research

GB200

As the GB200 NVL36 and NVL72 use direct connections internally, there is no need for network connections at the second layer of the fat tree architecture inside a single cabinet. Therefore, the NVLink interconnects that we see in NVL72 racks will be connected directly. According to the conversion between NVLink and SerDes we mentioned earlier, the GB200 uses 100GBps NVLink 5 per lane (i.e. 800Gbps) interconnects, equivalent to 400Gbps for a unidirectional link. The bandwidth for a pair of SerDes is 224Gbps, thus we can conclude that 400Gbps in one direction will be composed of 2 differential pairs. Each NVLink will be composed of four differential pairs. Therefore, the cable requirements will be four times the number of NVLink 5 interconnects. Based on the fact that the GB200 is a single-layer switch architecture, we note the number of NVLink interconnects in the GB200 is 72 x 18 = 1,296, and we can further calculate that the total number of copper cables will be 1,296 x 4 = 5,184. This calculation is also in line with Nvidia's March 2024 GTC, which stated that there are approximately 5,000 copper cables on the back of the GB200 NVL72.

As the NVLink 5 of the next-generation Blackwell platform will reach 1.8TBps, and the NVLink 6 used in the generation after Blackwell (Rubin platform) will reach 3.6TBps, we believe that the number of NVLink interconnects and SerDes specifications will continue to evolve. In addition to suppliers of 224Gbps SerDes chips, we believe network cable, connector, and splitter manufacturers in Taiwan, such as Browave (3163 TT, NT$110.5, NR), Jess-Link Products (6197 TT, NT$177.5, NR) and Bizlink (3665 TT, NT$304, OP) will benefit from migrations to the Blackwell server platform.

Figure 34: Product mix & 2024 development plan for optical communication-related stocks

Company	Ticker	Product Mix		Product Description	2024 Product Development Plan
		2022	2023		
LandMark Optoelectronics	3081 TT	- Epi-wafers 99.3% - Others 0.7%	- Epi-wafers 99% - Others 1%	- InP LD epi-wafers: mainly used in optical communication and data transmission - InP LD epi-wafers: mainly used in optical communication and data transmission - PD epi-wafers: mainly used in optical communication - GaAs LD epi-wafers: mainly used for high-power laser machining, sensor, and data	- more than 50Gbit/s DFB/EML epi-wafers - 6'' multi-structured VCSEL - High-power DFB Laser epi-wafers - Single photon APD epi-wafer - Over 1.7 long-wavelength PD epi-wafer - 1.45 long-wavelength LED epi-wafer
Browave	3163 TT	- Branch 30.5% - WDM 47.8% - OIN 13.3% - AMP 8.3% - Others 0.1%	- Branch 22.2% - WDM 21.2% - OIN 49.9% - AMP 6.3% - Others 0.3%	- Branch focuses on the XGS PON market - The WDM product group focuses on the telecommarket and Cable TV broadband - OIN focuses on data center applications, including MPO and AOC - AMP focuses on the telecommunications market	- Integrated wavelength mux/demux - Mini isolator/circulator - 2D fiber/collimator array - Fiber Harness - Automatic multi-channel alignment - CPO Fiber Connection Module
TrueLight	3234 TT	- Chips and components 71.3% - Optical transmission & connection modules 25.8% - Others 2.9%	- Chips and components 73% - Optical transmission & connection modules 24% - Others 3%	- VCSEL、FP/DFB、PIN/PINTIA for optical fiber communication, 4G/5G mobile communication base station interconnection, cloud data center, 3D Sensing/Near-Field Sensing/Flood Illumination	- VCSELs for various applications - 10Gbps and 25Gbps FP/DFB die, PIN/APD and OSA - High-Power DFB Light Sources - 100G QSFP28 SR4/400G QSFP56-DD SR8 - 56G/112G GaAs PD, 56G/112G InGaAs PD - CPO Package
FOCI	3363 TT	- Fiber jumpers 76.3% - Micro-optical fiber devices 11.2% - Fiber couplers 4.6% - Other passive products 3.3% - Fiber connectors 1.4% - Rental receipt 0.7% - Others 2.8%	- Fiber jumpers 78% - Micro-optical fiber devices 5.7% - Fiber couplers 5.5% - Other passive products 5.3% - Fiber connectors 1.3% - Rental receipt 0.8% - Others 3.4%	-Produces and sells optical passive components such as fiber jumpers couplers. Its main customers are optical fiber communication manufacturers and communication equipment manufacturers, etc.	- PM FA packaging - CPO Silicon photonic application connector product developmer - USB3.2、USB4、DP Alt Type C、HDMI2.1 AOC - Reflowable Lensed Fiber Array Connector
Elite Advanced Laser	3450 TT	- Power semiconductor 77.1% - Optical information & communication products 22.9%	- Power semiconductor 81.9% - Optical information & communication products 18.2%	- PA is used in computer products, handheld devices, automotive electronics, etc. - Optical communication products are mainly GPON and EPON TO-CAN packaging, mainly used in FTTx, DCs, 4G/LTE base stations, etc. - The main applications of optical information products are video recorders, 3D sensing, auto HUD, and auto LiDAR	- 5G mobile communication related components - Array low-power non-temperature-controlled ELS module - GaN on Silicon epi-wafers - Silicon base fiber splicing technologies - Tx components of DWDM thermal control system - 800G LPO
Apac Opto Electronics	4908 TT	- Optical transceiver module 93.7% - Others 6.3%	- Optical transceiver module 96.5% - Others 3.6%	- Produces and sells optical transceiver modules and connectors, mainly used in network and communication equipment, data transmission equipment and cable TV network equipment, etc.	- 800G OSFP SR8/DR8 - 800G QSFP-DD SR8/DR8
PCL Technologies	4977 TT	- SFP+ 51.7% - QSFP 37% - OSA 4.8% - SFP 1.8% - Others 4.7%	- SFP+ 53.6% - QSFP 37.5% - OSA 3% - SFP 0.5% - Others 5.4%	- Produces and sells optical transceiver modules and OSA - SFP+/QSFP are mainly used in Telecom/Ethernet/Datacom/Cloud computing /Storage - XFP is mainly used in Telecom/Ethernet/Datacom/Cloud computing - OSA/SFP are mainly used in Telecom/Ethernet	- 64G SMF SFP28 LR Transceiver - 64G MMF SFP28 SR Transceiver - 25G SMF SFP28 BIDI Transceiver - 32G MMF SFP28 SR Transceiver Gen2 - 1.6T CPO Remote Laser Module
LuxNet	4979 TT	- Component & transceiver 85.3% - Chip 9.4% - Others 5.4%	- Component & transceiver 92.5% - Chip 4.6% - Others 2.9%	- Specializes in products and services such as active components of optical communications (chips, TO-CAN, OSA), and OEM services for optical communications' transceiver module - Products are mainly used in 5G transmission and data centers applications	- Focuses 100mW CW LD and advanced packaging technology for 800G+ applications
JPC	6197 TT	- Smart Connection Industry (SCI) 46.5% - Datacenter/Networking/Telecom (DNT) - IoT 2.9% - Others 7.3%	- SCI 44.8% - DNT 44.6% - IoT 2.6% - Others 8%	- DNT is used for hyperscale data center, AI servers, storage, 5G telecom, edge computing, switch, and other high-speed transmission copper wires and optical fiber and optical module interconnection products.	- 800G 8X OSFP/QSFP-DD DAC/ACC/cable - 400G Data center DR4 transceiver - 800G OSFP SR8/SR4 transceiver
EZconn	6442 TT	- RF connectors 30.1% - Optical communication 69.9%	- RF connectors 23% - Optical communication 77%	- RF connectors can be classified into electronic and non-electronic categories. Electronics are mainly used in cable TV STB; non-electronics are mainly used in auto, aerospace, etc. - Optical transceiver modules and OSA are mainly used in network and communication, data transmission and cable TV network - Optical passive components (including jumpers, connectors, etc.) are used in data centers.	- DOCSIS 4.0 Filters - Photonic IC - 50Gbps PON ONU BOSA/Transceiver - 100Gbps C-PON - 1.6T OSFP-XD
ShunSin Technology	6451 TT	- Optical TXR 75% - SiP 21% - Others 4%	- Optical TXR 59% - SiP 33% - Others 8%	- SiP products are mainly high-frequency wireless communication modules, WiFi modules, LNA, sensors and automotive electronics. - Optical TXR packaging and testing services: mainly used for storage and transmission of enterprise servers and cloud servers	- 800G-2*FR4 Transceiver - 1.6T-OSFP-XD Transceiver - 800G DR8 Optical Engine

Source: Company data; KGI Research

Figure 35: Peer comparison

Company	Code	Market cap (US$ mn)	Share price (LCY)	EPS (LCY) 2024F	EPS (LCY) 2025F	EPS CAGR (%) (2023-2025F)	PER (x) 2024F	PER (x) 2025F	PBR (x) 2024F	PBR (x) 2025F	ROE (%) 2024F	ROE (%) 2025F	Dividend yield (%) 2024F	Dividend yield (%) 2025F
LandMark Optoelectronics*	3081 TT	399	140.5	0.67	7.15	N.M.	208.2	19.7	3.3	3.2	1.6	16.7	0.4	4.3
Browave	3163 TT	257	110.5	N.A.	N.A.	N.M.	N.M.	N.M.	N.M.	N.M.	N.A.	N.A.	N.A.	N.A.
TrueLight	3234 TT	160	46.3	N.A.	N.A.	N.M.	N.M.	N.M.	N.M.	N.M.	N.A.	N.A.	N.A.	N.A.
FOCI	3363 TT	485	159.0	N.A.	N.A.	N.M.	N.M.	N.M.	N.M.	N.M.	N.A.	N.A.	N.A.	N.A.
Elite Advanced Laser	3450 TT	430	95.4	N.A.	N.A.	N.M.	N.M.	N.M.	N.M.	N.M.	N.A.	N.A.	N.A.	N.A.
BizLink*	3665 TT	1,536	304.0	20.86	26.71	36.4	14.6	11.4	1.9	1.7	13.3	15.8	4.1	5.3
Apac Opto Electronics	4908 TT	267	110.5	N.A.	N.A.	N.M.	N.M.	N.M.	N.M.	N.M.	N.A.	N.A.	N.A.	N.A.
PCL Technologies*	4977 TT	186	75.0	0.52	3.92	(6.9)	144.0	19.1	1.5	1.3	1.0	7.5	1.3	4.7
LuxNet*	4979 TT	636	146.0	4.23	5.51	28.4	34.5	26.5	5.9	4.9	17.9	19.3	1.4	1.6
Jess-Link Products	6197 TT	670	177.5	8.12	11.81	53.7	21.9	15.0	N.M.	N.M.	27.8	34.5	2.8	3.2
Ezconn	6442 TT	602	257.5	N.A.	N.A.	N.M.	N.M.	N.M.	N.M.	N.M.	N.A.	N.A.	N.A.	N.A.
ShunSin Technology Holding	6451 TT	705	212.0	N.A.	N.A.	N.M.	N.M.	N.M.	N.M.	N.M.	N.A.	N.A.	N.A.	N.A.
Peer Average							**84.6**	**18.3**	**3.1**	**2.8**	**12.3**	**18.8**	**2.0**	**3.8**

*Source: Bloomberg; KGI Research (*KGI estimates)*

All the above named KGI analyst(s) is SFC licensed person accredited to KGI Asia Ltd to carry on the relevant regulated activities. Each of them and/or his/her associate(s) does not have any financial interest in the respectively covered stock, issuer and/or new listing applicant.