An Overview of NVIDIA NVLink

Posted on Jan 29, 2024 by **Howard** 2.0k

NVIDIA NVLink has emerged as a crucial technology in the fields of high-performance computing (HPC) and artificial intelligence (AI). This article delves into the intricacies of NVLink, and learns about NVSwitch chips, NVLink servers, and NVLink switches, shedding light on its significance in the ever-evolving landscape of advanced computing.

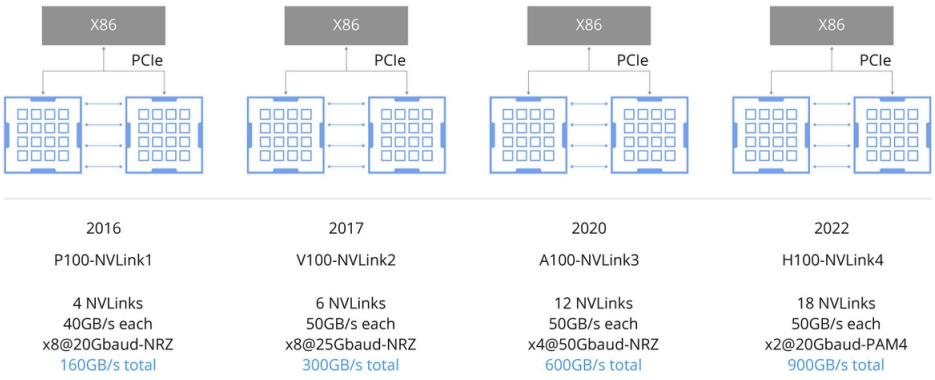
What Is NVIDIA NVLink?

NVLink is a protocol that addresses the communication limitations between GPUs within a server. Unlike traditional PCIe switches, which have limited bandwidth, NVLink enables high-speed direct interconnection between GPUs within the server. The fourth-generation NVLink offers significantly higher bandwidth—112Gbps per lane—compared to PCIe Gen5 lanes, which is three times faster.

PCI Express link performance

Version	Introduced	Line Code		Trandfer rate Per lane	Throughput				
					X1	X2	X4	X8	X16
1.0	2003	NRZ	8b/10b	2.5 GT/s	0.250 GB/s	0.500 GB/s	1.000 GB/s	2.000 GB/s	4.000 GB/s
2.0	2007			5.0 GT/s	0.500 GB/s	1.000 GB/s	2.000 GB/s	4.000 GB/s	8.000 GB/s
3.0	2010		128b/130b	8.0 GT/s	0.985 GB/s	1.969 GB/s	3.938 GB/s	7.877 GB/s	15.754 GB/s
4.0	2017			16.0 GT/s	1.969 GB/s	3.938 GB/s	7.877 GB/s	15.754 GB/s	31.508 GB/s
5.0	2019			32.0 GT/s	3.938 GB/s	7.877 GB/s	15.754 GB/s	31.508 GB/s	63.015 GB/s
6.0	2022	PAM-4 FFC	242B/256B FLIT	64.0 GT/s 32.0 GBd	7.563 GB/s	15.125 GB/s	30.250 GB/s	60.500 GB/s	121.000 GB/s
7.0	2025 (planned)			128.0 GT/s 64.0 GBd	15.125 GB/s	30.250 GB/s	60.500 GB/s	121.000 GB/s	242.000 GB/s

NVLink aims to offer a streamlined, high-speed, point-to-point network for direct GPU interconnections, minimizing overhead compared to traditional networks. By providing CUDA acceleration across different layers, NVLink reduces communication-related network overhead. NVLink has evolved alongside GPU architecture, progressing from NVLink1 for P100 to NVLink4 for H100, as depicted in the figure. The key difference among NVLink 1.0, NVLink 2.0, NVLink 3.0, and NVLink 4.0 lies in the connection method, bandwidth, and performance.

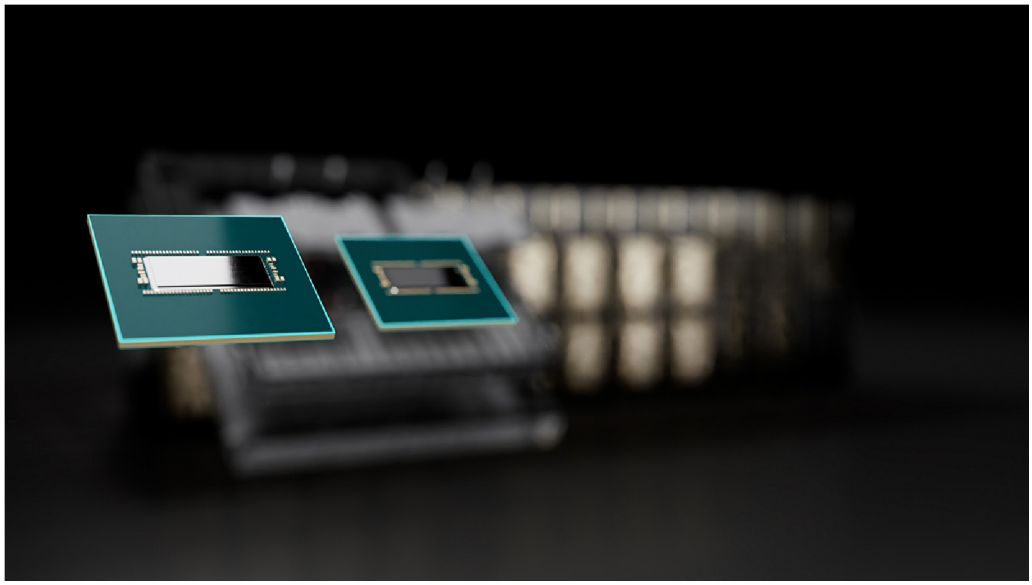


NVSwitch Chip

The NVSwitch chip is a physical chip similar to a switch ASIC that connects multiple GPUs with high-speed NVLink interfaces, improving communication and bandwidth within a server. The third generation of NVIDIA NVSwitch has been proposed and can interconnect each pair of GPUs at a staggering 900 GB/s.

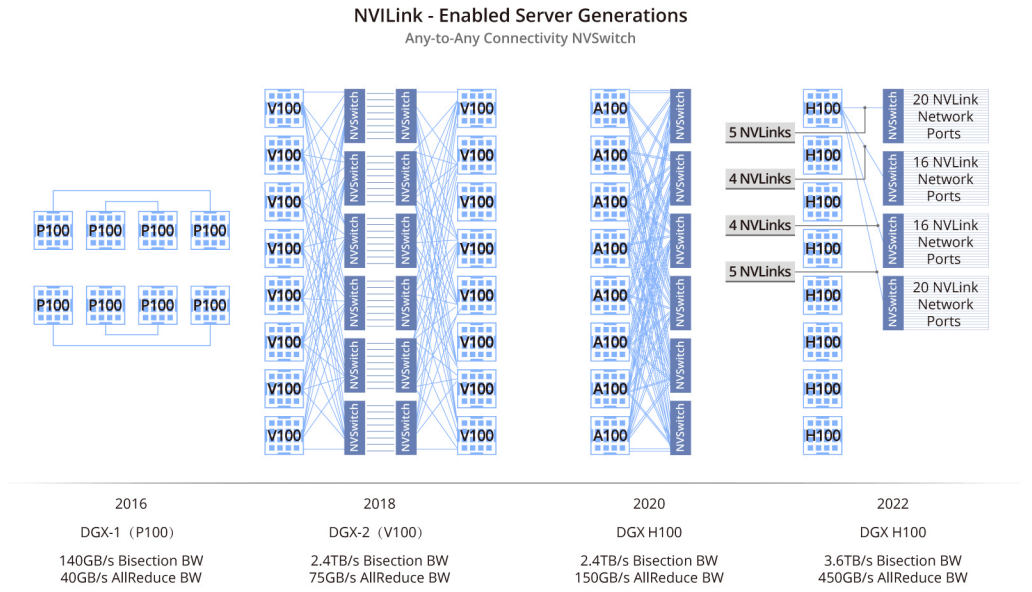
Number of GPUs with direct connection / node	Up to 8	Up to 8	Up to 8
NVSwitch GPU-to-GPU bandwidth	300GB/s	600GB/s	900GB/s
Total aggregate bandwidth	2.4TB/s	4.8TB/s	7.2TB/s
Supported NVIDIA architectures	NVIDIA Volta architecture	NVIDIA Ampere architecture	NVIDIA Hopper Architecture

The latest NVSwitch3 chip, with 64 NVLink4 ports, offers a total of 12.8 Tbps of unidirectional bandwidth or 3.2 TB/s of bidirectional bandwidth. What sets the NVSwitch3 chip apart is its integration of the SHARP function, which aggregates and updates computation results across multiple GPU units during all reduced operations, reducing network packets and enhancing computational performance.

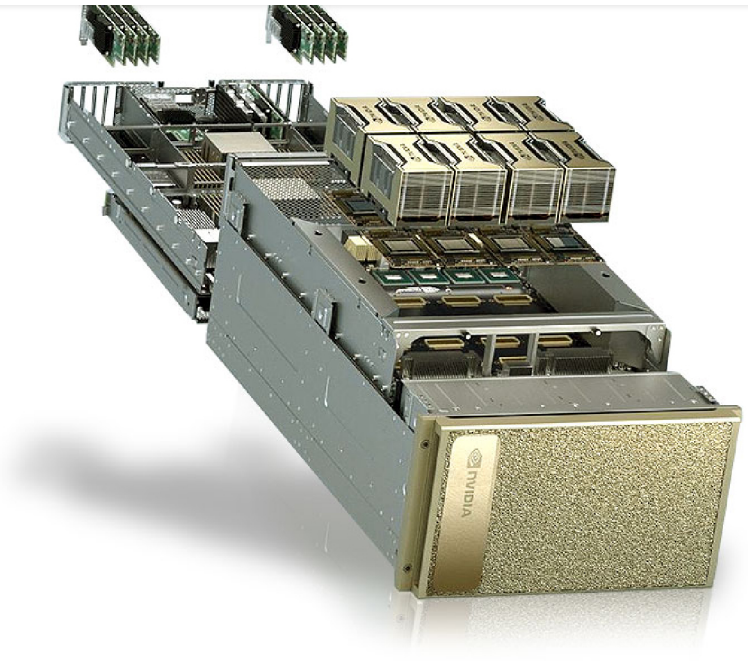


NVLink Server

NVLink servers incorporate NVLink and NVSwitch technologies to connect GPUs, typically found in NVIDIA's DGX series servers or OEM HGX servers with similar architectures. These servers utilize NVLink technology, delivering exceptional GPU interconnectivity, scalability, and HPC capabilities. In 2022, NVIDIA announced the fourth-generation NVIDIA® DGX™ system, the world's first AI platform to be built with new NVIDIA DGX H100 server.

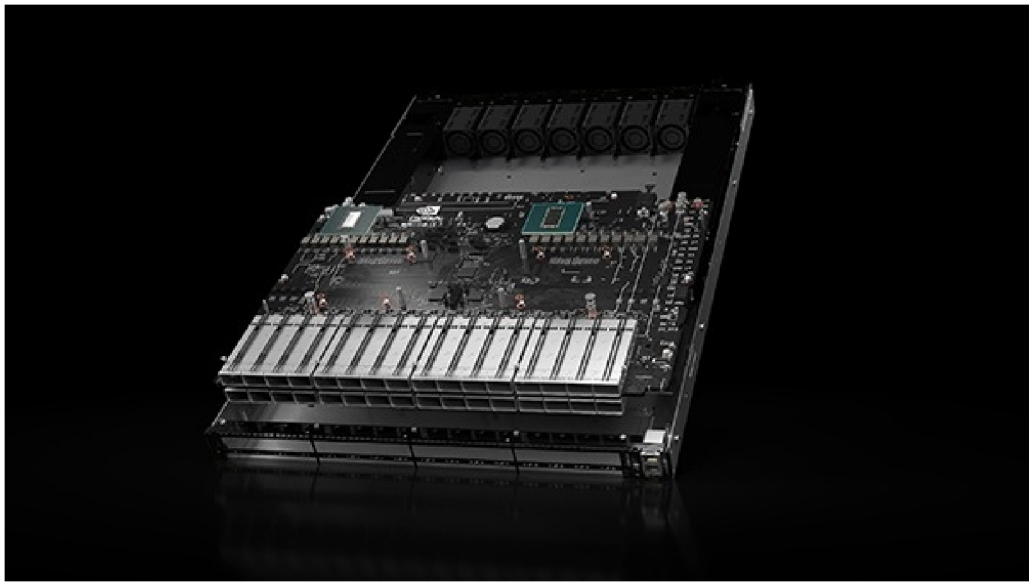


Consequently, NVLink servers have become indispensable in crucial domains such as scientific computing, AI, big data processing, and data centers. By providing robust computing power and efficient data processing, NVLink servers not only



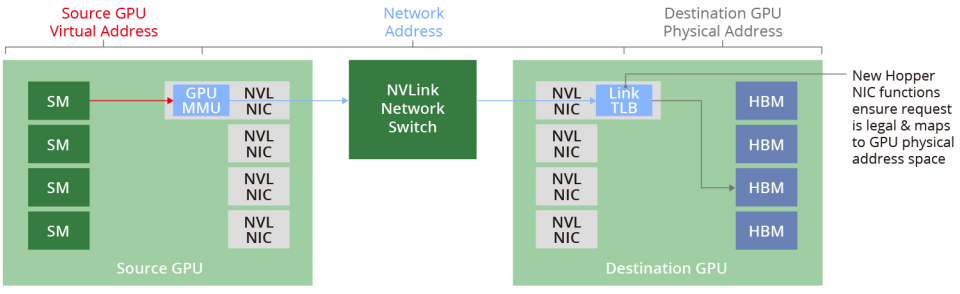
NVLink Switch

In 2022, NVIDIA took out the NVSwitch chip and made it into a switch called the NVLink Switch, which connects GPU devices across hosts. It adopts a 1U size design with 32 OSFP ports; each OSFP comprises 8 112G PAM4 lanes, and each switch has 2 built-in NVSwitch3 chips.



NVLink Network

The NVSwitch physical switch connects multiple NVLink GPU servers into a large Fabric network, which is the NVLink network, solving high-speed communication bandwidth and efficiency issues between GPUs. Each server has its own independent address space, providing data transmission, isolation and security protection for GPUs in the NVLink network. When the system starts, the NVLink network automatically establishes a connection through the software API and can change the address during operation.



	NVLink	NVLink Network
Address Spaces	1 (shared)	N (independent)
Request Addressing	GPU physical address	Network address
Connection Setup	During boot process	Runtime API call by software
Isolation	No	Yes

The figure compares NVLink networks with traditional Ethernet networks, demonstrating the creation of an NVLink network independent of IP Ethernet and dedicated to GPU service.

Concept	Traditional Example	NVLink Network
Physical Layer	400G electrical/optical media	Custom-FW OSFP
Data Link Layer	Ethernet	NVLink custom on-chip HW and FW
Network Layer	IP	New NVLink Network Addressing and Management Protocols
Transport Layer	TCP	NVLink custom on-chip HW and FW
Session Layer	Sockets	SHARP groupsCUDA export of Network addresses of data-structures
Presentation Layer	TSL/SSL	Library abstractions (e.g., NCCL, NVSHMEM)
Application Layer	HTTP/FTP	AI Frameworks or User Apps
NIC	PCIe NIC (card or chip)	Functions embedded in GPU and NVSwitch
RDMA OffLoad	NIC Off-Load Engine	GPU-internal Copy Engine
Collectives OffLoad	NIC/Switch Off-Load Engine	NVSwitch-internal SHARP Engines
Security Off-Load	NIC Security Features	GPU-internal Encryption and "TLB" Firewalls
Media Control	NIC Cable Adaptation	NVSwitch-internal OSFP-cable controllers

Table: Traditional networking concepts mapped to their counterparts with the NVLink Switch System

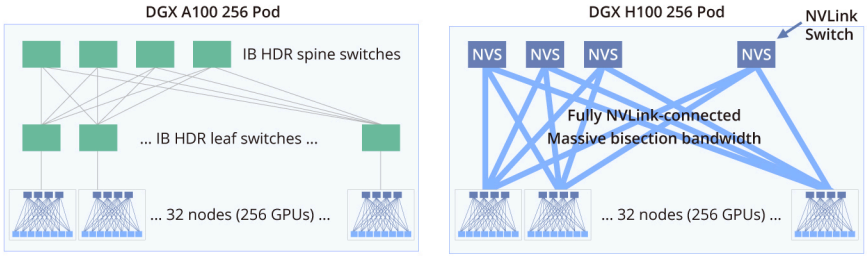
InfiniBand Network VS NVLink Network

InfiniBand Network and NVLink Network are two different networking technologies used in high-performance computing and data center applications. They have the following differences:

Architecture and Design: InfiniBand Network is an open-standard networking technology that utilizes multi-channel, high-speed serial connections, supporting point-to-point and multicast communication. NVLink Network is a proprietary technology by NVIDIA, designed for high-speed direct connections between GPUs.

Application: InfiniBand Network is widely used in HPC clusters and large-scale data centers. NVLink Network is primarily used in large-scale GPU clusters, HPC, AI and other fields.

support fast data exchange and collaborative computing. The following is the bandwidth comparison between the H100 using NVLink network and the A100 using IB network.



	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	2.5	2,400	150	16	3,600	450	1.5x	3x
32 DGXs / 256 GPUs	80	6,400	100	512	57,600	450	9x	4.5x

Also check-Getting to Know About InfiniBand.

Conclusion

NVIDIA NVLink stands as a groundbreaking technology that has revolutionized the fields of HPC and AI. Its ability to enhance GPU communication, improve performance, and enable seamless parallel processing has made it an indispensable component in numerous HPC and AI applications. As the landscape of advanced computing continues to evolve, NVLink's significance and impact are set to expand, driving innovation and pushing the boundaries of what is possible.

Tags

Wiki # Networking Devices # Switches # High-performance Computing # Networking # Artificial Intelligence

You might be interested in

Knowledge

Moris

Layer 2 vs Layer 3 Switch: Which One Do You Need?

Oct 6, 2021 543.5k

Knowledge

John

Fiber Optic Cable Types: Single Mode vs Multimode Fiber Cable

May 10, 2022 846.0k

Knowledge

Sheldon

Running 10GBASE-T Over Cat6 vs Cat6a vs Cat7 Cabling?

Sep 29, 2021 414.0k