NVLINK

FALL 2024

Arian Afzalzadeh

NVLINK

Motivation

NVLink is NVIDIA's high-speed interconnect technology for GPU-accelerated computing. NVLink was first implemented in Tesla P100 accelerator boards and Pascal GP100 GPUs. It significantly increases performance for both GPU-to-GPU communications and for GPU access to system memory. Multiple GPUs are commonly used in the nodes of high-performance computing clusters. Up to eight GPUs per node was typical when NVLink was devised (much more today), and in multiprocessing systems, a powerful interconnect is extremely valuable. The motivation behind NVLink was to create an interconnect for GPUs that would offer much higher bandwidth than PCI Express Gen 3 (PCIe), and be compatible with the GPU ISA to support shared memory multiprocessing workloads.

Today, unlocking the full potential of exascale computing and trillion-parameter AI models hinges on the need for swift, seamless communication among every GPU within a server cluster. The latest generation of NVLink can scale up to 576 GPUs to accelerate performance for trillion- and multi-trillion-parameter AI models. Blackwell GPUs provide 1.8 TB/sec total bandwidth, 900 GB/sec in each direction, which is over 14X the bandwidth of PCIe Gen5. That's nearly seven petabytes of data transfer in an hour from one GPU, or more data than 18 years of streaming 4K movies, or the entire Internet bandwidth processed by just 11 Blackwell GPUs.

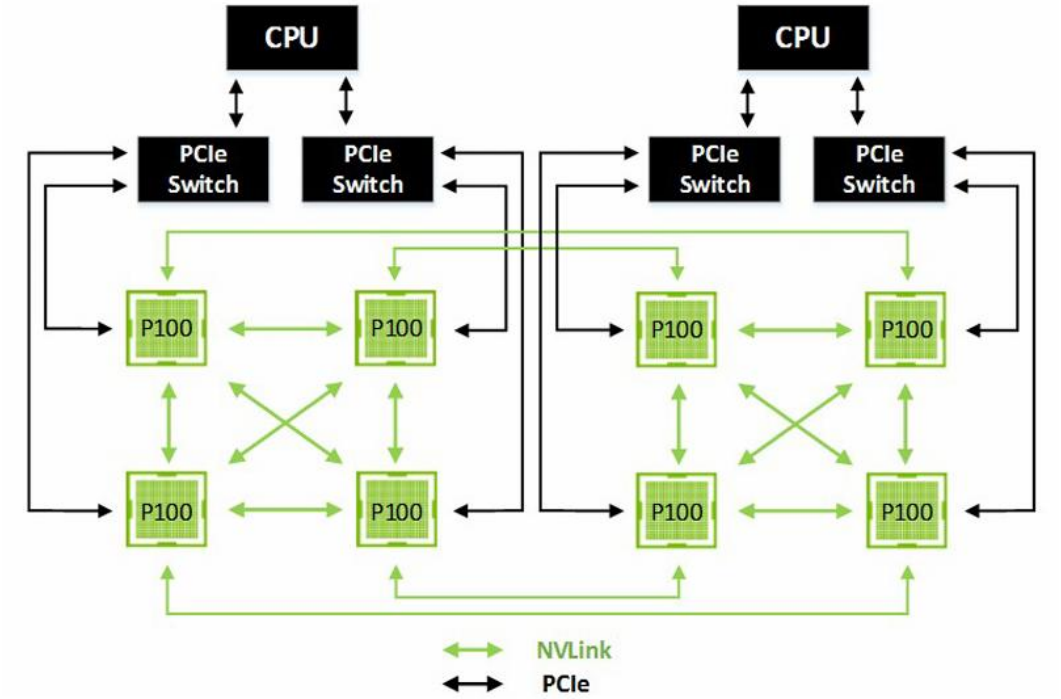
With NVLink-connected GPUs, programs can execute directly on memory that is attached to another GPU as well as on local memory, and the memory operations remain correct (for example providing full support for atomic operations).

NVLink Architecture

A single Link in the P100 supports up to 40 GB/sec of bidirectional bandwidth between the endpoints. Multiple Links can be combined to form Gangs for even higher bandwidth connectivity between processors. The NVLink implementation in Tesla P100 supports up to four Links, enabling ganged configurations with aggregate maximum bidirectional bandwidth of 160 GB/sec.

GPU-TO-GPU CONNECTIVITY

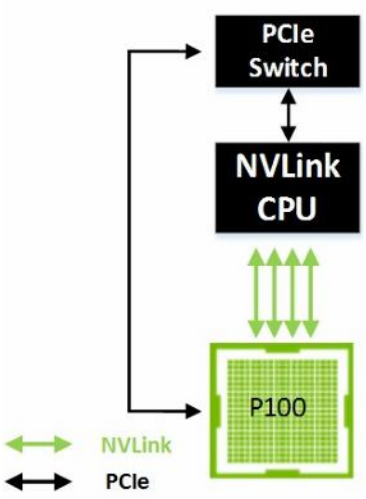
This figure shows two fully NVLink-connected quads of GPUs, with NVLink connections between the quads, and GPUs within each quad connected to their respective CPUs directly through PCIe. By using separate NVLink connections to span the gap between the two quads, it relieves pressure on the PCIe uplink to each CPU, and likewise avoids routing transfers through system memory and over an inter-CPU link.



Each half of the 8-GPU Hybrid Cube Mesh can operate as a shared memory multiprocessor, while the remote nodes can also share memory with DMA through peers. With all GPU-to-GPU traffic flowing over NVLINK, PCIe is now entirely available for either connection to a NIC or for accessing system memory traffic.

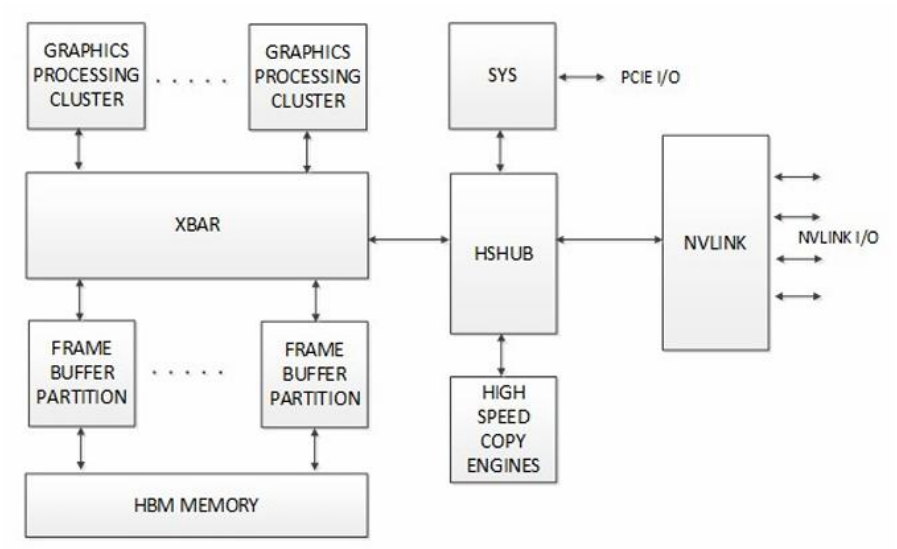
CPU-TO-GPU CONNECTIVITY

It is also possible to use as a CPU-to-GPU interconnect. This figure shows a single GPU connected to an NVLink-enabled CPU. In this case, the GPU can access system memory at up to 160 GB/sec bidirectional bandwidth —5x higher bandwidth than available over PCIe.



NVLINK RELATIONSHIP TO OTHER MAJOR BLOCKS IN GP100

At the level of the GPU architectural interface, the NVLink controller communicates with the GPU internals through another new block called the High-Speed Hub (HSHUB). The HSHUB has direct access to the GPU-wide crossbar and other system elements, such as the High-Speed Copy Engines (HSCE). which can be used to move data into and out of the GPU at peak NVLink rates.



NVLink Signaling and Protocol

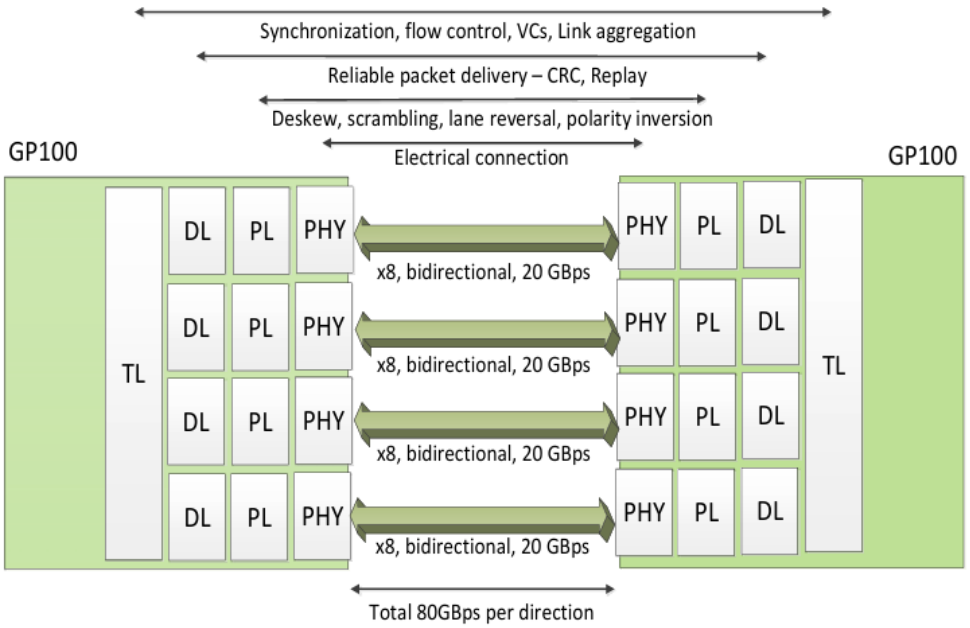
NVLink uses NVIDIA's High Speed Signaling technology (NVHS). Data is sent differentially at 20 Gbit/sec per signal pair. Eight differential pairs in each direction are combined to form a single link. This is the basic building block. A single link has a raw bidirectional bandwidth of 40 GB/sec. Signaling is NRZ (Non-Return to-Zero). The link is DC-coupled and has a differential impedance of 85 Ohms. Links can tolerate polarity inversion and lane reversal to support effective PCB routing. On die, data is sent from the PHY (physical level circuit) to the NVLink controller using a 128-bit Flit (Flow control digit) at 1.25GHz data rate. NVHS uses an embedded clock. At the receiver, the recovered clock is used to capture the incoming data.

NVLINK CONTROLLER LAYERS

The NVLink controller consists of three layers—the Physical Layer (PL), Data Link Layer (DL), and Transaction Layer (TL). The protocol uses a variable length packet with packet sizes ranging from 1 (simple read request command for example) to 18 (write request with data for 256B data transfer with address extension) flits.

PHYSICAL LAYER (PL)

The PL interfaces with the PHY. The PL is responsible for framing (figuring out the start of each packet), scrambling/descrambling (to ensure adequate bit transition density to support clock recovery), polarity inversion, lane reversal and for delivering the received data to the Data Link Layer. This figure shows the NVLink Layers and Links.



DATA LINK LAYER (DL)

The Data Link Layer is primarily responsible for reliable transmission of packets across the link. Packets to be transmitted are protected using a 25-bit CRC (Cyclic Redundancy Check). The transmitted packets are stored in a replay buffer until they have been positively acknowledged (ACK) by the receiver at the other end of the link. If the DL detects a CRC error on an incoming packet, it does not send an ACK, and prepares for reception of the retransmitted data. The transmitter meanwhile, in the absence of an ACK, times-out and initiates data retransmission from the replay buffer. A packet is retired from the replay buffer only when it has been acknowledged. The 25-bit CRC allows detection of up to 5 random bit errors or up to 25-bit bursts of errors on any lane. The CRC is calculated over the current header and the previous payload (if any). The DL is also responsible for link bring-up and maintenance. The DL sends data on to the Transaction Layer (TL).

TRANSACTION LAYER (TL)

The Transaction Layer handles synchronization, link flow control, virtual channels, and can aggregate multiple links together to provide very high communication bandwidth between processors.

NVLink Generations

	SECOND GENERATION	THIRD GENERATION	FOURTH GENERATION	FIFTH GENERATION
NVLINK BANDWIDTH PER GPU	300GB/s	600GB/s	900GB/s	1800GB/s
MAXIMUM NUMBER OF LINKS PER GPU	6	12	18	18
DIFFERENTIAL PAIRS PER LINK	8	4	2	2
SUPPORTED NVIDIA ARCHITECTURES	Volta	Ampere	Hopper	Blackwell