NVLink

AUTHOR: ARIAN AFZALZADEH

COURSE INSTRUCTOR: DR. FOSHATI

Topics

1. MOTIVATION
2. NVLINK FEATURES
3. NVLINK ARCHITECTURE
4. NVLINK SIGNALING AND PROTOCOL
5. NVLINK GENERATIONS

Motivation

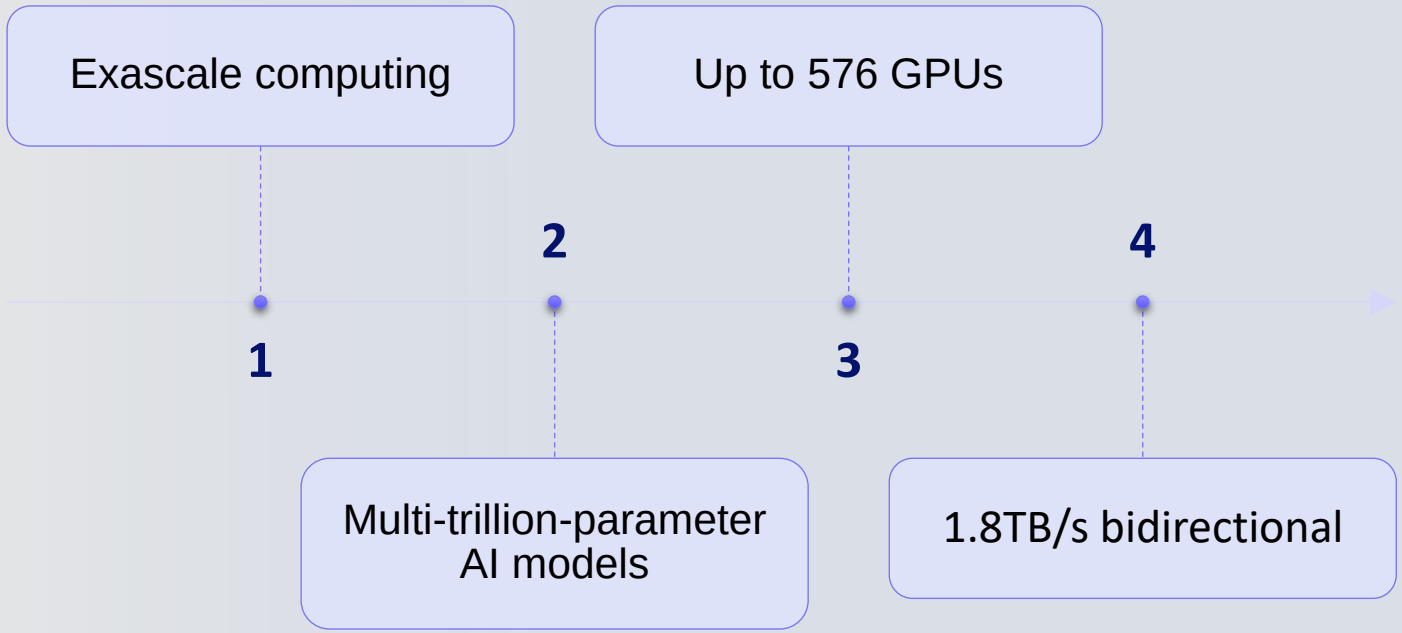
Why NVLink Over PCIe?

- The need of a powerful interconnect in multiprocessing systems
- Achieve higher scalability for deep learning workloads
- Create an interconnect for GPUs that would offer much higher bandwidth than PCIe Gen 3
- Compatibility with the GPU ISA to support shared memory multiprocessing workloads (full support for Pascal's atomic operations)

NVLink Achievements Today

Direct communication between every GPU within a server cluster

- High-bandwidth
- Energy-efficient
- Low-latency
- Lossless



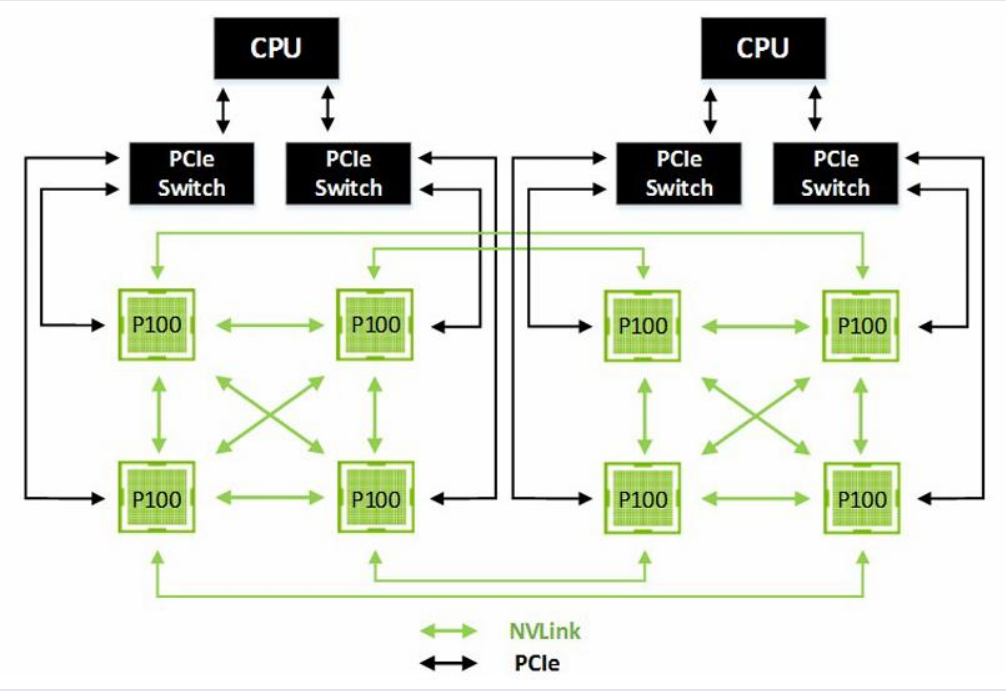
NVLink Architecture

First Generation of NVLink in The Tesla P100

- Bi-directional link at up to 40 Gb/s (GPU-to-GPU or GPU-to-CPU)
- Combination of four links for higher bandwith

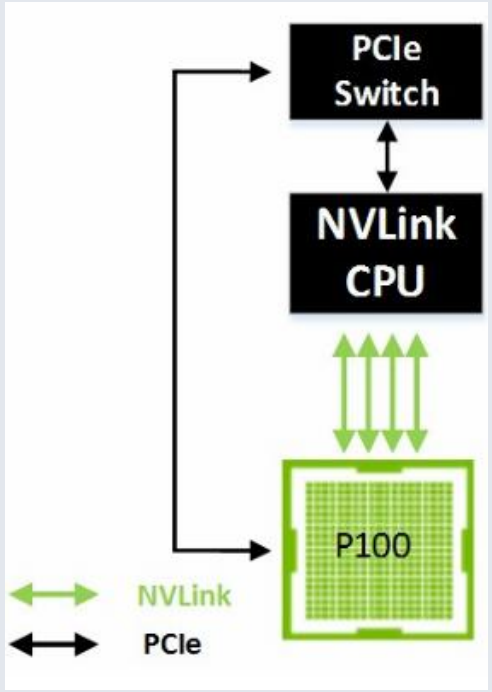
GPU-to-GPU Connectivity

- Two fully NVLink-connected quads of GPUs, with NVLink connections between the quads
- GPUs within each quad connected to their respective CPUs through PCIe
- Less pressure on the PCIe uplink to CPUs and avoiding routing transfers through system memory

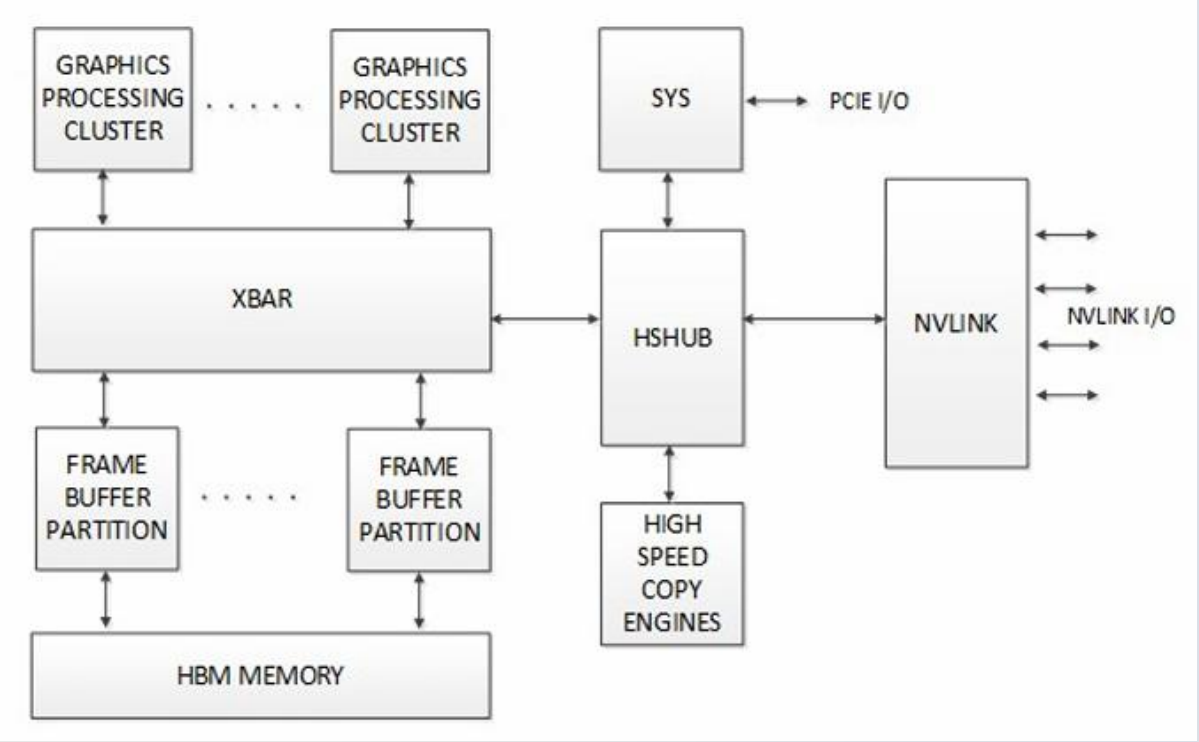


CPU-to-GPU Connectivity

Access system memory at up to 160 GB/s bidirectional—5x higher than PCIe gen 3



NVLink Relationship to Other Major Blocks in GP100



NVLink Signaling And Protocol

Signaling

- NVIDIA High-Speed Signaling interconnect (NVHS)
- Differential pair running at up to 20 Gb/s
- Combination of eight differential pairs in each direction to form a link
- Non-Return to-Zero
- Embedded clock
- Data sent from the PHY (physical level circuit) to the NVLink controller

Physical Layer And Transaction Layer

The PL handles:

- Framing
- Scrambling/descrambling
- Polarity inversion
- Lane reversal
- Delivering the received data to the Data Link Layer

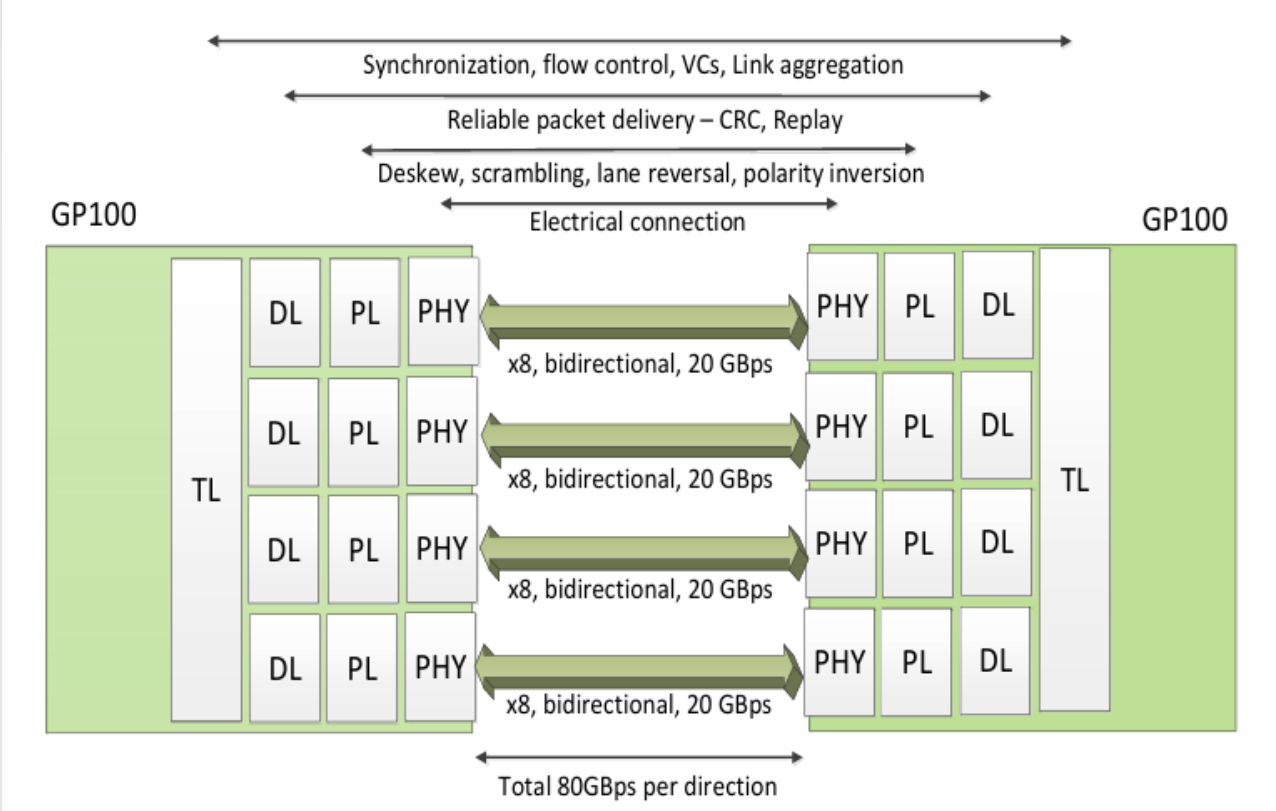
The TL handles:

- Synchronization
- Link flow control
- Aggregate multiple links together

Data Link Layer

- Reliable transmission of packets across the link
- Protection of packets using a 25-bit CRC against up to 5 random bit errors or up to 25-bit bursts of errors
- Storing the transmitted packets in a replay buffer until they have been acknowledged by the receiver at the other end of the link
- On CRC error on incoming packet, retransmission of data from the replay buffer and no ACK
- A packet is retired from the replay buffer only when it has been acknowledged.
- Sending data on to the TL

Overview of Layers And Links



NVLink Generations

Comparison of Different 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

Thank you

ARIAN AFZALZADEH

ARIAN.AFZALZADEH01@SHARIF.EDU