4th-Generation NVSwitch Chip

1. Brief History of NVLink
2. 4th-Generation New Features
3. Chip Details

Hopper-Generation SuperPODs

1. NVSwitch-Enabled Platforms
2. NVLink Network SuperPODs
3. SuperPOD Performance
**NVLINK MOTIVATIONS**

**Bandwidth and GPU-Synergistic Operation**

**GPU Operational Characteristics Match NVLink Spec**
- Thread-Block execution structure efficiently feeds parallelized NVLink architecture
- NVLink-Port Interfaces match data-exchange semantics of L2 as closely as possible

**Faster than PCIe**
- 100Gbps-per-lane (NVLink4) vs 32Gbps-per-lane (PCIe Gen5)
- Multiple NVLinks can be “ganged” to realize higher aggregate lane counts

**Lower Overheads than Traditional Networks**
- Target system scales (256 Hopper GPUs) allow complex features (e.g., end-to-end retry, adaptive routing, packet reordering) to be traded-off against increased port counts
- Simplified Application/Presentation/Session-layer functionality allows all to be embedded directly in CUDA programs/driver
NVLINK GENERATIONS
Evolution In-step with GPUs

2016
P100-NVLink1
- 4 NVLinks
- 40GB/s each
- x8@20Gbaud-NRZ
- 160GB/s total

2017
V100-NVLink2
- 6 NVLinks
- 50GB/s each
- x8@25Gbaud-NRZ
- 300GB/s total

2020
A100-NVLink3
- 12 NVLinks
- 50GB/s each
- x4@50Gbaud-NRZ
- 600GB/s total

2022
H100-NVLink4
- 18 NVLinks
- 50GB/s each
- x2@50Gbaud-PAM4
- 900GB/s total

Listed bandwidths are full-duplex (total of both directions). Whitepaper: http://www.nvidia.com/object/nvlink.html
NVLINK-ENABLED SERVER GENERATIONS
Any-to-Any Connectivity with NVSwitch

<table>
<thead>
<tr>
<th>Year</th>
<th>Model</th>
<th>Bisection BW</th>
<th>AllReduce BW</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>DGX-1 (P100)</td>
<td>140GB/s</td>
<td>40GB/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bisection BW</td>
<td>AllReduce BW</td>
</tr>
<tr>
<td>2018</td>
<td>DGX-2 (V100)</td>
<td>2.4TB/s</td>
<td>75GB/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bisection BW</td>
<td>AllReduce BW</td>
</tr>
<tr>
<td>2020</td>
<td>DGX A100</td>
<td>2.4TB/s</td>
<td>150GB/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bisection BW</td>
<td>AllReduce BW</td>
</tr>
<tr>
<td>2022</td>
<td>DGX H100</td>
<td>3.6TB/s</td>
<td>450GB/s</td>
</tr>
</tbody>
</table>
**NEW FEATURES**

**Expanding Server Performance**

**NVLink Network Support**
- PHY-electrical interfaces compatible with 400G Ethernet/InfiniBand
- OSFP support (4 NVLinks per cage) with custom FW for active modules
- Additional Forward Error Correction (FEC) modes for optical-cable performance/reliability

**Doubling of Bandwidth**
- 100Gbps-per-diff-pair (50Gbaud PAM4)
- x2 NVLinks and 64 NVLinks-per-NVSwitch (1.6TB/s internal bisection BW)
- More BW with fewer chips

**SHARP Collectives/Multicast Support**
- NVSwitch-internal duplication of data avoid need for multiple access from/by source GPU
- Embedded ALUs allow NVSwitches to perform AllReduce (and similar) calculations on behalf of GPUs
- Roughly doubles data throughput on communication-intensive-operations in AI-applications

**2022**
DGX H100
- 3.6TB/s Bisection BW
- 450GB/s AllReduce BW
**NVLink4 NVSwitch**

**Chip Characteristics**

- **32 PHY Lanes**
- **PORT Logic (including SHARP accelerators)**
- **XBAR**
- **PORT Logic (including SHARP accelerators)**
- **32 PHY Lanes**

**Largest NVSwitch Ever**
- TSMC 4N process
- 25.1B transistors
- 294mm²
- 50mmX50mm package (2645 balls)

**Highest Bandwidth Ever**
- 64 NVLink4 ports (x2 per NVLink)
- 3.2TB/s full-duplex bandwidth
- 50Gbaud PAM4 diff-pair signaling
- All ports NVLink Network capable

**New Capabilities**
- 400GFLOPS of FP32 SHARP (other number formats are supported)
- NVLink Network management, security and telemetry engines
ALLREDUCE IN AI TRAINING
Critical Communication-Intensive Operation

BASIC TRAINING FLOW

ALLREDUCE IN MULTI-GPU TRAINING
TRADITIONAL ALLREDUCE CALCULATION
Data-Exchange and Parallel Calculation

ALLREDUCE IN MULTI-GPU TRAINING
## NVLINK SHARP ACCELERATION

### Step 1: Read and reduce

<table>
<thead>
<tr>
<th>A100</th>
<th>H100 + NVLink SHARP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Send Partials</strong></td>
<td><strong>Send Partials</strong></td>
</tr>
<tr>
<td>N reads</td>
<td>N reads</td>
</tr>
<tr>
<td><strong>Receive Partials</strong></td>
<td><strong>In-Switch Sum</strong></td>
</tr>
<tr>
<td>N reads</td>
<td>--</td>
</tr>
</tbody>
</table>

**Receive Partials**
1 reduced read

### Step 2: Broadcast result

<table>
<thead>
<tr>
<th>A100</th>
<th>H100</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Send New Partial</strong></td>
<td><strong>Send New Partial</strong></td>
</tr>
<tr>
<td>N writes</td>
<td>1 write</td>
</tr>
<tr>
<td><strong>Receive New Partials</strong></td>
<td><strong>In-Switch MultiCast</strong></td>
</tr>
<tr>
<td>N writes</td>
<td>N duplications</td>
</tr>
</tbody>
</table>

**Receive New Partials**
N writes

### Traffic summary (at each GPU interface)

- 2N send, 2N receive
- N+1 send, N+1 receive

~2x effective NVLink bandwidth
NVLink Network for Raw BW
4.5X More BW than Maximum InfiniBand (IB)

Neural Recommender Engine
- Linear layers
- Data parallel
- Replicated across GPUs

Example Recommender with 14 TB Embedding Tables
- Redistribute: Model-Parallel -> Data-Parallel
- Embedding tables: Model parallel
- Distributed across GPUs

Projected performance subject to change. Example model assumes DLRM with a mix of 300-hot and 1-hot embedding tables with total capacity of 14TB. Different recommender models may show different performance characteristics.
New Hopper NIC functions ensure request is legal & maps to GPU physical address space.
# MAPPING TO TRADITIONAL NETWORKING

NVLink Network is Tightly Integrated with GPU

<table>
<thead>
<tr>
<th>Concept</th>
<th>Traditional Example</th>
<th>NVLink Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Layer</td>
<td>400G electrical/optical media</td>
<td>Custom-FW OSFP</td>
</tr>
<tr>
<td>Data Link Layer</td>
<td>Ethernet</td>
<td>NVLink custom on-chip HW and FW</td>
</tr>
<tr>
<td>Network Layer</td>
<td>IP</td>
<td>New NVLink Network Addressing and Management Protocols</td>
</tr>
<tr>
<td>Transport Layer</td>
<td>TCP</td>
<td>NVLink custom on-chip HW and FW</td>
</tr>
<tr>
<td>Session Layer</td>
<td>Sockets</td>
<td>SHARP groups</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CUDA export of Network addresses of data-structures</td>
</tr>
<tr>
<td>Presentation Layer</td>
<td>TSL/SSL</td>
<td>Library abstractions (e.g., NCCL, NVSHMEM)</td>
</tr>
<tr>
<td>Application Layer</td>
<td>HTTP/FTP</td>
<td>AI Frameworks or User Apps</td>
</tr>
<tr>
<td>NIC</td>
<td>PCIe NIC (card or chip)</td>
<td>Functions embedded in GPU and NVSwitch</td>
</tr>
<tr>
<td>RDMA Off-Load</td>
<td>NIC Off-Load Engine</td>
<td>GPU-internal Copy Engine</td>
</tr>
<tr>
<td>Collectives Off-Load</td>
<td>NIC/Switch Off-Load Engine</td>
<td>NVSwitch-internal SHARP Engines</td>
</tr>
<tr>
<td>Security Off-Load</td>
<td>NIC Security Features</td>
<td>GPU-internal Encryption and “TLB” Firewalls</td>
</tr>
<tr>
<td>Media Control</td>
<td>NIC Cable Adaptation</td>
<td>NVSwitch-internal OSFP-cable controllers</td>
</tr>
</tbody>
</table>
New SHARP Blocks
- ALU matched to Hopper unit
- Wide variety of operators (logical, min/max, add) and formats (S/U integers, FP16, FP32, FP64, BF16)
- SHARP Controller can manage up to 128 SHARP groups in parallel
- XBAR BW uprated to carry additional SHARP-related exchanges

New NVLink Network Blocks
- Security Processor protects data and chip configuration from attacks
- Partitioning features isolate subsets of ports into separate NVLink Networks
- Management controller now also handles attached OSFP cables
- Expanded telemetry to support InfiniBand-style monitoring
NVLink4-Generation NVSwitch Chip

1. Brief History of NVLink
2. NVLink4-Generation New Features
3. Chip Details

Hopper-Generation SuperPODs

1. NVSwitch-Enabled Platforms
2. NVLink Network SuperPODs
3. SuperPOD Performance
DGX H100 SERVER

8-H100 4-NVSwitch Server

- 32 PFLOPS of AI Performance
- 640 GB aggregate GPU memory
- 18 NVLink Network OSFPs
- 3.6 TBps of full-duplex NVLink Network bandwidth (72 NVLinks)
- 8x 400 Gb/s ConnectX-7 InfiniBand/Ethernet ports
- 2 dual-port Bluefield-3 DPUs
- Dual Sapphire Rapids CPUs
- PCIe Gen5
DGX H100: DATA-NETWORK CONFIGURATION

Full-BW Intra-Server NVLink
- All 8 GPUs can simultaneously saturate 18 NVLinks to other GPUs within server
- Limited only by over-subscription from multiple other GPUs

Half-BW NVLink Network
- All 8 GPUs can half-subscribe 18 NVLinks to GPUs in other servers
- 4 GPUs can saturate 18 NVLinks to GPUs in other servers
- Equivalent of full-BW on AllReduce with SHARP
- Reduction in All2All BW is a balance with server complexity and costs

Multi-Rail InfiniBand/Ethernet
- All 8 GPUs can independently RDMA data over its own dedicated 400 Gb/s HCA/NIC
- 800 GBps of aggregate full-duplex to non-NVLink Network devices
DGX H100 SUPERPOD: NVLINK SWITCH

NVLink Switch

- Standard 1RU 19-inch formfactor highly leveraged from InfiniBand switch design
- Dual NVLink4 NVSwitch chips
- 128 NVLink4 ports
- 32 OSFP cages
- 6.4 TB/s full-duplex BW
- Managed switch with out-of-band management communication
- Support for passive-copper, active-copper and optical OSFP cables (custom FW)
DGX H100 SUPERPOD: AI EXASCALE

DGX H100 SuperPOD Scalable Unit

- 32 DGX H100 nodes + 18 NVLink Switches
- 256 H100 Tensor Core GPUs
- 1 ExaFLOP of AI performance
- 20 TB of aggregate GPU memory
- Network optimized for AI and HPC
- 128 L1 NVLink4 NVSwitch chips + 36 L2 NVLink4 NVSwitch chips
- 57.6 TB/s bisection NVLink Network spanning entire Scalable Unit
- 25.6 TB/s full-duplex NDR 400 Gb/s InfiniBand for connecting multiple Scalable Units in a SuperPOD
SCALE-UP WITH NVLINK NETWORK

### DGX A100 256 POD
- IB HDR spine switches
- ... IB HDR leaf switches ...
- ... 32 nodes (256 GPUs) ...

### DGX H100 256 POD
- Fully NVLink-connected
- Massive bisection bandwidth
- ... 32 nodes (256 GPUs) ...

<table>
<thead>
<tr>
<th></th>
<th>A100 SuperPOD</th>
<th>H100 SuperPOD</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 DGX / 8 GPUs</td>
<td>2.5</td>
<td>2,400</td>
<td>150</td>
</tr>
<tr>
<td>32 DGXs / 256 GPUs</td>
<td>80</td>
<td>6,400</td>
<td>100</td>
</tr>
</tbody>
</table>
NVLINK NETWORK BENEFITS
Dependent on Communication Intensity

Projected performance subject to change. A100 cluster: HDR IB network. H100 cluster: NDR IB network with NVLink Network where indicated.

# GPUs: Climate Modelling 1K, LQCD 1K, Genomics 8, 3D-FFT 256, MT-NLG 32 (batch sizes: 4 for A100, 60 for H100 at 1sec, 8 for A100 and 64 for H100 at 1.5 and 2sec), MRCNN 8 (batch 32), GPT-3 16B 512 (batch 256), DLRM 128 (batch 64K), GPT-3 175B 16K (batch 512), MoE 8K (batch 512, one expert per GPU)
SUMMARY
Cutting-Edge Speeds and Capabilities

NVLink4-Generation NVSwitch
- 64 NVLink4 ports and 3.2 TB/s full-duplex BW
- NVLink SHARP (multi-cast and reductions off-load)
- Inter-Server NVLink Network support
- Custom FW OSFP NVLink Network cable support
- Basis of new NVLink Switch

Hopper-Generation SuperPOD
- 32 DGX H100 servers
- 18 NVLink Switches
- 1 ExaFLOP of AI performance
- 57.6TB/s NVLink Network bisection BW
- NVLink Network can more than double performance for communication-intensive applications
- Scalable to thousands of GPUs using InfiniBand to connect multiple Scalable Units