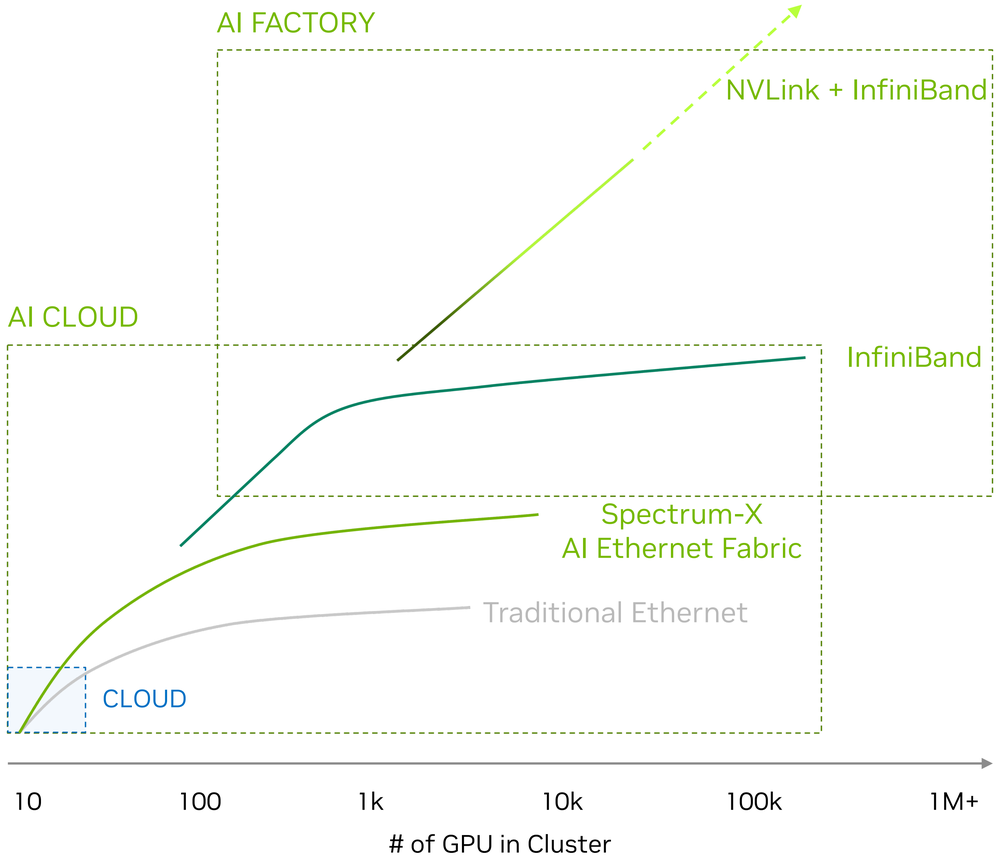
Networking Update

Sungta Tsai, Sr. Solution Architect | April 2024

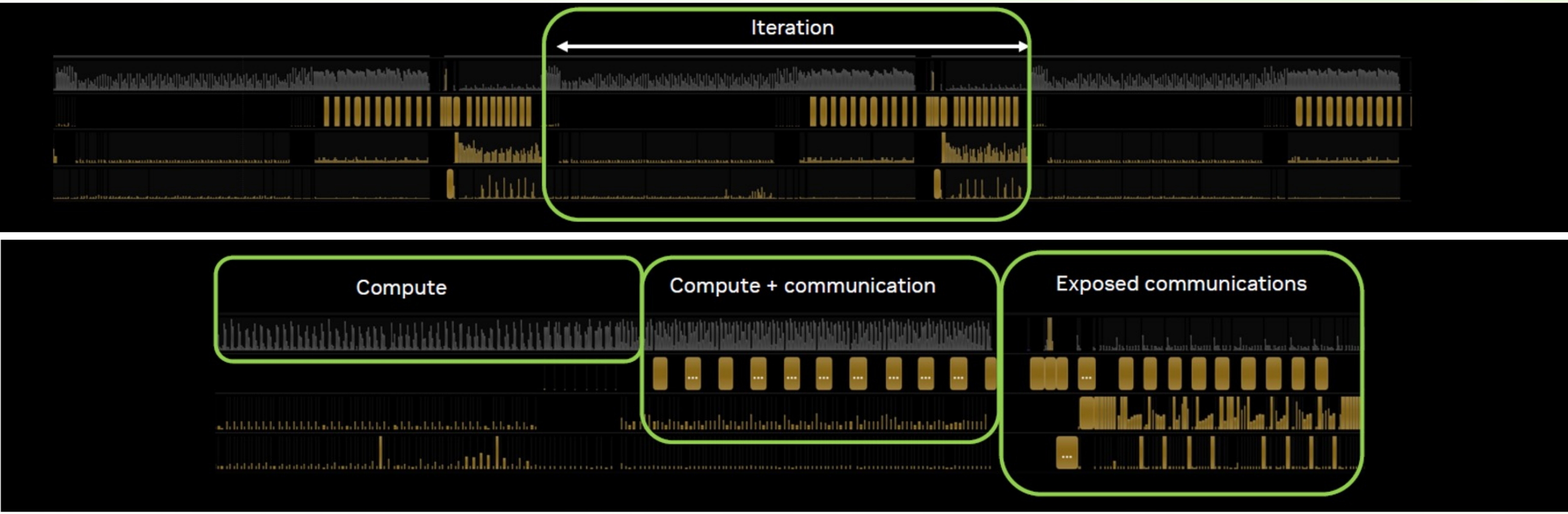
The Data Center is The Computer

The network defines the data center



- Cloud
  - Multi-tenant
  - Variety of small-scale workloads
  - Traditional ethernet network can suffice
- Generative AI Cloud
  - Multi-tenant
  - Variety of workloads including larger scale Generative AI
  - Traditional ethernet network for North-South traffic
  - NVIDIA Spectrum-X ethernet for AI fabric (East-West)
- AI Factories
  - Single or few users
  - Extremely large AI models
  - NVLink and InfiniBand gold standard for AI fabric

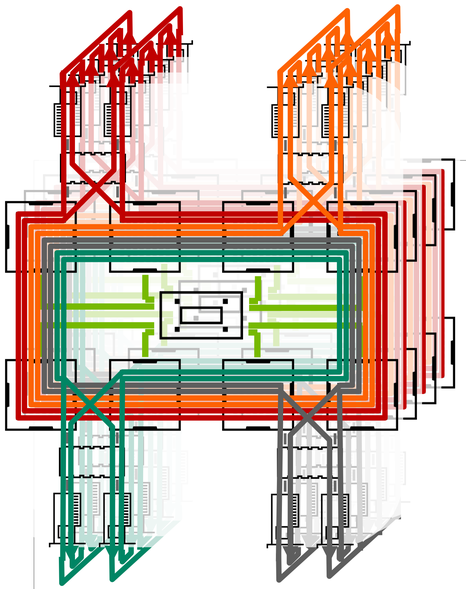
LLM Compute and Communication Profiling



Representative profile form a large scale LLM training run
Communications is bursty in nature, an average bandwidth utilization is not a good network criteria

Hardware & Software Accelerated In-Network Computing

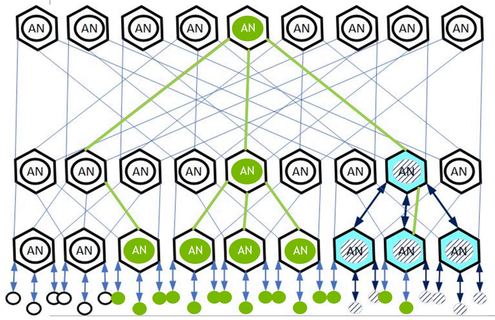
AI Network Considerations



Software Acceleration

NCCL — NVIDIA Collective Communication Library

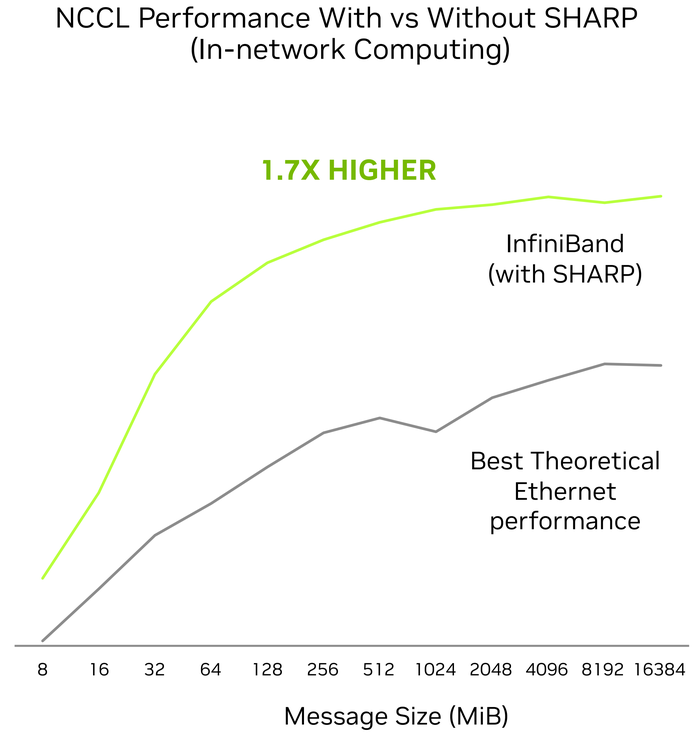
The SDK library for AI communications - connects the GPUs and the network for the AI network operations.



Hardware Acceleration

SHARP —Scalable Hierarchical Aggregation and Reduction Protocol Technology

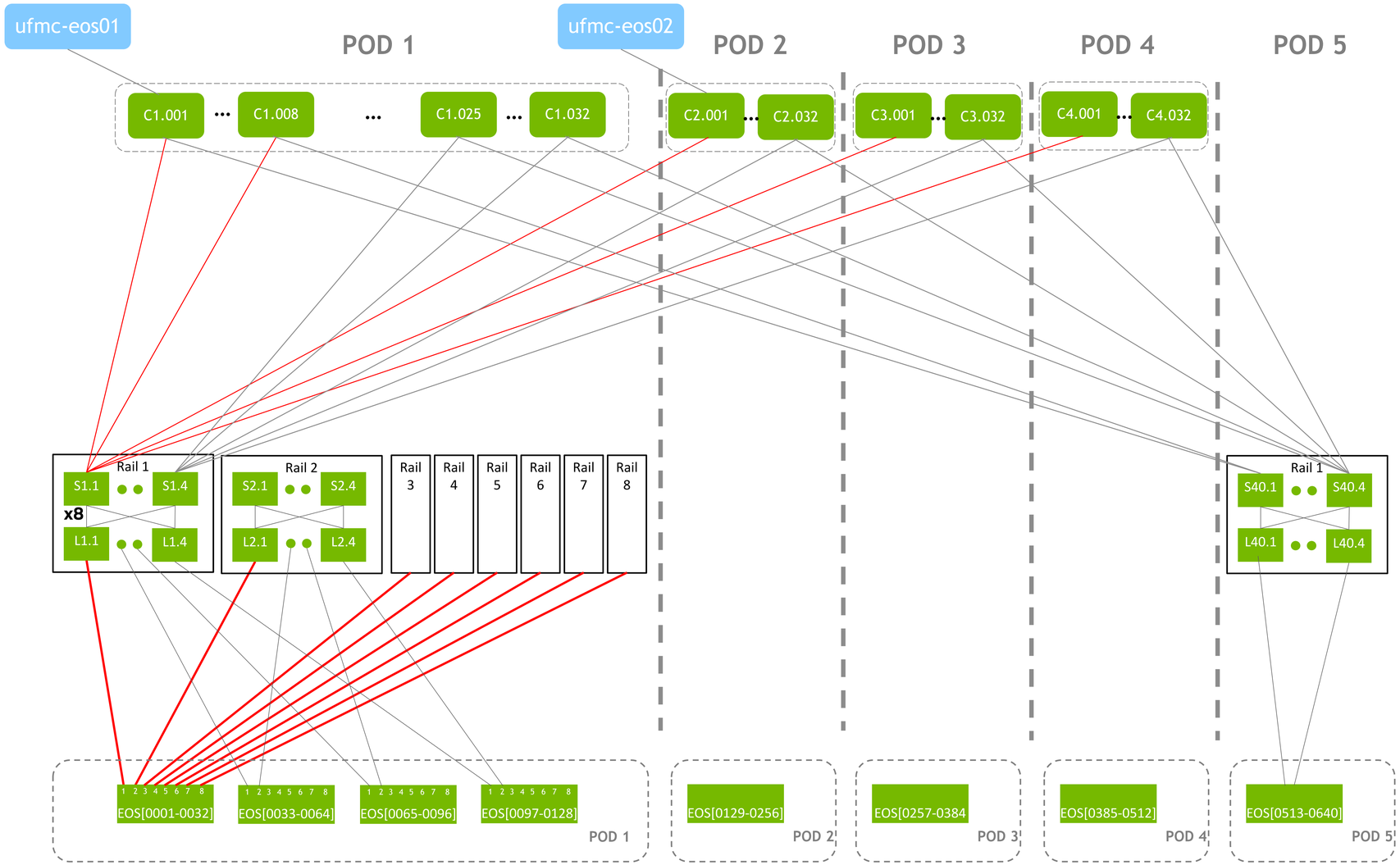
SHARP is part of the InfiniBand and NVLink switch ASICs. It enables the network to perform data reduction operations, an important element of AI workloads. This decreases the amount of data traversing the network and dramatically reduces collective operations time.



Compute InfiniBand Architecture

Fully plan production scale-out deployment

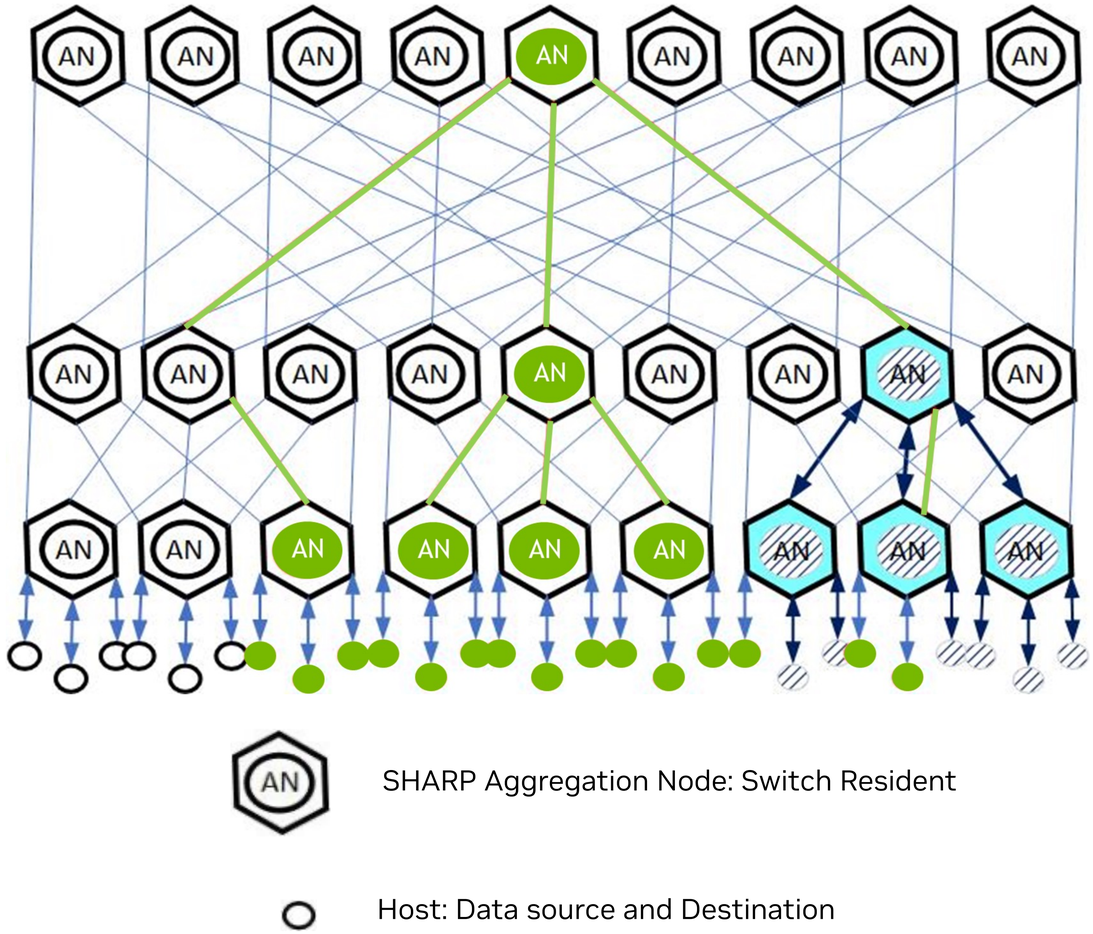
- Full Fat Tree for performance
- “Rail-optimized” to minimize latency and avoid congestion. “Rail” refers to the network link associated with a particular GPU index
  - Groups of 32 nodes have each rail connected to a single switch (“Leaf Rail-Optimized”)
  - Rail Groups are made of 4 leaf switches and 4 spine switches. There are 8 Rail Groups per POD, one per rail
  - The core switches are installed in conjunction with the first 4 POD only, 32 switches each. Empty ports are left on the core switches to support 3 additional PODs without recabling the existing ones



NVIDIA SHARP

Scalable Hierarchical Aggregation and Reduction Protocl Technology

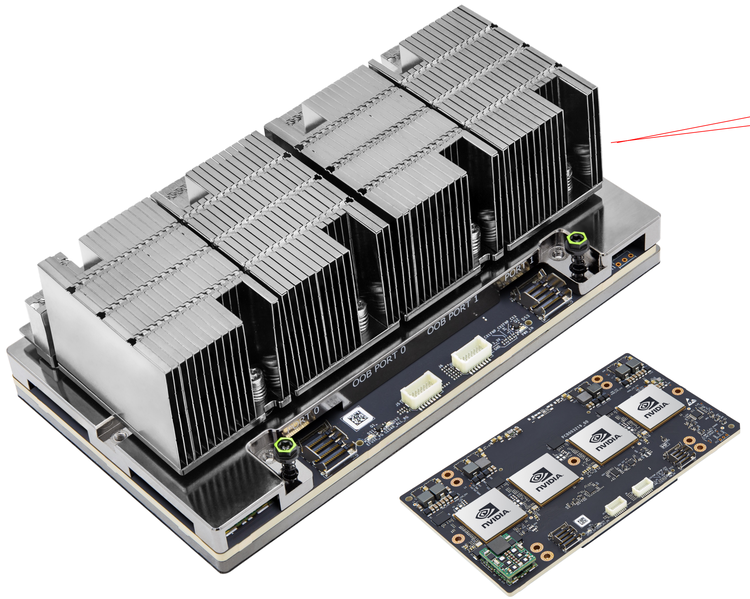
- In-Network data aggregation mechanism
- Multiple simultaneous outstanding operations
- Barrier, reduce, all-reduce, broadcast and more
- Sum, min, min-loc, max-loc, or, xor, and
- Integer and floating-point, 8/16/32/64 bits



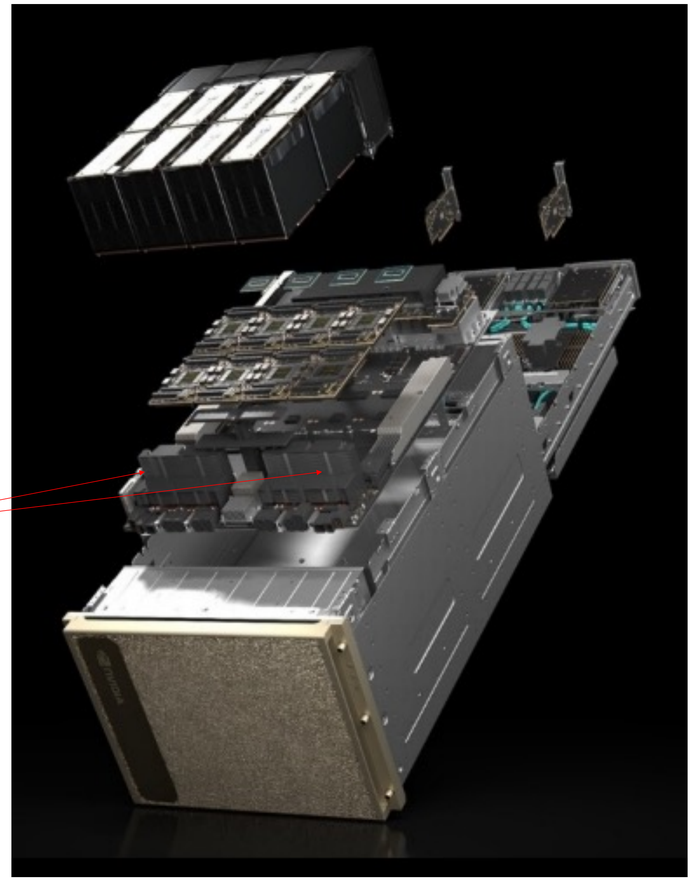
InfiniBand NDR in DGX H100 / B200

Inside DGX H100

- 8x ConnectX-7 (within 2x Mezz boards) for compute
  - Mezz board with 4 ConnectX-7 VPI (internal name: CEDAR)
  - 4 OSFP connectors, delivering 8 400Gb/s ports
- 2x ConnectX-7 VPI for storage & management
  - 2 QSFP112 each



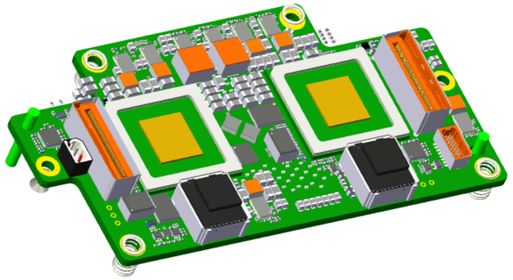
NVIDIA DGX H100



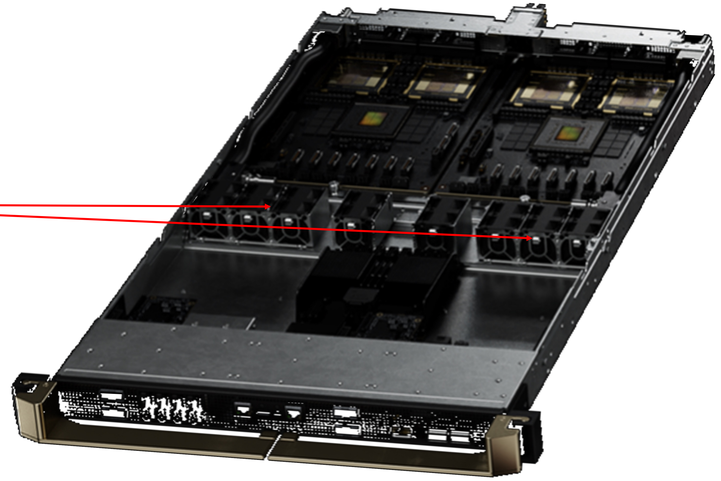
NVIDIA DGX B200

InfiniBand XDR in GB200 NVL72

- GB200 NVL72 is a rack scale solution for GPUs connected via NVLink
- GB200 Architecture supports 72 GPUs NVL domains



x2 C8280Z mezzanine network board.
Each one contain 2x ConnectX-8 ASICs



GB200 SUPERCHIP COMPUTE TRAY

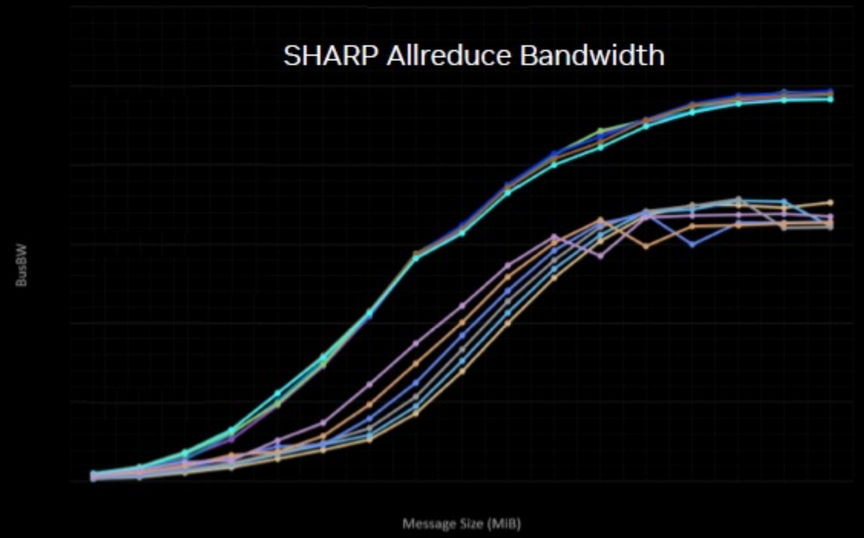
2x GB200
80 PETAFLOPS FP4 AI INFERENCE
40 PETAFLOPS FP8 AI TRAINING
1728 GB FAST MEMORY
1U Liquid Cooled
18 Per Rack

GB200 NVL72
Compute for Trillion Parameter Scale AI
Maximum Performance and Lowest TCO

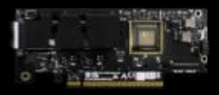
Quantum-X800 InfiniBand Switch

Highest-Performance AI-Dedicated Infrastructure

- 144 ports of 800G, 5x higher switch capacity
- SHARP v4 with 14.4 TFlops of In-Network Computing, 9x higher
- Adaptive routing, congestion control



Quantum-X800 switch
- 144X 800G ports,
- SHARPv4 In-Network Computing
- Adaptive routing
- congestion control, and noise isolation



ConnectX-x800 InfiniBand SuperNIC
- PCIe Gen 6, PCIe switch
- Multi-host

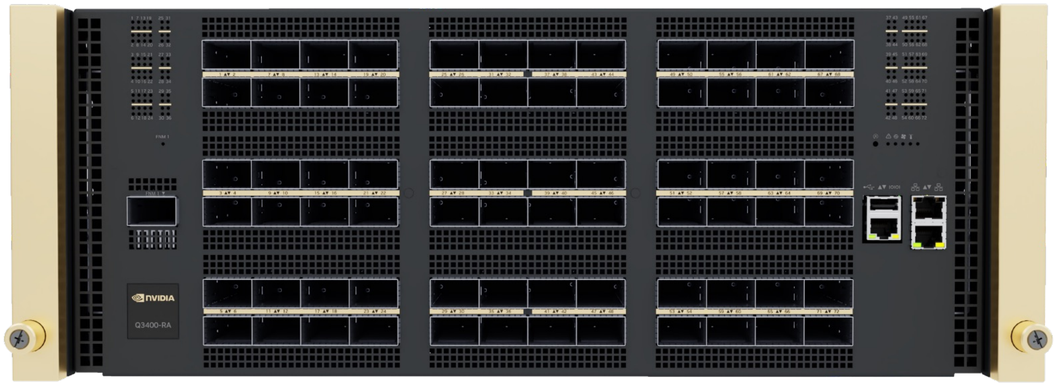
Quantum-X800

ConnectX-800 SuperNIC

Quantum-X800 Q3400 Systems

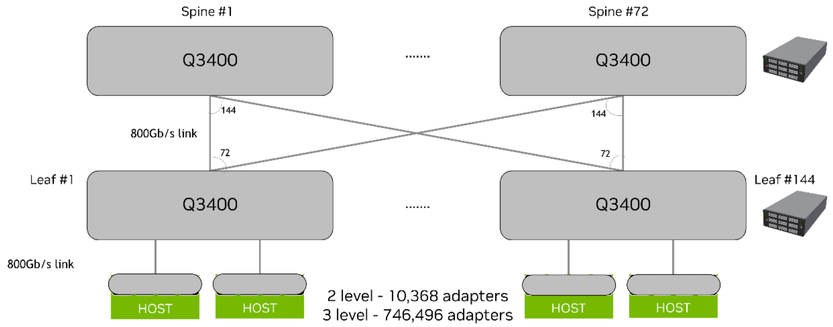
Quantum-X800 Q3400 4U switch platform

- 144 x 800Gb/s ports, over 72 OSFP cages
- Port for management (UFM)
- 10,368 nodes with two level fat tree non-blocking topology
- Managed switches
- Air-cooled (19” rack) and liquid-cooled (OCP rack) systems



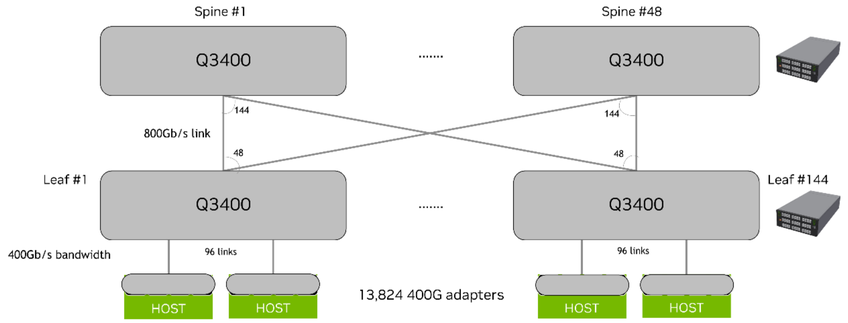
- Best performance for HPC & AI workloads
  - SHARP
  - Adaptive Routing
  - Telemetry based congestion control and performance isolation
  - Advanced power features

Quantum-X800 Topology Design



Quantum-X800 Topology Design

400G to the host by 2:1 port ratio at Leaf level



Quantum-X800 Q3200 2U Systems

- Best for
  - Small scale platforms
  - Connectivity to previous IB generations (NDR, HDR)

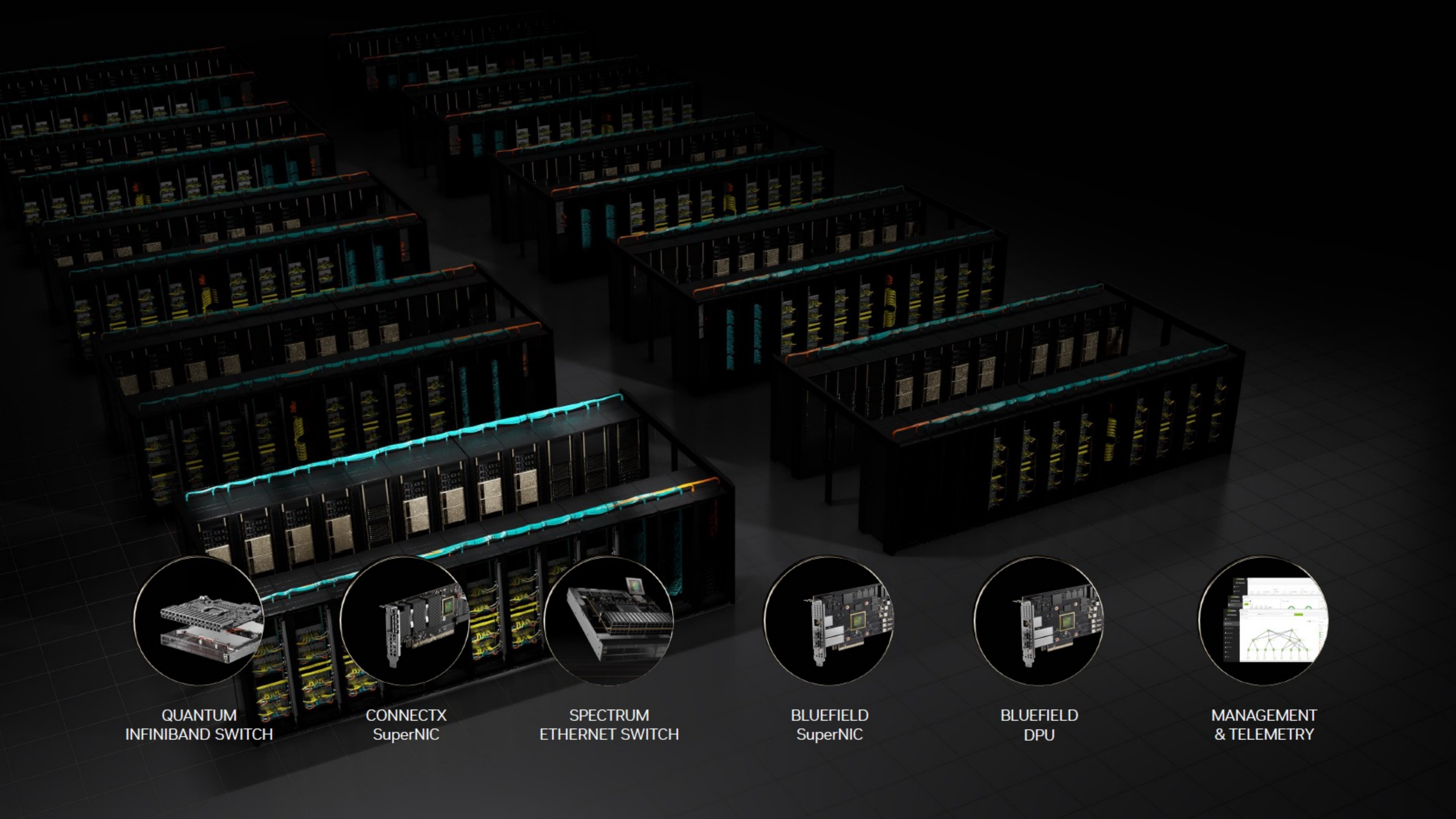
- Two independent 36-port switches in a single 2U enclosure
  - 18 OSFP cages per switch
  - 36 ports of 800G per switch
  - Managed switch
  - Air cooled

- InfiniBand port for out-of-band IB management

Switch-1 Switch-2

Switch Radix Capability		
Speed	**Max Ports per Switch**	**Max Ports per Q3200-RA**
XDR	36	72
NDR	36	72
NDR200	72	144
HDR	36	72

Images for illustration only


QUANTUM
INFINIBAND SWITCH
CONNECTX
SuperNIC
SPECTRUM
ETHERNET SWITCH
BLUEFIELD
SuperNIC
BLUEFIELD
DPU
MANAGEMENT
& TELEMETRY