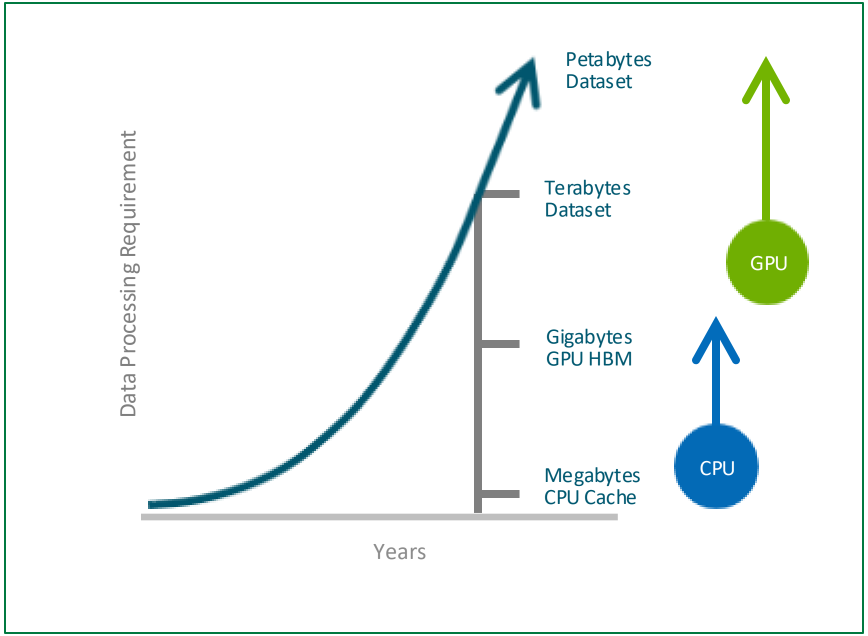
Mature at Scale Memory Fabrics for all Performance and Price Points

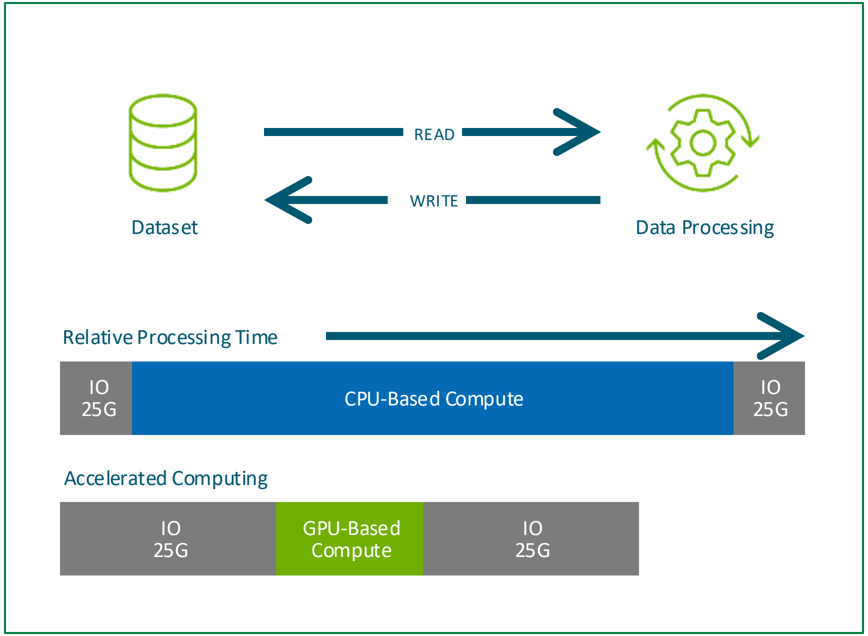
Rob Davis, VP Storage Technology, NVIDIA

MEMORY FABRIC FORUM | OCP GLOBAL SUMMIT | OCT 15-17, 2024 SAN JOSE, CA

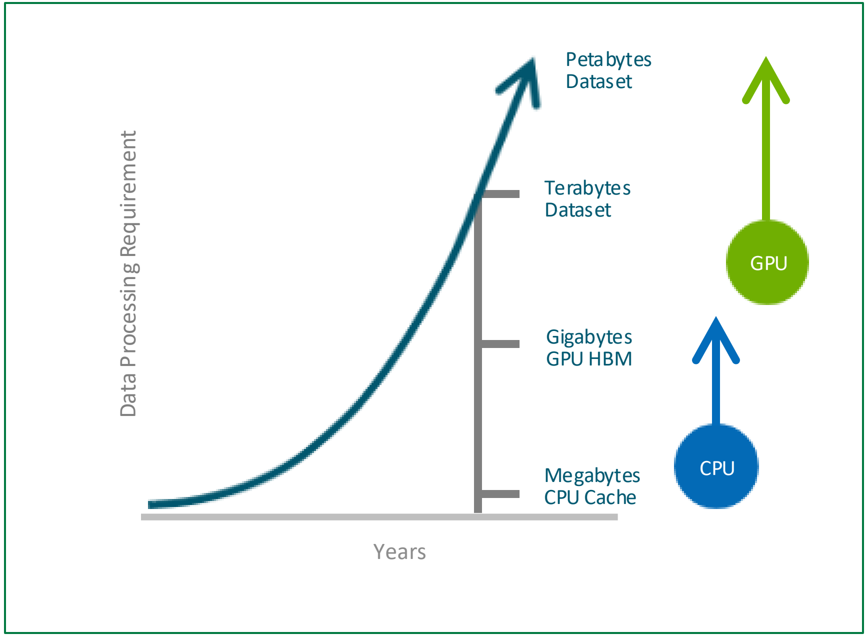
Networking Challenges for AI



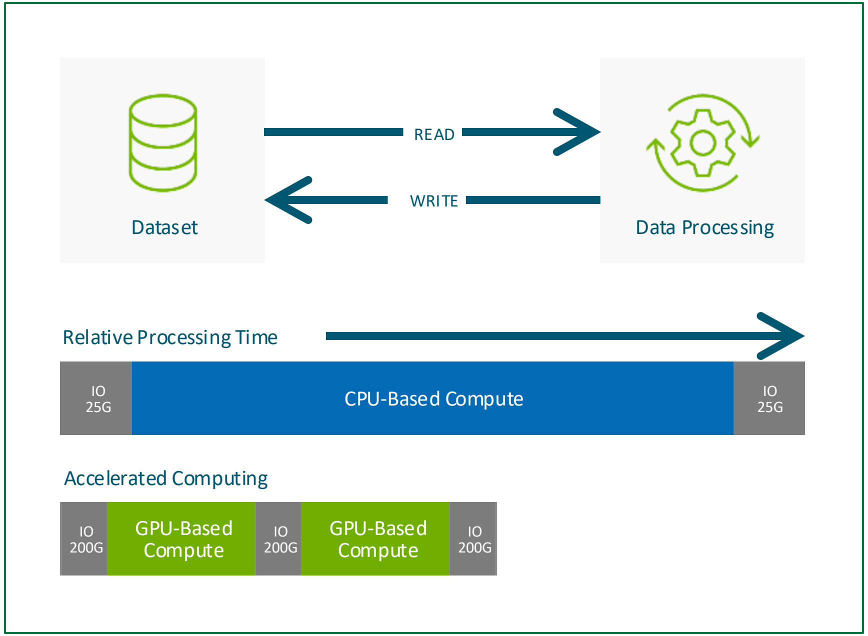
GPU Application Data Sets



Networking Challenges for AI



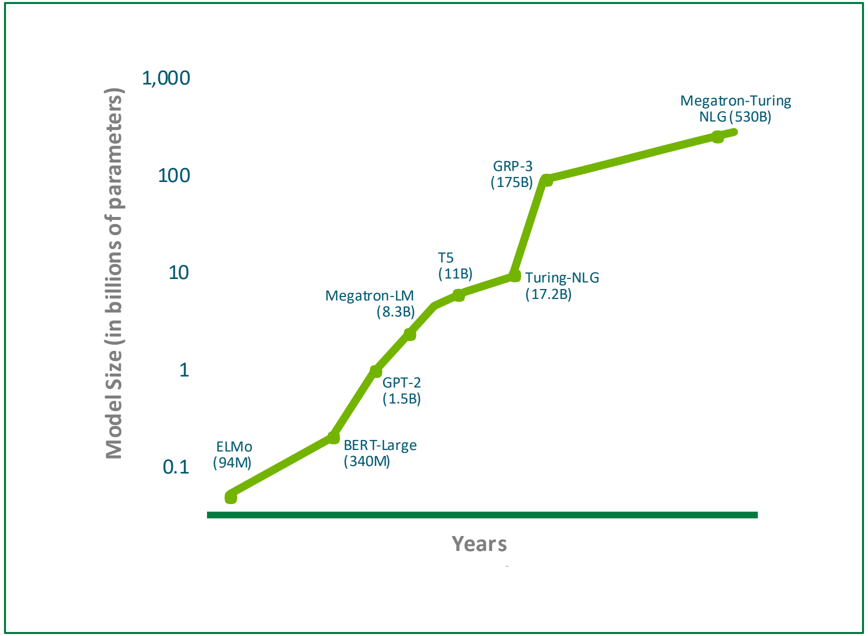
GPU Application Data Sets



GPU Direct Storage (RDMA)

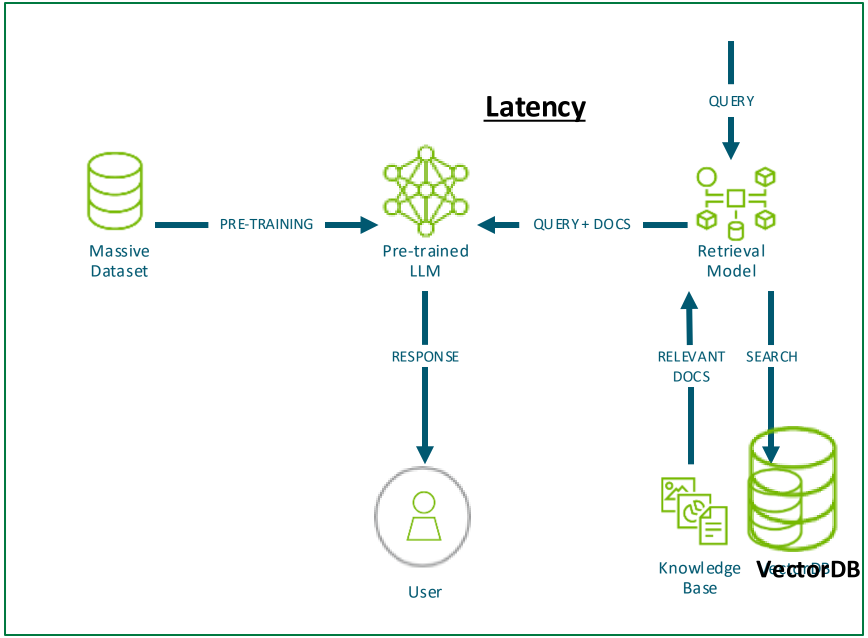
Next Wave of Performance and Scalability

Training



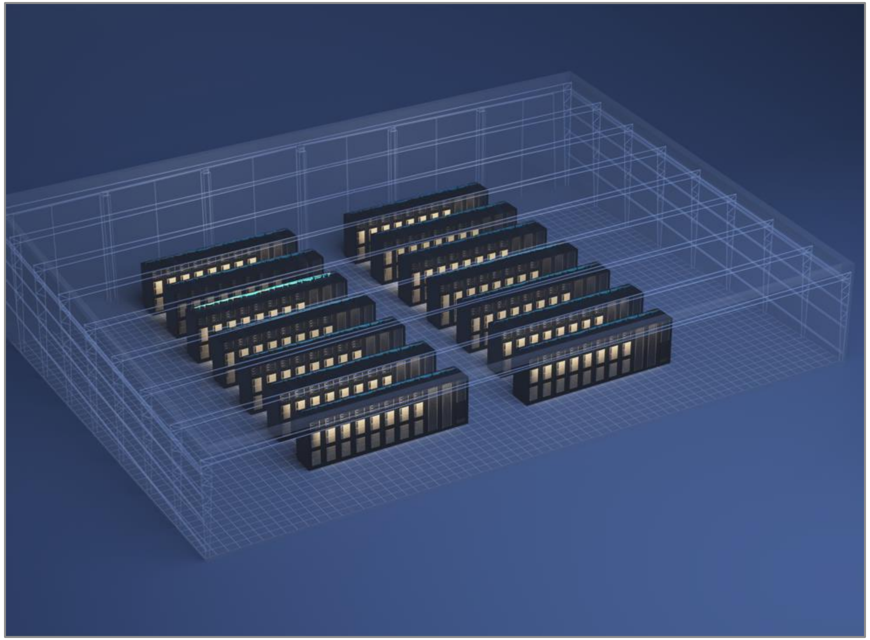
Exploding Model Size

Inferencing



Retrieval Augmented Generation (RAG)

Two Types of AI Data Centers



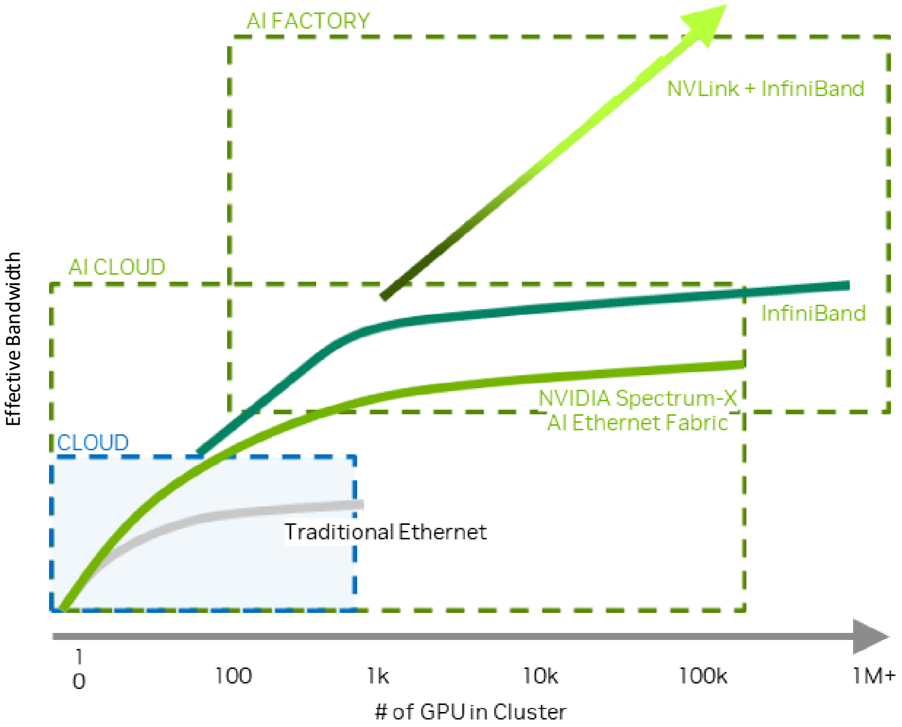
AI Factories

Single or few users | Extremely large AI models | NVLink and InfiniBand AI fabric

AI Cloud

Multi-tenant | Variety of workloads | Ethernet network

The Right Network for the Right AI Workload and Scale



NVLink

Fastest Compute Fabric Connectivity

InfiniBand

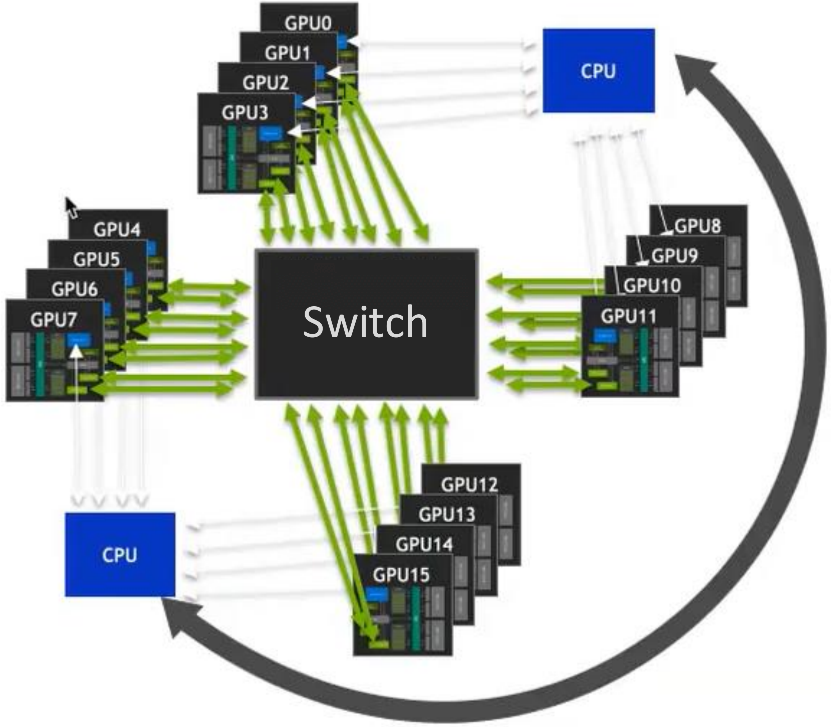
For Fast Scale-Out AI Network

Spectrum-X

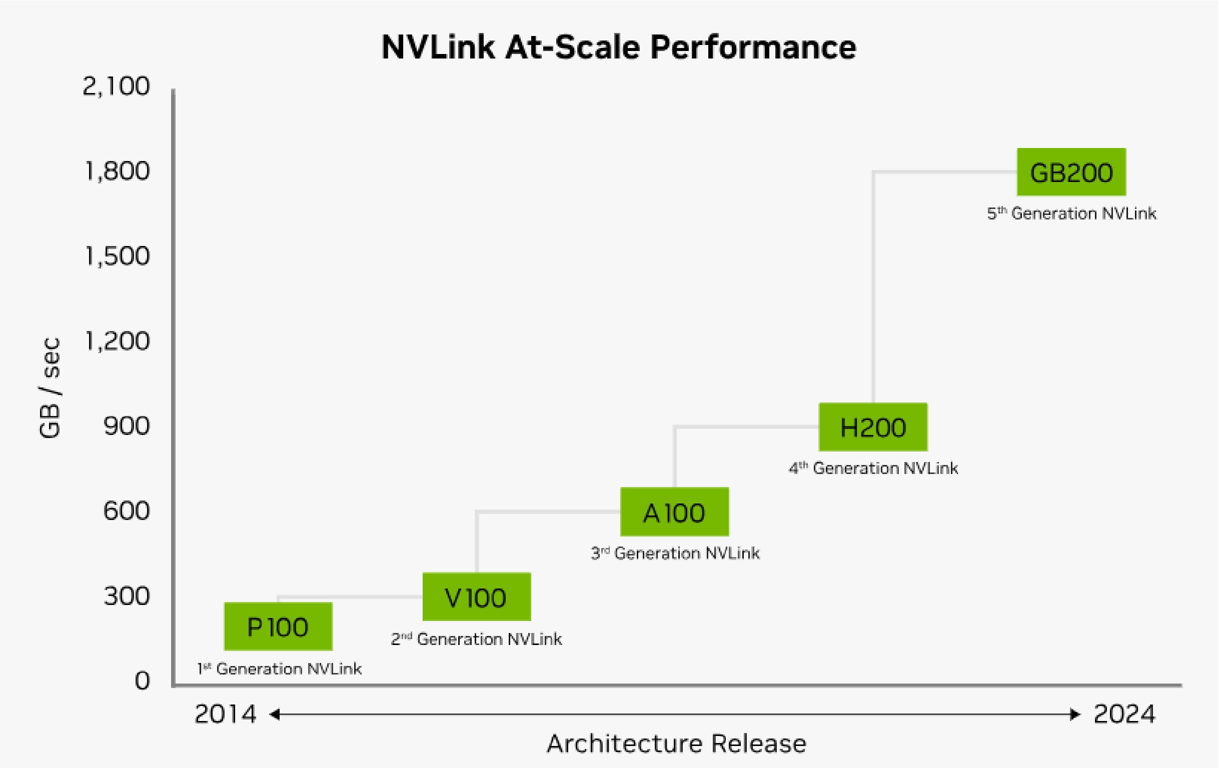
Ethernet Optimized for AI

NVLINK Creates One Gigantic GPU Memory

- From the perspective of the GPUs, all HBMs can be accessed without intervention by other processes
- Load/Store, DMA



NVLink is Tried and True and VERY Performant



Fifth Generation NVLink Switch

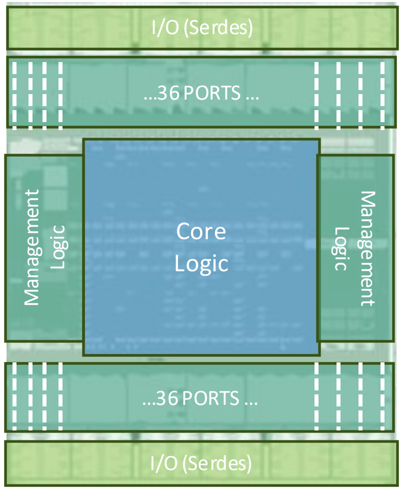
Single ASIC

72 NVLink Ports

100GB/s port speed

7.2TB/s Total Bandwidth

3.6 TF In-Network Compute



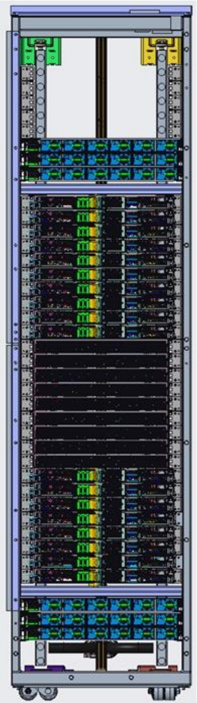
1.8 TB/s Bidirectional Bandwidth

Rack-Scale to Data Center Scale

Up to 576 GPU NVLink Domain

Unified Fabric Management

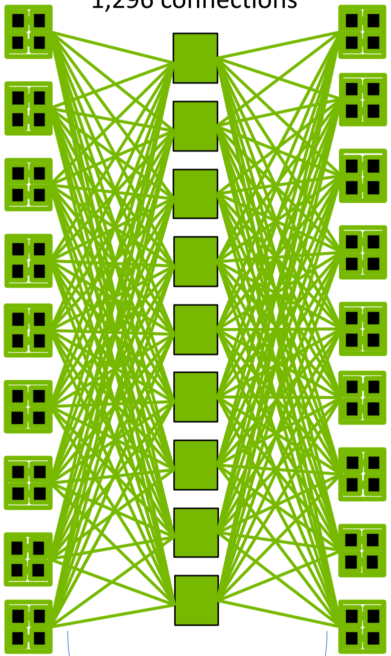
NVLink Deployed at Rack Scale



10 compute trays

9 NVLink Switch trays

8 compute trays



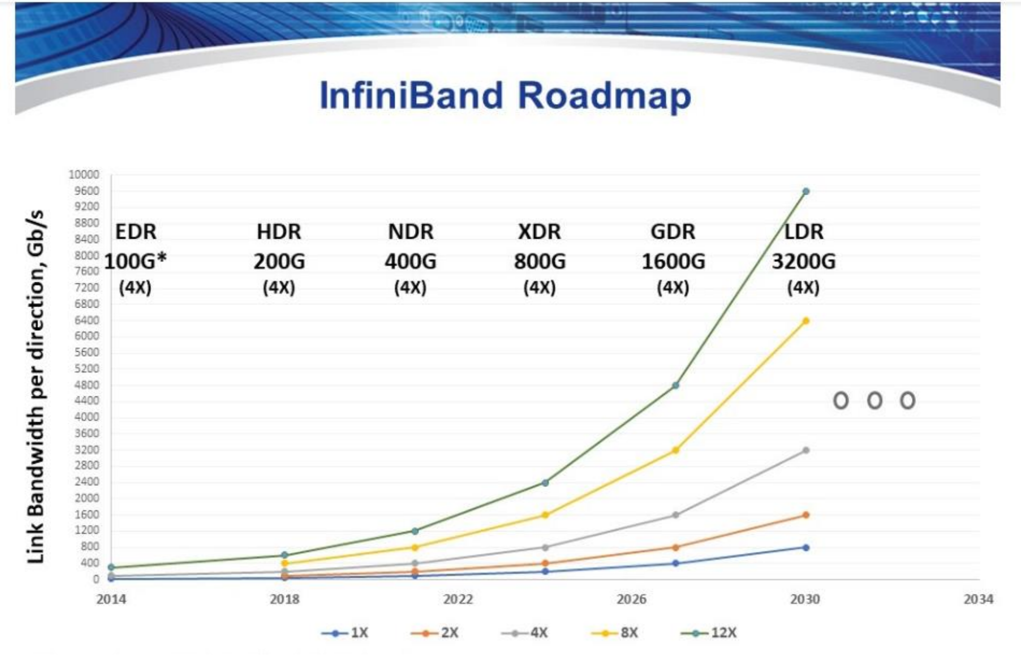
1,296 connections

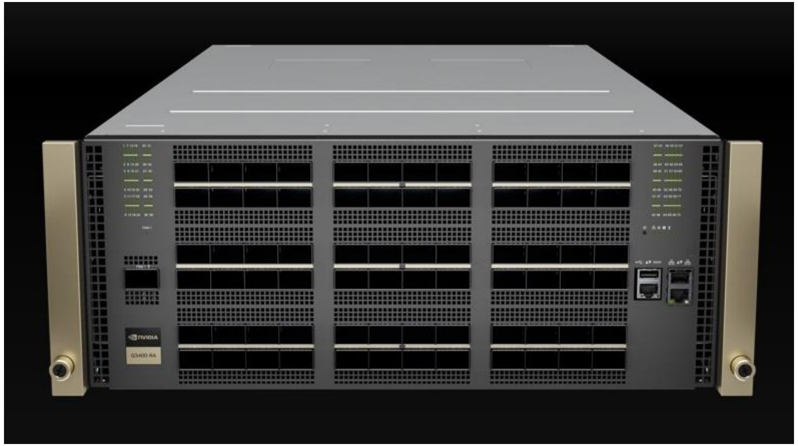
4 GPUs per compute tray

18 compute trays

9 NVLink Switch trays
2 NVLink Switches with 72 ports each

InfiniBand Creates Data Center Scale Memory with RDMA



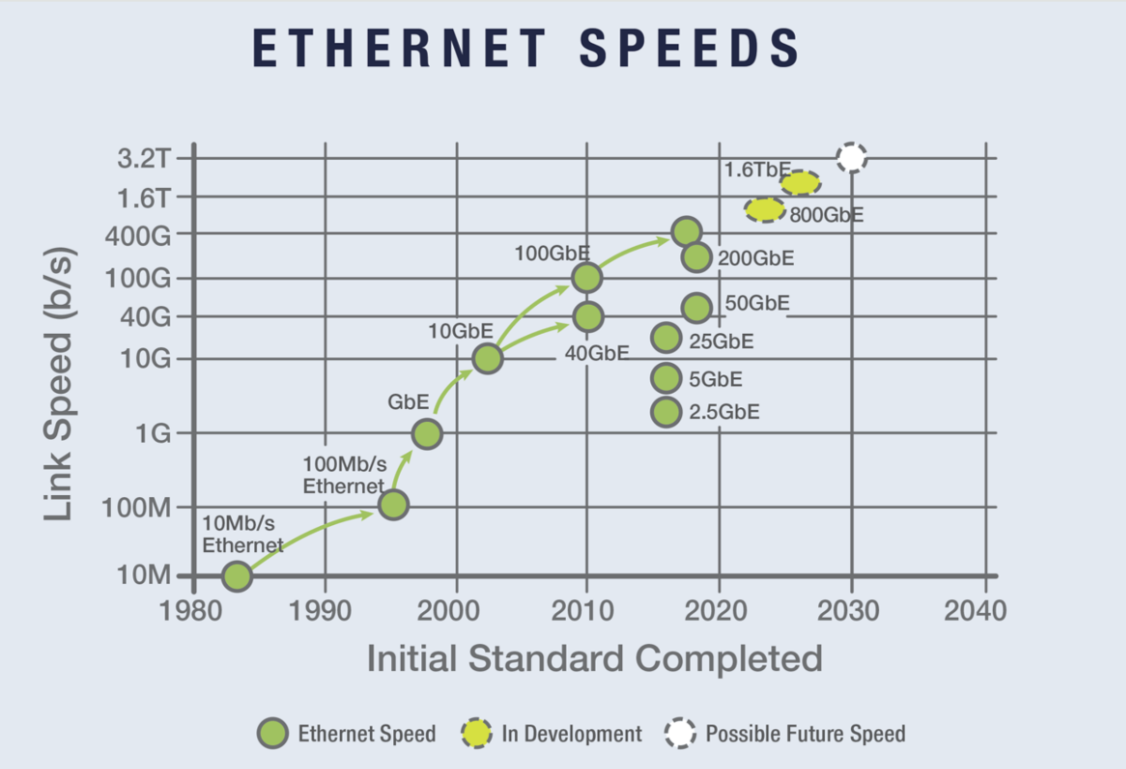


144 port 800Gb/port InfiniBand Switch

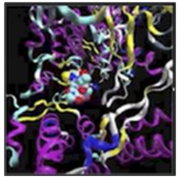
InfiniBand Scales to Hundreds of Thousands of Nodes

- 10,368 ports (2 levels)
- 746,496 ports (3 levels)
- Adaptive routing and congestion control
- Self-Healing
- Copper between switches (up to 1.5m)

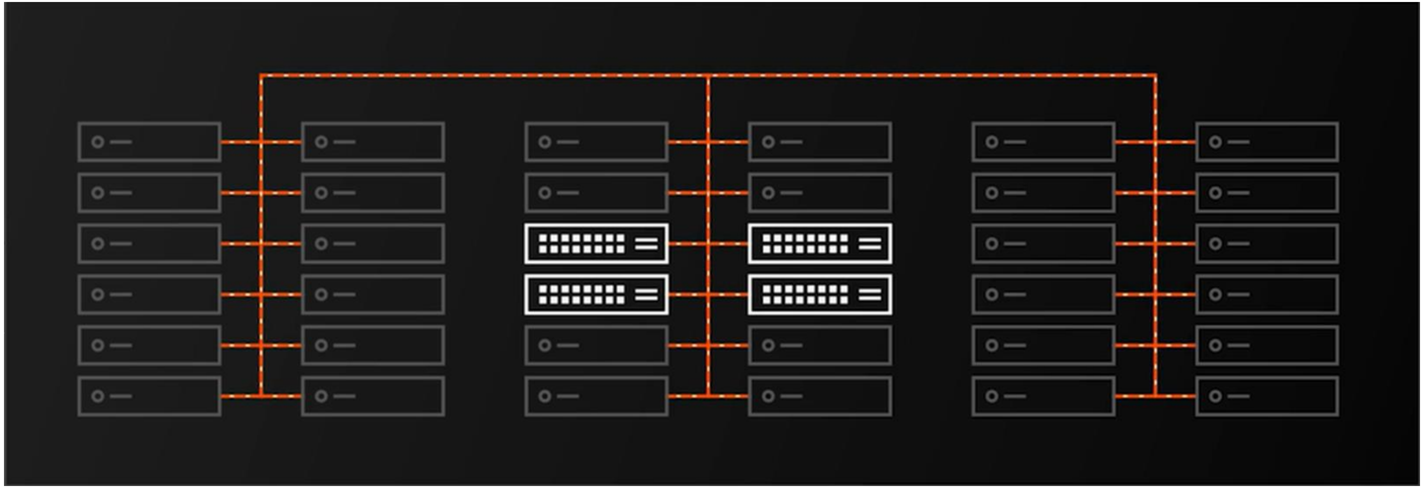
High Speed RoCE with Enhancements for AI



Challenges to Running AI Workloads on Traditional Ethernet



AI Workload

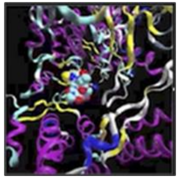


Significant Congestion

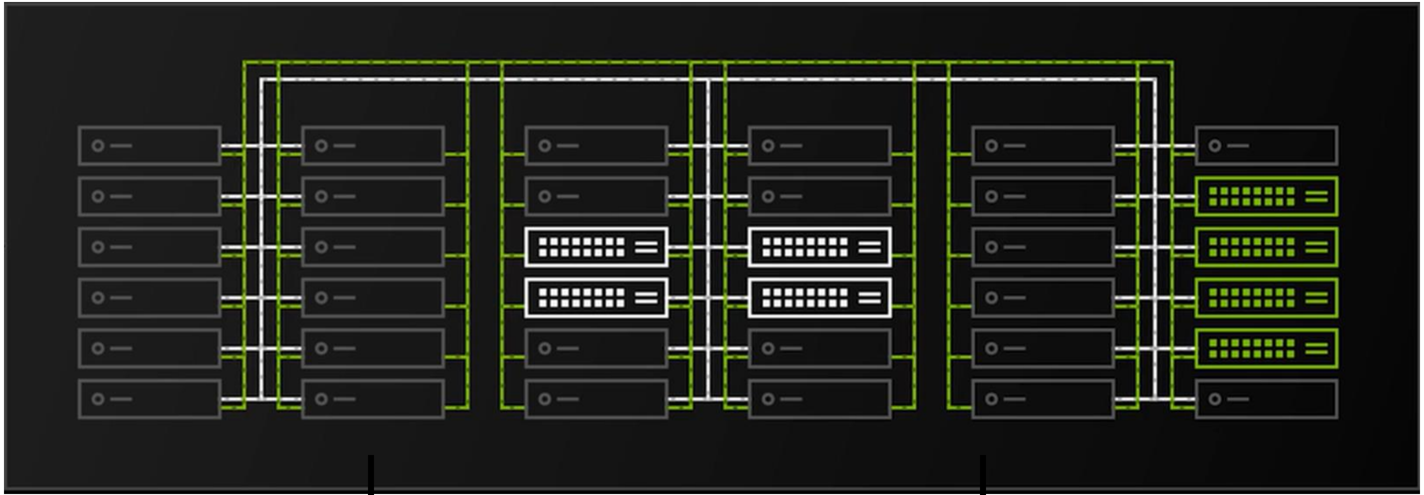
Increased Latency

Bandwidth Unfairness

Ethernet Enhancements for AI – Spectrum-X



AI Workload



95%

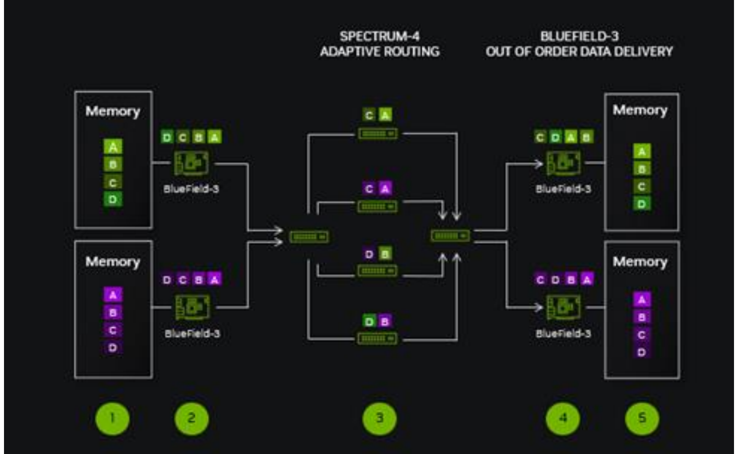
Higher
Effective
Bandwidth

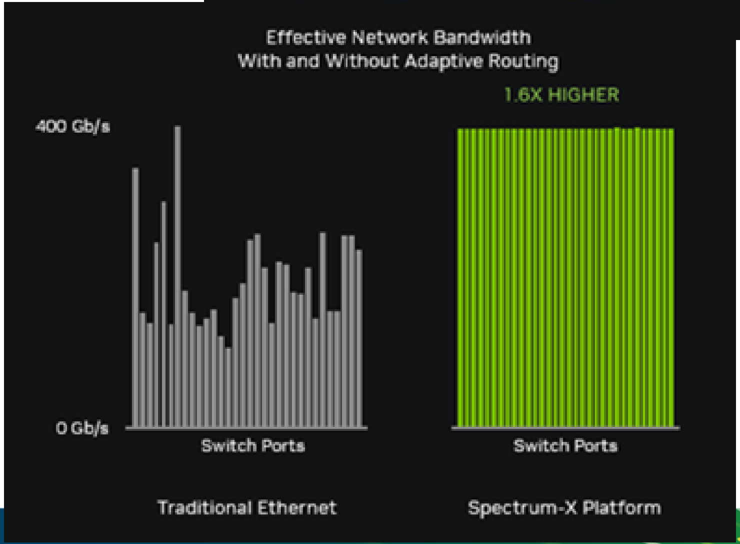
1.6X

Increased
AI Network
Performance

Packet Level Adaptive Routing

- The NIC sends data into the switch network
- The switch spreads the data packets across all available routes
- The NIC ensures in-order data delivery
- Increase from typical 60% to 95% effective bandwidth




Effective Network Bandwidth
With and Without Adaptive Routing

Call to Action

- NVLink, InfiniBand and Spectrum-X Ethernet solutions are here today to improve GPU efficiency at different performance and price points
- NVLink, InfiniBand and Spectrum-X Ethernet products are available and welcome AI solution partners to test them and show the advantages
- Reach out to us for any questions
- Where to find additional information (URL links)
  - https://www.nvidia.com/en-us/data-center/nvlink/
  - https://www.infinibandta.org/infiniband-roadmap/
  - https://www.nvidia.com/en-us/networking/spectrumx/

Thank you!