NVIDIA Networking Solutions for AI: NVLink, InfiniBand and Spectrum-X

Randy Splinter Ph.D., Senior Solution Architect | MOREnet Tech Summit/Feb 18, 2025

Randy@NVIDIA.com

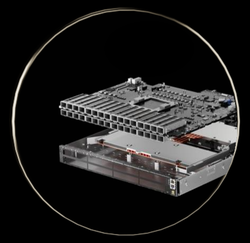
Agenda

- Generative AI is a Data Center Scale Computing Problem
- NVIDIA NVLink
- NVIDIA InfiniBand
- NVIDIA Spectrum-X
- Enabling Software: SHARP and NCCL
- NVIDIA Reference Architectures

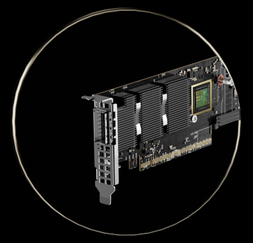
NVIDIA Networking Platforms

Over 100M Ports Shipped

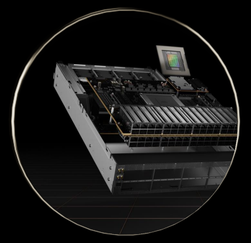
Scale-Out East-West InfiniBand | Scale-Out East-West AI Ethernet | Scale-Up NVLINK | North-South



QUANTUM
InfiniBand Switch



CONNECTX
SuperNIC



SPECTRUM
Ethernet AI Switch

BLUEFIELD
SuperNIC

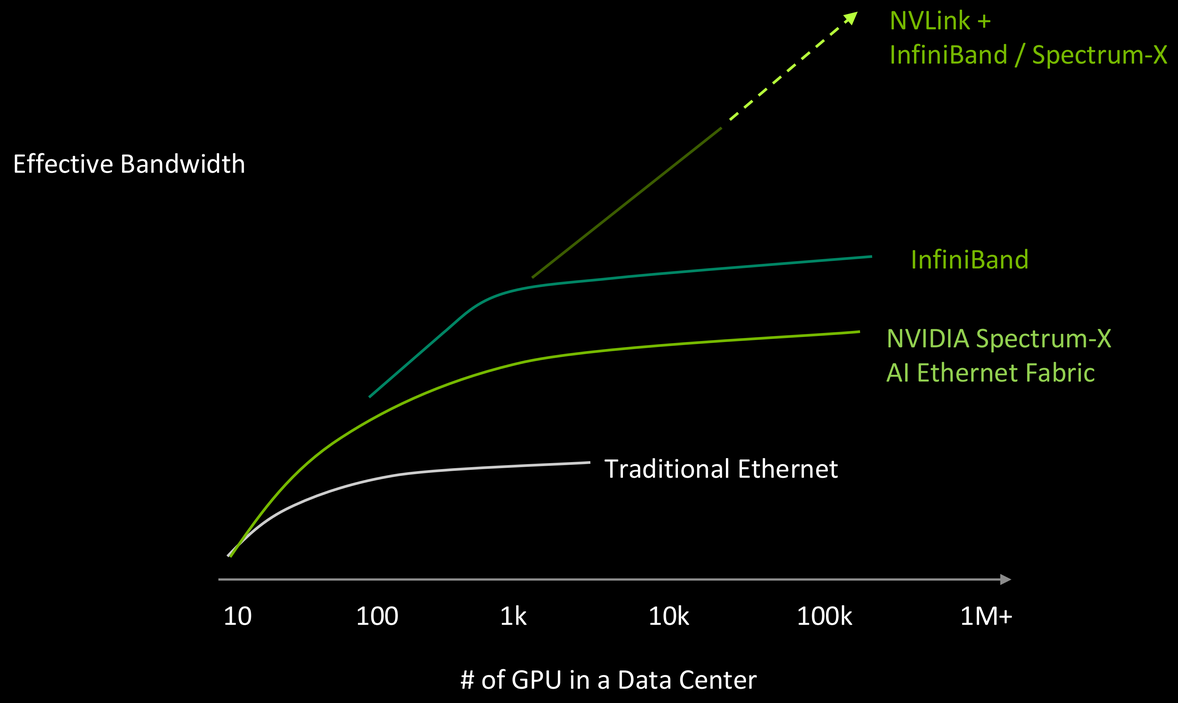
NVLINK
Switch

BLUEFIELD
DPU

Generative AI is a Data Center Scale Computing Problem

AI factories emerge as a new class of generative AI data centers

Purpose-built high performance networking is necessary to effectively scale AI



NVIDIA NVLink

Fastest Compute Fabric Connectivity

NVIDIA InfiniBand

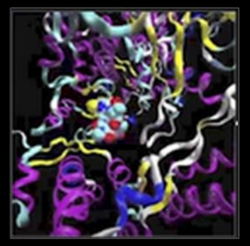
Gold Standard for Scale-Out AI Fabrics

NVIDIA Spectrum-X

Ethernet Optimized for Multi-Tenant AI Factories

Running AI Workloads on Traditional Networks

Sub-Optimal for Addressing Needs of AI



AI Workload

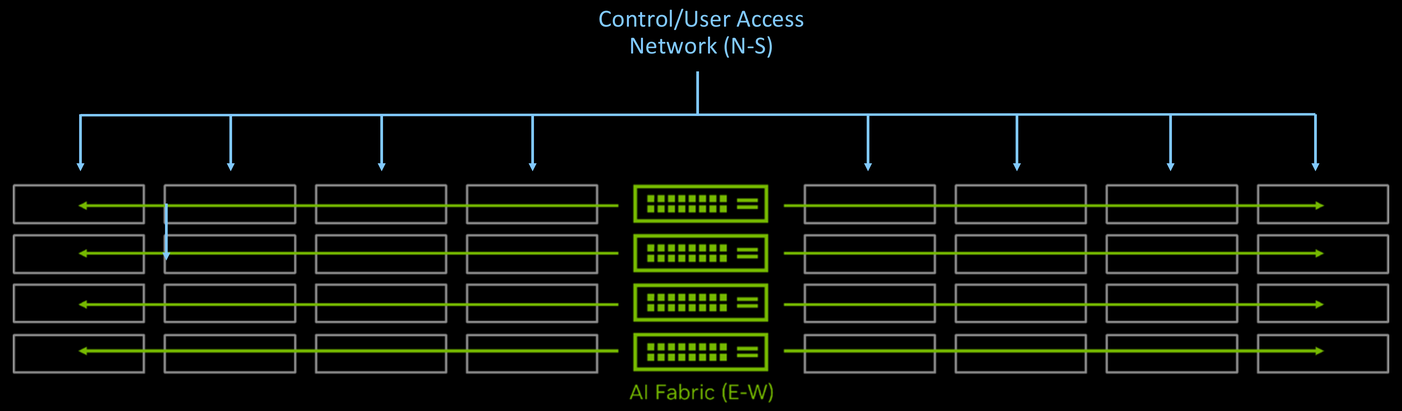
Significant Congestion

Increased Latency

Bandwidth Unfairness

Peak AI Performance Demands Optimized Networking

AI Workloads Require an AI Fabric



Control / User Access Network (N-S)		AI Fabric (E-W)
Loosely-Coupled Applications		Tightly-Coupled Processes
TCP (Low Bandwidth Flows and Utilization)		RDMA (High Bandwidth Flows and Utilization)
High Jitter Tolerance		Low Jitter Tolerance
Oversubscribed Topologies		Nonblocking Topologies
Heterogeneous Traffic, Statistical Multi-Pathing		Bursty Network Traffic, Predictive Performance

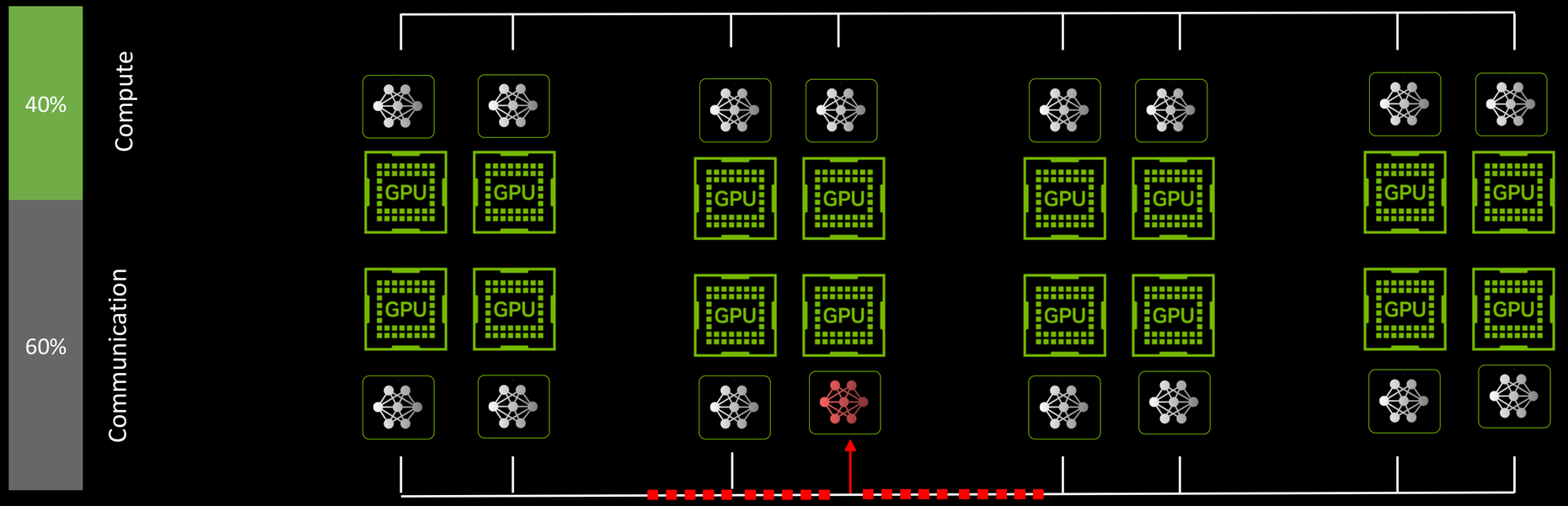
NVIDIA NVLink

Next Generation Models Communication Bottleneck

More Compute Needs to Balance with More Communications

Mixture of Expert Modes

GPT MoE1.8T Parameters



XDR Infiniband 800 GBps

15 GPUs Sending to 1 GPU

What is NVLink?

Brief Overview of the Technology

NVLink is a wire-based serial multi-lane near-range communications link developed by Nvidia. Unlike PCI Express, a device can consist of multiple NVLinks, and devices use mesh networking to communicate instead of a central hub. The protocol was first announced in March 2014 and uses a proprietary high-speed signaling interconnect (NVHS). [1]

- NVLink is developed by Nvidia for data and control code transfers in processor systems between CPUs and GPUs and solely between GPUs. NVLink specifies a point-to-point connection with data rates of 20, 25 and 50 Gbit/s (v1.0/v2.0/v3.0+ resp.) per differential pair. [1]

[1] Wikipedia NVLink Page

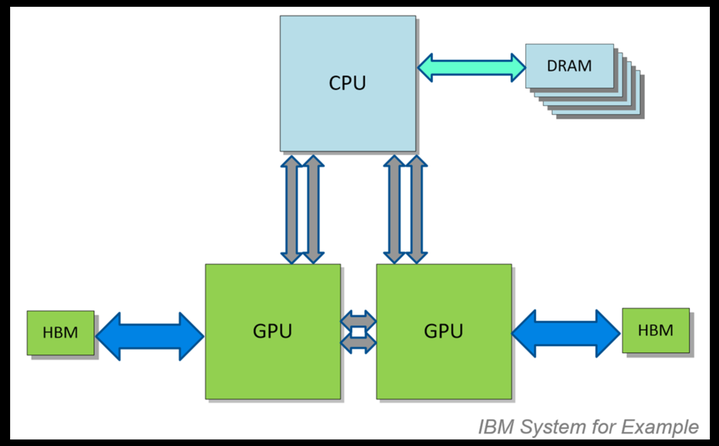
Interconnect	Transfer Rate	Modulation	Effective Payload Rate per Lane	Total Links (NVLink)	Total Bandwidth (bi-directional)	Realized in Design
NVLink 1.0	20 GT/s	PAM4 w/ FEC	20GB/s	4	160GB/s	Pascal, POWER8+
NVLink 2.0	25GT/s	NRZ	25GB/s	6	300GB/s	Volta, POWER9
NVLink 3.0	50GT/s	NRZ	25GB/s	12	600GB/s	Ampere
NVLink 4.0	50GT/s	PAM4	25GB/s	18	900GB/s	Hopper, NVIDIA Grace
NVLink 5.0	100GT/s	PAM4	50GB/s	18	1800GB/s	Blackwell, NVIDIA Grace

Motivation

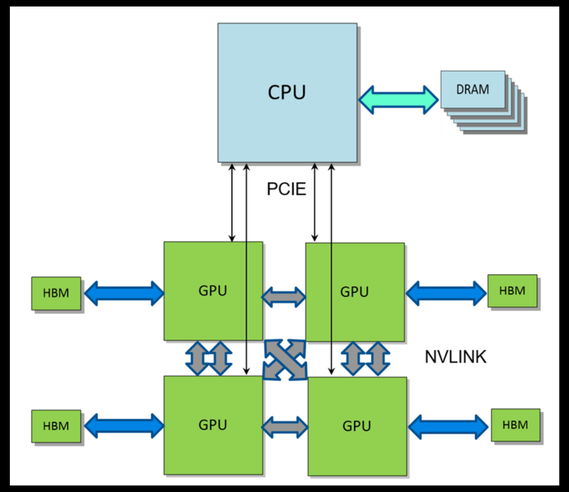
High BW Access to a Large Memory Footprint

- Workloads (Machine Learning, HPC) require large memory footprints
- GPU counts on high BW access to memory for throughput
- PCIe was a bottleneck to CPU's large memory pool

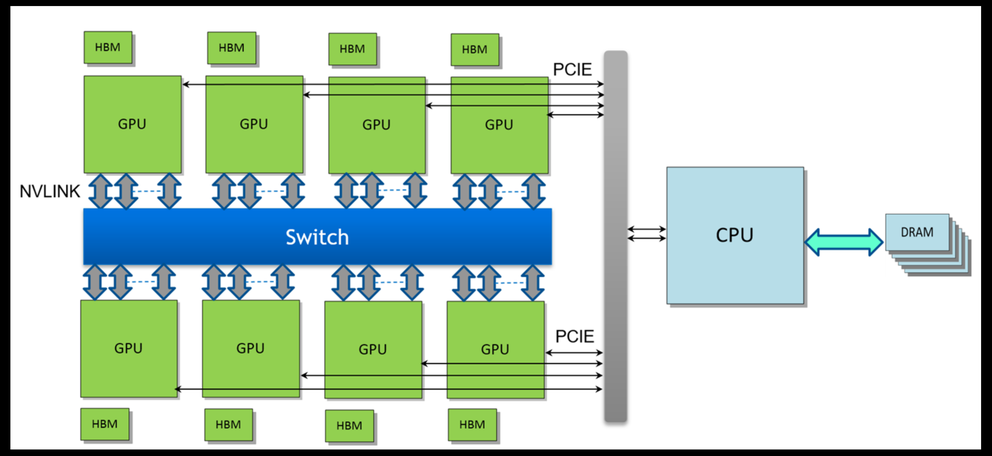
Larger Effective Compute Nodes



Links can be used for GPU⇔GPU connections



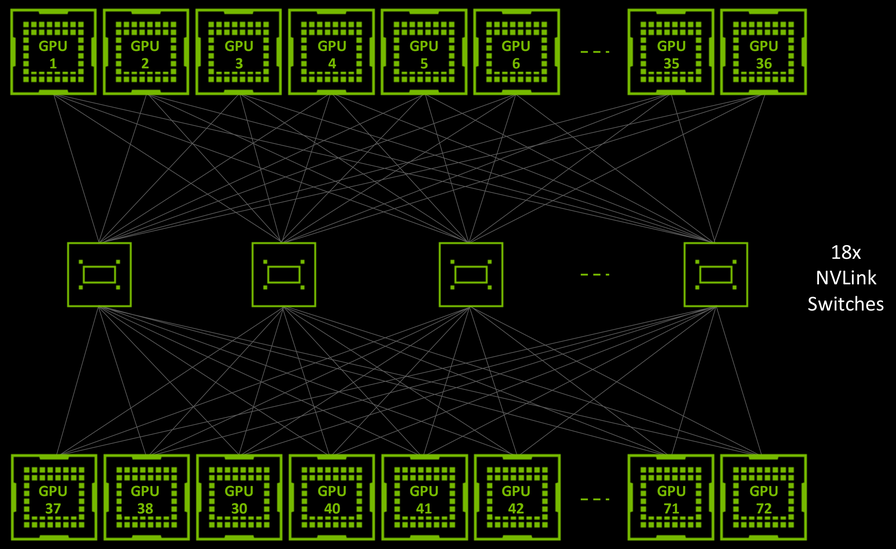
GPU Clustering allows for tight cooperation between GPUs working on the same problem



Switch Based Connections for higher/full all-to-all GPU Bandwidth

Minimizing Multi-GPU Communication Latency

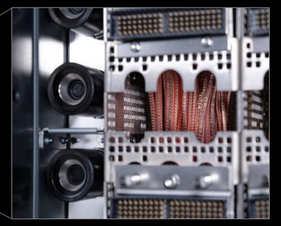
NVIDIA NVLink Switch and NVLink Create a Unified GPU



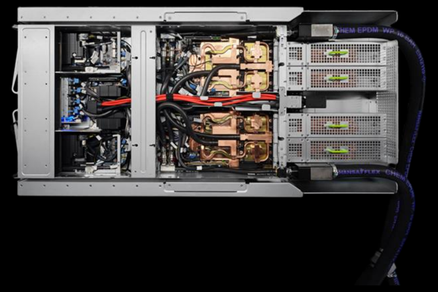
GB200 NVL72 Rack

NVLink Spine

NVLink Cables



NVLink Switch Tray



NVLink	NVL Domain	All-to-All	All-Reduce
1.8 TB/s	72 GPUs	130 TB/s	260 TB/s

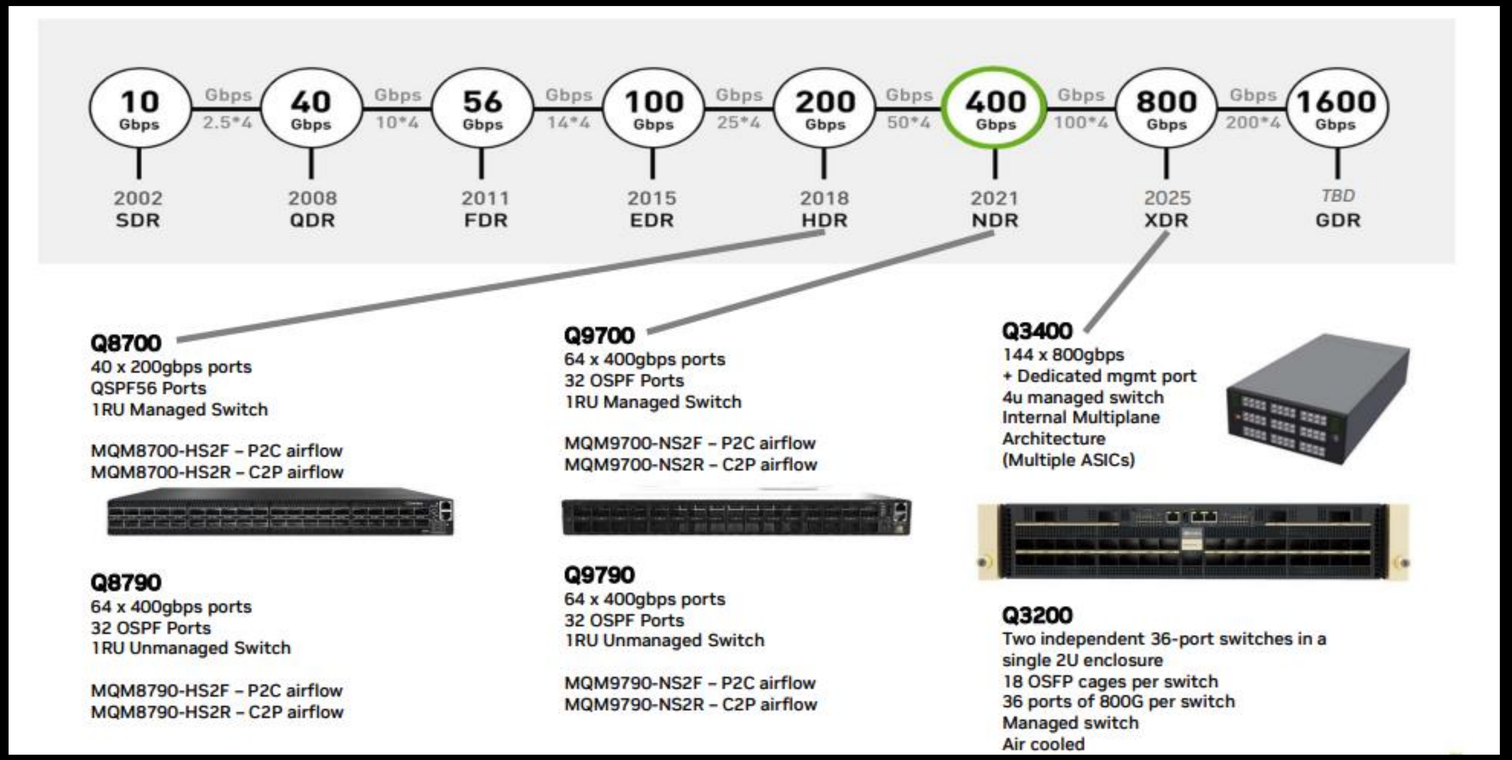
NVIDIA InfiniBand Components

What is InfiniBand?

Brief Overview of the Technology

- **InfiniBand** (**IB**) is a computer networking communications standard used in high-performance computing that features very high throughput and very low latency. It is designed to be scalable and uses a switched fabric network topology. [1]
  - Each link is duplex. Links can be aggregated: most systems use a 4 link/lane connector (QSFP). HDR often makes use of 2x links (aka HDR100, 100 Gb link using 2 lanes of HDR, while still using a QSFP connector). 8x is called for with NDR switch ports using OSFP (Octal Small Form Factor Pluggable) connectors. [1]

[1] Wikipedia InfiniBand Page



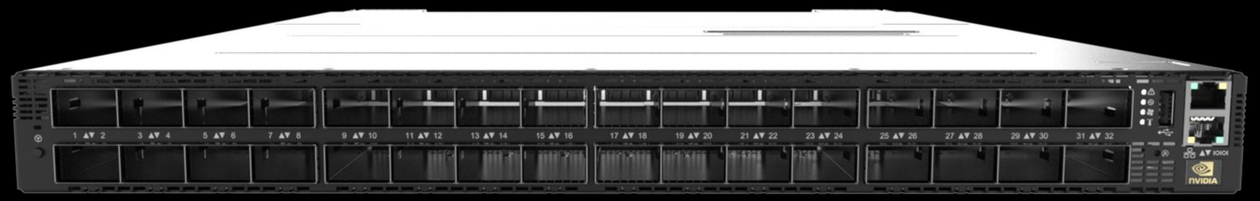
Quantum-2 Switch

QM9700 and QM9790 Family of 1U Switches

- 64 ports of 400Gb/s (NDR) over 32 OSFP cages
- 128 ports of 200Gb/s (NDR200)
- Secured Boot

- 51.2Tb/s aggregate bandwidth
- 66.5 billion packets per second
- SHARPv3 - low latency data reduction and streaming aggregation

- Internally managed (QM9700), and externally managed (QM9790) SKUs
- 26’’ depth, Air cooled
- 2 power supplies (1+1), hot swappable



NVIDIA Quantum-X800 InfiniBand Platform

Highest-Performance for Scientific Computing
And Large-Scale Machine Learning 800Gbps End-to-End

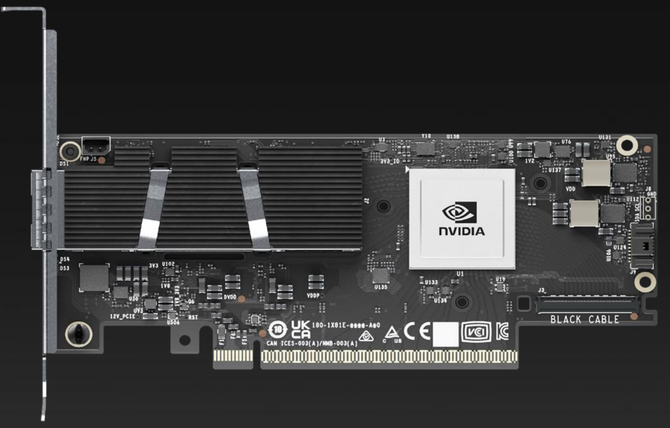
Quantum-X800 Q3400-RA Switch

- 115 Terabits per second bandwidth (5X higher)
- SHARP v4 with 14.4 TFlops of In-Network Computing (9X higher)
- Adaptive routing, congestion control, advanced power management

Quantum-X800 Q3400-RA Switch

ConnectX-8 SuperNIC

- In-Network Computing
- 32 lanes of PCIe Gen6.1, PCIe switch, Socket Direct

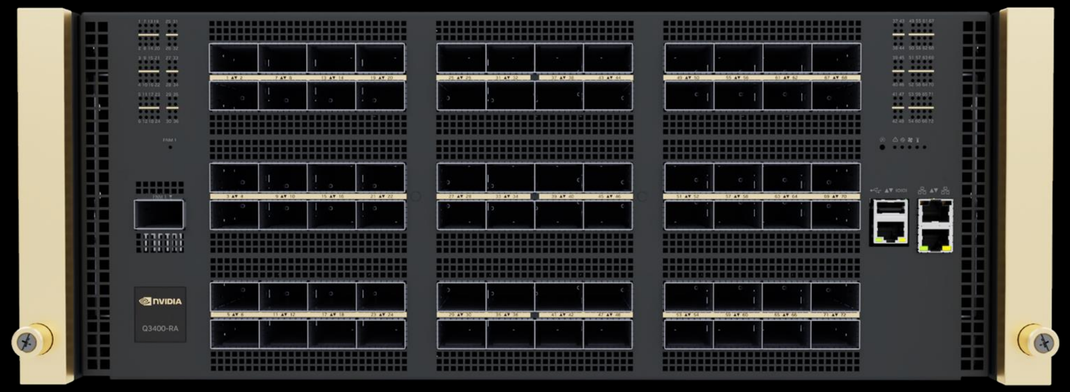


ConnectX-8 SuperNIC

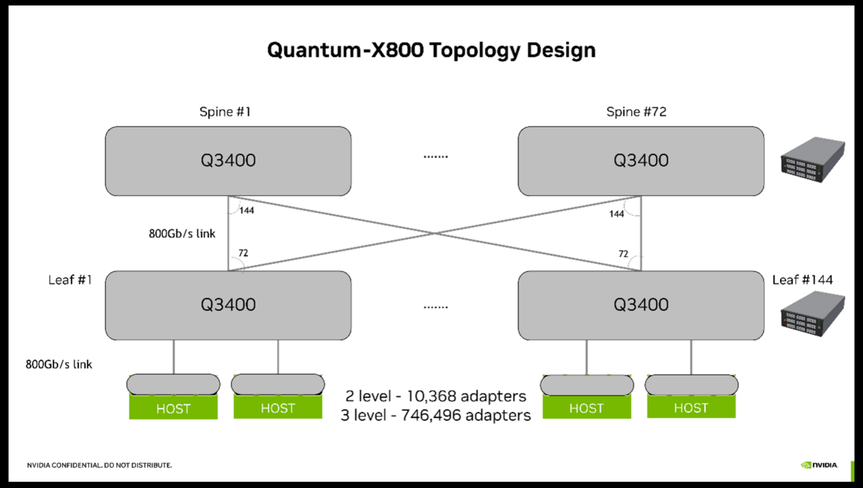
Quantum-X800 Q3400 Systems

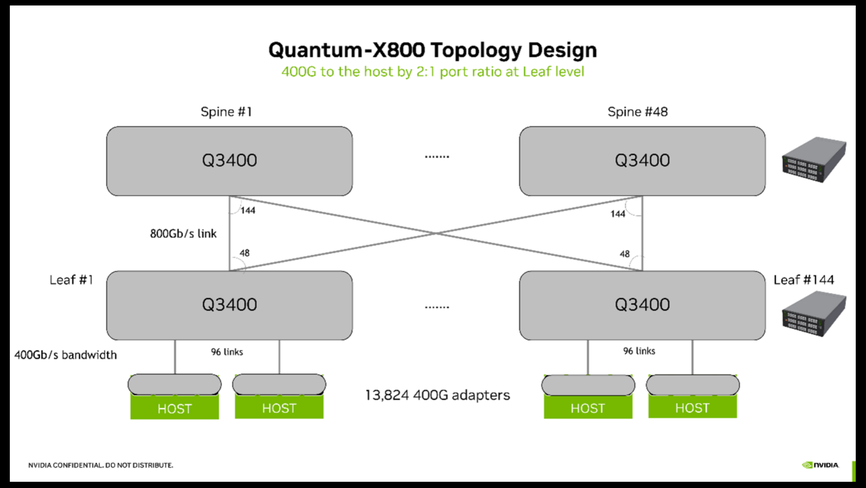
Quantum-X800 Q3400 4U switch platform

- 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 and liquid-cooled systems



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



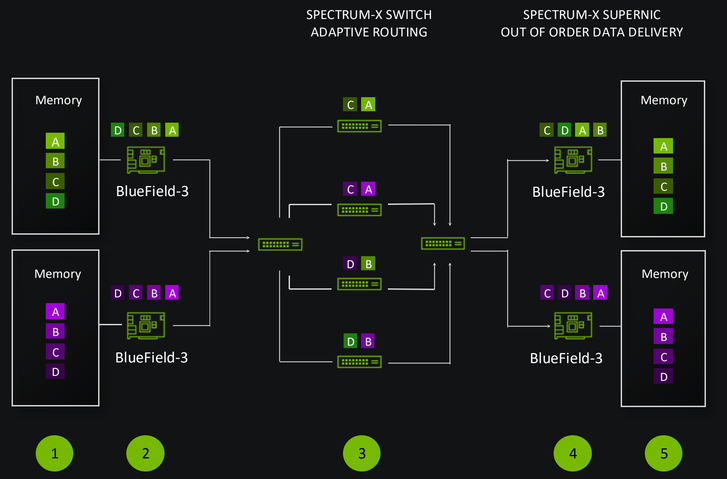


Spectrum-X Platform Components

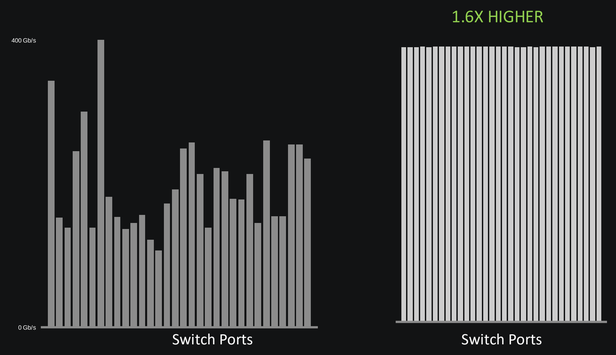
End-to-End Adaptive RDMA Routing With Lossless Network

Increases effective data throughout by 1.6X

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



Effective Network Bandwidth
With and Without Adaptive Routing

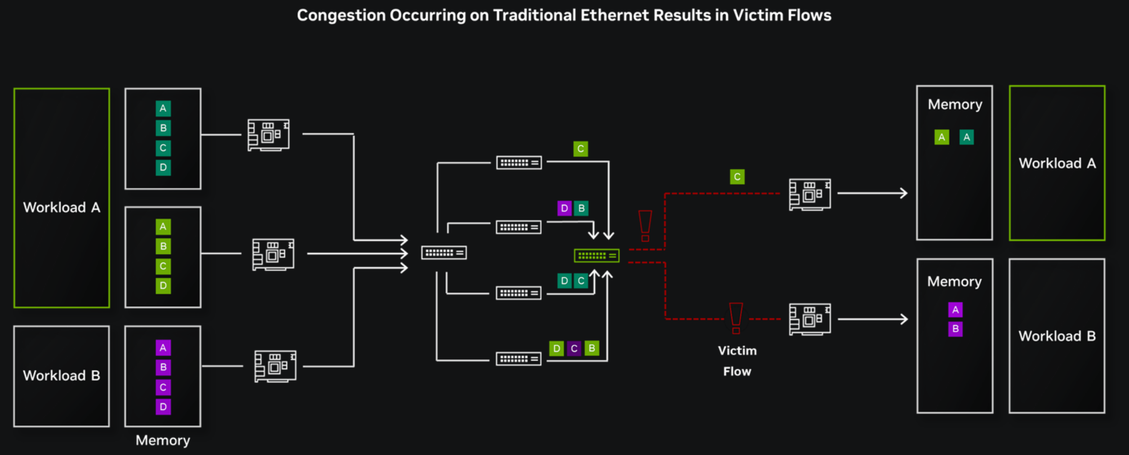


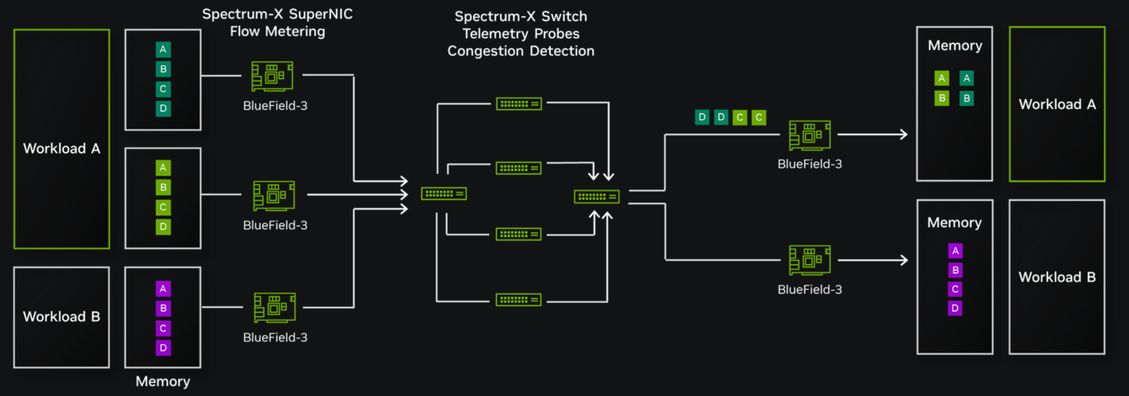
Traditional Ethernet

Spectrum-X Platform

Noise Isolation With Programable Congestion Control

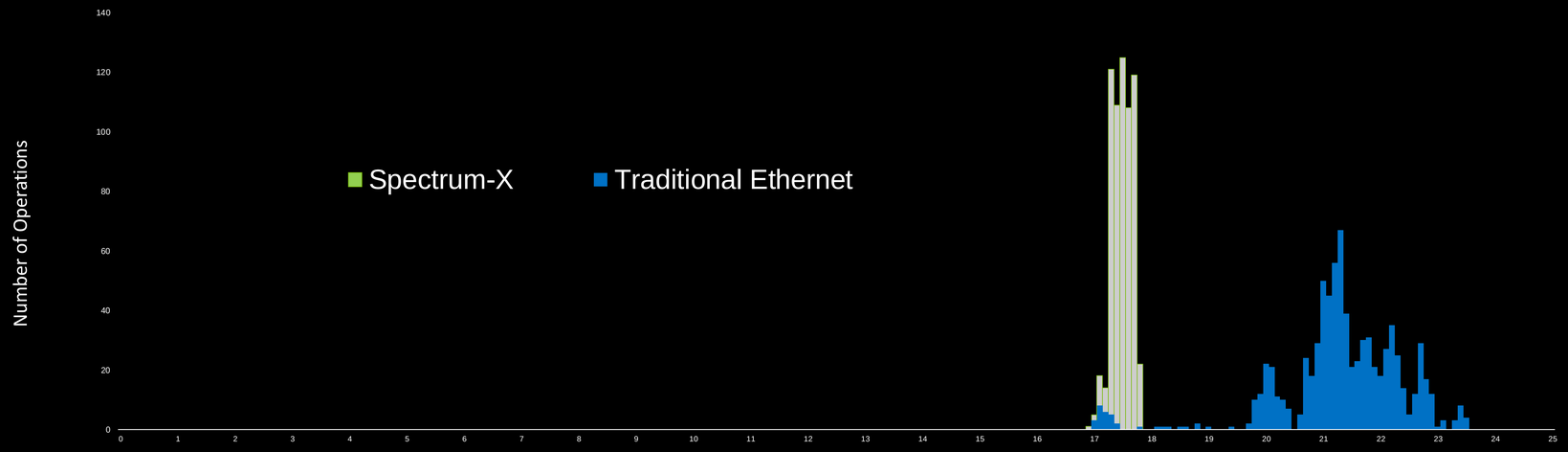
- Diverse workloads can impact each other’s performance
- Spectrum-X detects congestion spots in real time
- Programmable congestion control meters the data flow
- Results in performance isolation across workloads

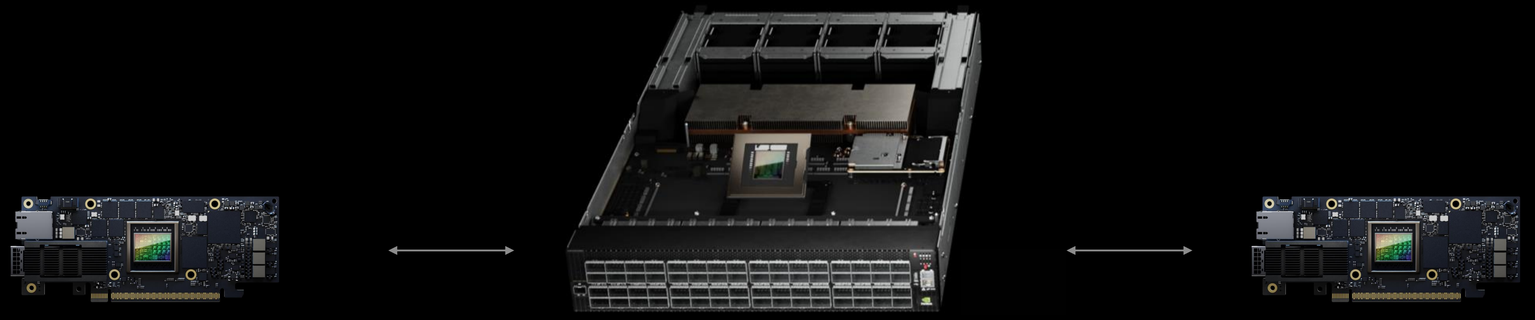




What Makes Spectrum-X Special

Switch-to-SuperNIC, End-to-End Network Processing, Bringing High Performance to Ethernet





Spectrum-X800 and BlueField-3 SuperNIC AI on Ethernet

Schedule Data Transmission to Avoid Congestion

Ultra-High-Speed Traffic Monitoring Distribute Data Across All Switch Ports Ignoring Data Ordering

Reordering - Receive Data and Place it Back in Order

Spectrum-X Ethernet Accelerates World's Largest AI Supercomputer

122 Days

vs.
Months

Time to build supporting facility and state-of-the-art supercomputer

19 Days

vs.
Months

Time from the first rack rolled onto the floor until training began

Zero Collisions

vs.
Thousands

No data flow collisions on a three-tier network

95%

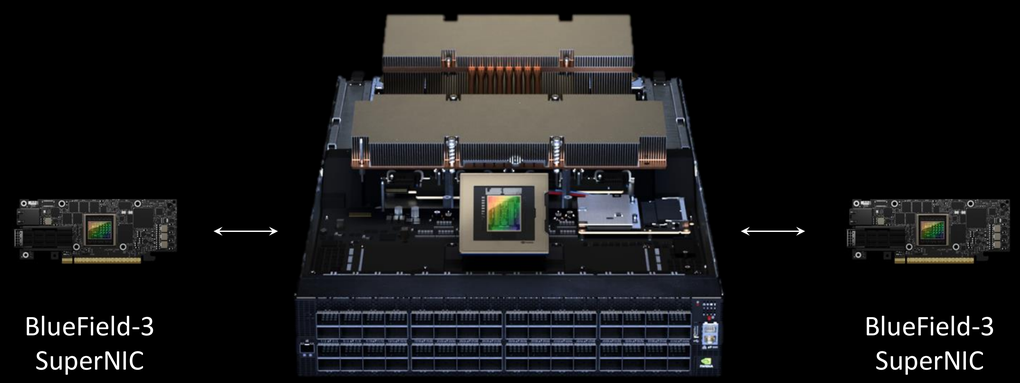
vs.
60%

Effective data throughput

NVIDIA Spectrum-X: World's First Ethernet Platform for AI

Combining Specialized High-Performance Architecture with Standard Ethernet Connectivity

- Nearly Perfect Effective Bandwidth at Scale
- Extremely Low Latency
- NVIDIA RoCE Extensions for Scalable AI Communications
- Full Stack and End-to-End Optimization
- Deterministic Performance and Performance Isolation
- Open NOSes: SONiC, Cumulus
- Standard Ethernet Connectivity



BlueField-3 SuperNIC

Spectrum SN5600 Switch

BlueField-3 SuperNIC

Spectrum-X

- 100 billion transistors, TSMC 4N
- 51.2T bandwidth, 100G SerDes
- 64 X 800G Ports, 128 X 400G ports
- 8K GPUs in Two-Tiers
- End-to-end optimized with BlueField-3

BlueField-3

- 16 Arm 64-Bit Cores
- 16 Core / 256 Threads Datapath Accelerator
- 400Gb/s Ethernet Networking
- DDR memory interface
- PCIe switch

NVIDIA Spectrum-X Ethernet Platform

High-Performance Ethernet for AI

Spectrum SN5600 Switch

- 51.2 Terabits per second bandwidth (4X higher)
- 128x 400Gbps; 64x 800Gbps
- Adaptive routing, congestion control, high frequency telemetry

Spectrum SN5600 Switch

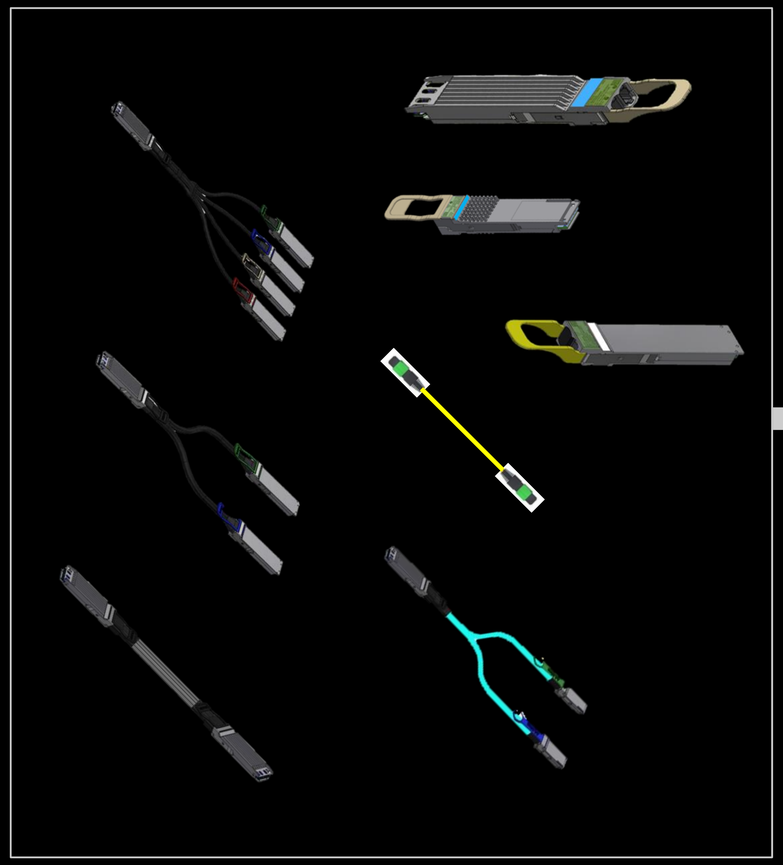
BlueField-3 SuperNIC

- Best-in-class RoCE for AI workloads
- Multi-tenancy at massive scale
- Power efficient, low-profile design

BlueField-3 SuperNIC

NVIDIA AI Networking: LinkX Cables and Transceivers

Optimized End-to-End Connectivity for NVIDIA AI Solutions



LinkX Cables & Transceivers

Performance Optimized for NVIDIA AI Solutions

Minimal bit errors, low error correction delays, high data rate performance, and optimal thermal management

- 25G/40G/100G/200G/400G/800G/1.6T
- Ethernet & InfiniBand

Best Signal Integrity & Reliability

Work seamlessly out of the box with rigorous design, simulation, and extensive live testing processes

- Rigorous quality control
- Multi-source supply chain

Higher ROI with End-to-End NVIDIA Network

Works in unison with NVIDIA's specialized hardware, software and firmware to monitor links ensuring maximum uptime

- UFM Integration
- Remote In-Service Software Installation

Passive Copper

Active Copper

Active Optical

Multimode transceivers

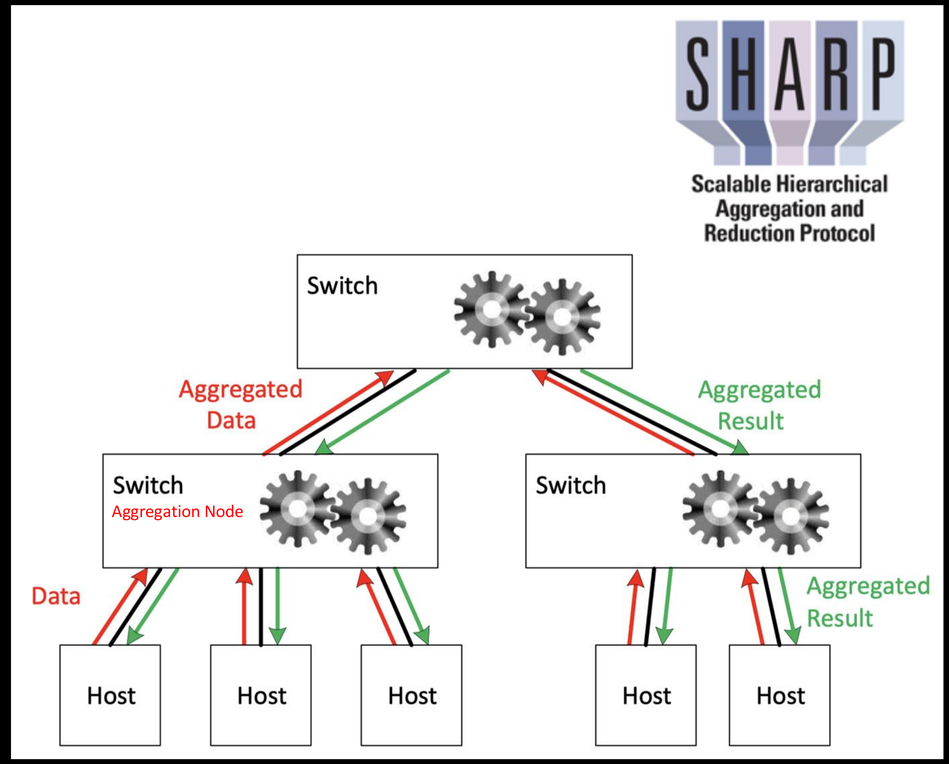
Single Mode Transceivers

NVIDIA Enabling Software Technologies: SHARP and NCCL

SHARP

In Network Computing

- Scalable Hierarchical and Reduction Protocol
  - SHARP is the in-network compute capability of Nvidia's InfiniBand switches.
    - Best-suited for hierarchical collective operations such as scatter, gather, reductions or barriers.
  - Offloading computations to the network has a dual benefit:
    - Freeing up compute resources.
    - Augmenting the algorithmic design space to unlock better new algorithms.
- Algorithmic View of SHARP
  - SHARP represents every compute-capable network element as an Aggregation Node.
  - Aggregation Nodes form SHARP Trees, which span the participating compute nodes.
  - Compute nodes push collective operations onto the SHARP tree. Receive the result when it is ready.
  - SHARP operations function in an up-down, hop-by-hop-synchronous fashion.

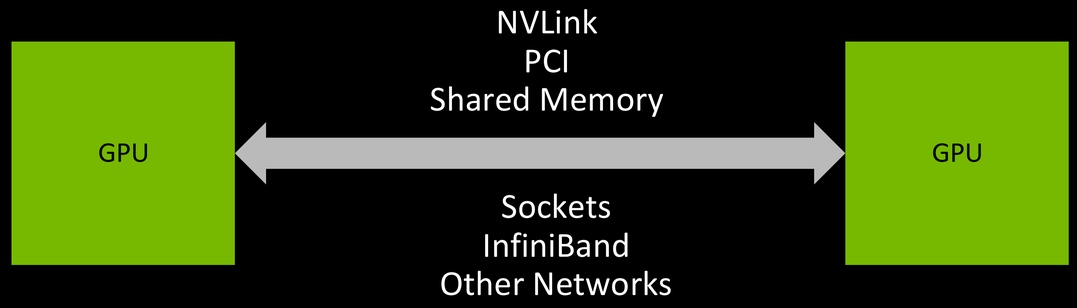


Optimized Inter-GPU Communication

NVIDIA Collective Communication Library (NCCL)

NVIDIA NCCL (NVIDIA Collective Communications Library) is designed for fast and efficient communication between GPUs within and across nodes.

- NCCL provides collective operations like AllReduce, Broadcast, Reduce, AllGather, and ReduceScatter for synchronizing data across GPUs.
- NCCL is topology-aware, optimizing communication paths based on interconnects like PCIe, NVLink, or InfiniBand for high bandwidth and low latency.
- NCCL integrates seamlessly with deep learning frameworks (e.g., PyTorch, TensorFlow) to accelerate distributed training on multi-GPU systems.
- NCCL achieves high performance by combining communication and computation in a single CUDA kernel.



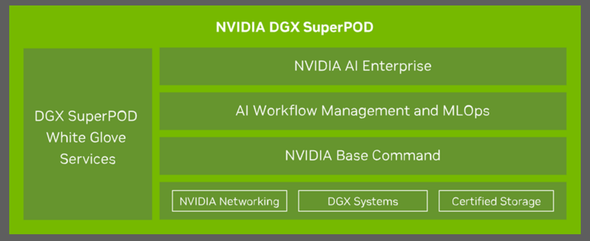
Binaries : https://developer.nvidia.com/nccl and in NGC containers Source code : https://github.com/nvidia/nccl
Perf tests : https://github.com/nvidia/nccl-tests

NVIDIA Architectures for AI

NVIDIA DGX SuperPOD

The fastest path to AI-innovation at scale

- Benefits:
  - World’s fastest commercially-available AI infrastructure
  - Turnkey solution that eliminates design complexity
  - Integrated, optimized software that keeps getting faster
  - Intelligently designed for your business, without upheaval
  - Complete full-stack offering, backed by NVIDIA
  - White-glove implementation / ramp-up service



DGX SuperPOD Software Stack

32 node DGX SuperPOD Scalable Unit

Compute
- Start with (32) DGX H100, H200 or DGX B200, 1 exaflop, scale modularly as needed

Storage
- High performance storage

Network
- 800 Gbps NVIDIA InfiniBand fabric
- NVL72 (coming)

Software
- NVIDIA Base Command
- NVIDIA AI Enterprise

Services
- NVIDIA Advanced Services
- NVIDIA Partner Services
- NVIDIA Support Services including Technical Account Manager
- Optional colocation via DGX-Ready Data Center Program

PLAN / DEPLOY

- Capacity planning
- Data center design
- Performance projection
- Site eval /prep
- Installation
- Post-install testing
- Provisioning /management

RAMP / OPTIMIZE

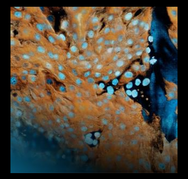
- Application performance testing
- Site documentation package
- User / DevOps training
- Workload-based DLI(s)
- Custom system runbook
- Hand-over session

NVIDIA DGX SuperPOD delivers the “SuperPOD Experience” to every organization that needs leadership-class infrastructure, with a white-glove implementation that’s optimized to your business so your team can deliver results sooner.

DGX Full-Stack Software

Enterprise software that drives the value of AI investments

Modern AI Use Cases

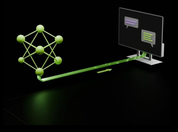


Biomedical

Speech & Translation

Route Planning

Content Generation



Reasoning

More….

End-to-End AI Development Tools
NVIDIA NIM Microservices, NVIDIA CUDA-X Microservices, and AI Application Frameworks

NVIDIA AI Enterprise

Comprehensive AI infrastructure software
for streamlined orchestration & cluster management, acceleration libraries, and optimized DGX OS from edge to core to cloud

NVIDIA Base Command

NVIDIA Advanced Cluster Management

Next-gen AI infrastructure software
optimized for new architectures to deliver Base Command capabilities + enhanced infrastructure reliability

DGX with Hopper

DGX with Blackwell

Accelerated Computing
with world-class DGX infrastructure

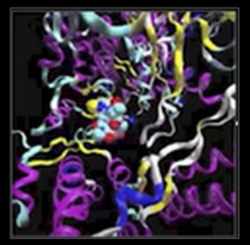
NVIDIA DGX GB200

Always-available enterprise infrastructure for mission-critical AI

- The building block of DGX SuperPOD with DGX GB200 systems
- Based on the NVIDIA GB200 NVL72
- Provides a fully-integrated, ready-to-scale infrastructure solution for generative AI
- Built with 36 GB200 Superchips and fifth-gen NVLink
- Connects 36 Grace CPUs and 72 Blackwell GPUs for compute intensive workloads
- 1.4 exaFLOPS of AI performance and 30TB of fast memory
- Handles the most complex generative AI workloads

The Problems of E-W Compute Networks for AI

Why Tail-End Latency is Critical to Resolve.



AI Workload

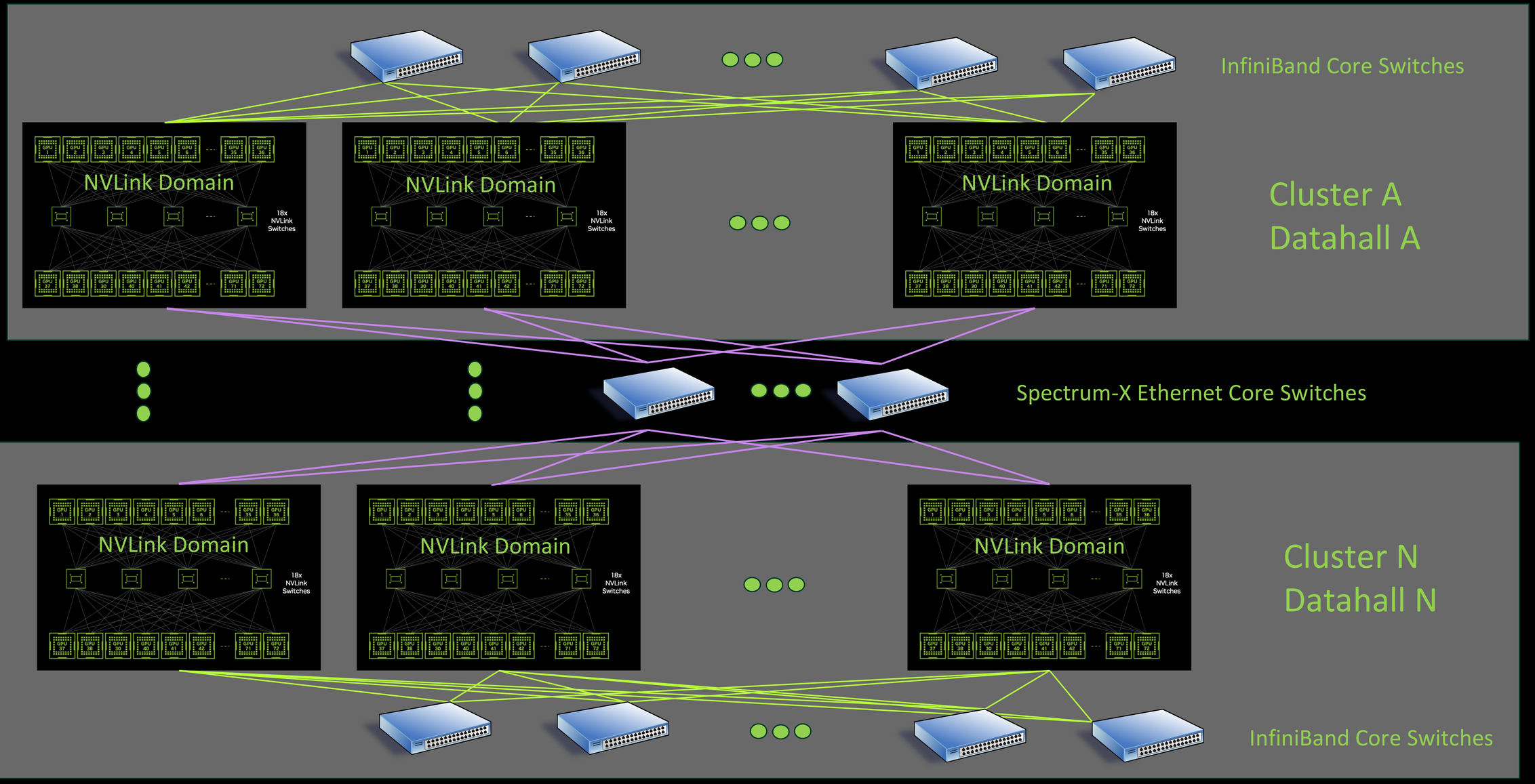
Significant Congestion

Increased Latency

Bandwidth Unfairness

The Challenge

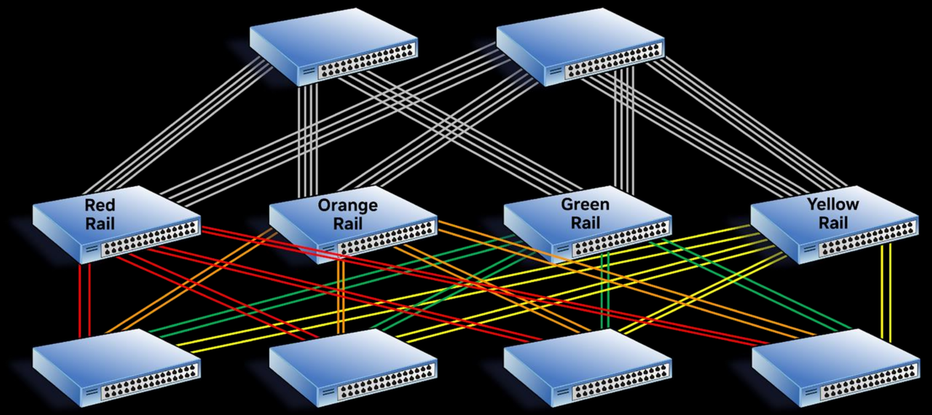
To Solve the Problems of Congestion, Increasing Latency and Bandwidth Fairness While Enable Scaling Out…



Understanding Rail-Optimized vs. Top of Rack

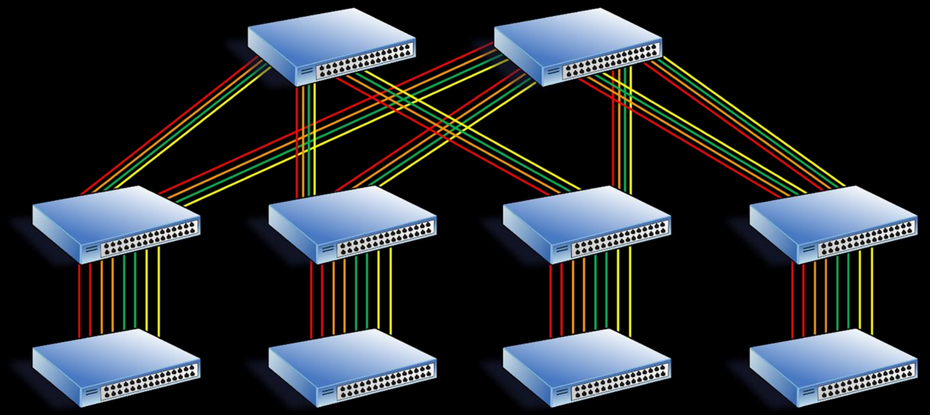
Rail Optimized Architectures for Peak Learning Performance

Latency Optimized



Rail-Optimized End of Row Design:

- Defined by GPU connectivity
- SHARP collectives can be executed on multiple rails simultaneously
- NCCL-optimized topology
- Copper cables from leaf to spine
- Optics between leaf and servers
- Higher AI performance
- Lowers latency between GPUs
- Reduces spine traffic



Top of Rack Topology:

- Defined by physical proximity (rack)
- Collectives on multiple rails increases network bandwidth requirements
- Cable-optimized topology
- Copper cables from server to leaf
- Optics between leaf and spine
- Lower AI performance
- 3x higher switch latency between GPUs
- Higher spine congestion

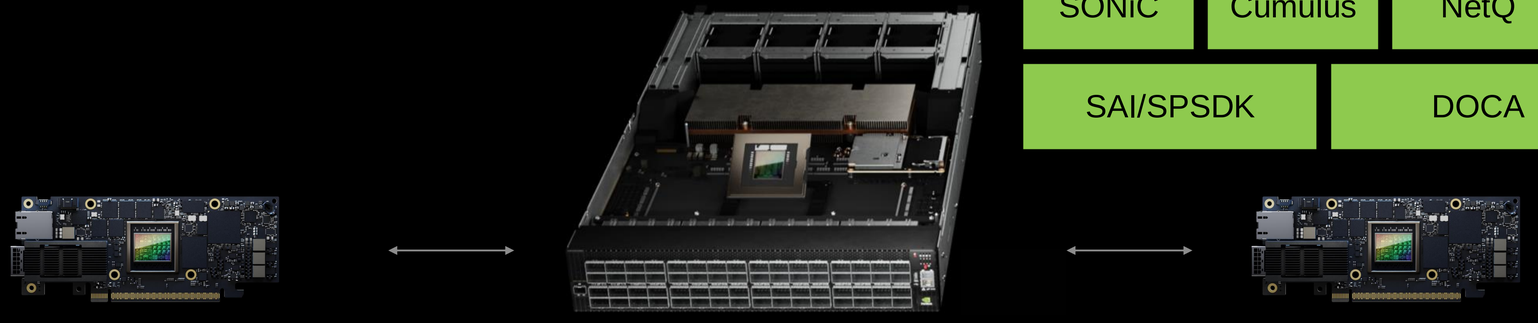
What Makes Spectrum-X Special

Switch-to-SuperNIC, End-to-End Network Processing, Bringing High Performance to Ethernet

Load Balancing	Tail Latency	Noise Isolation	Resilient Performance	High-Frequency Telemetry	Datapath Programmability
1.6X	**1.3X**	**2.2X**	**1.3X**	**1,000X**	**16 Cores 256 Threads**
Higher Effective Bandwidth	**Higher Collectives Bandwidth**	**Higher All-Reduce Bandwidth**	**Higher All-to-All Bandwidth**	**Faster Telemetry Collection**	**In-Network Configurable Computing**
Optimal usage of all network connections	Reducing Collectives Tail Latency	Eliminating network hot spots	Rebalances data flows avoiding failed links	Real-time adaption to network conditions	Support for customized end-user AI frameworks

SONiC | Cumulus | NetQ | NCCL | Air

SAI/SPSDK | DOCA | Magnum IO



Spectrum-X800 and BlueField-3 SuperNIC AI on Ethernet

Schedule Data Transmission to Avoid Congestion

Ultra-High-Speed Traffic Monitoring Distribute Data Across All Switch Ports Ignoring Data Ordering

Reordering - Receive Data and Place it Back in Order

Additional Resources

Networking for AI

Networking for AI Video

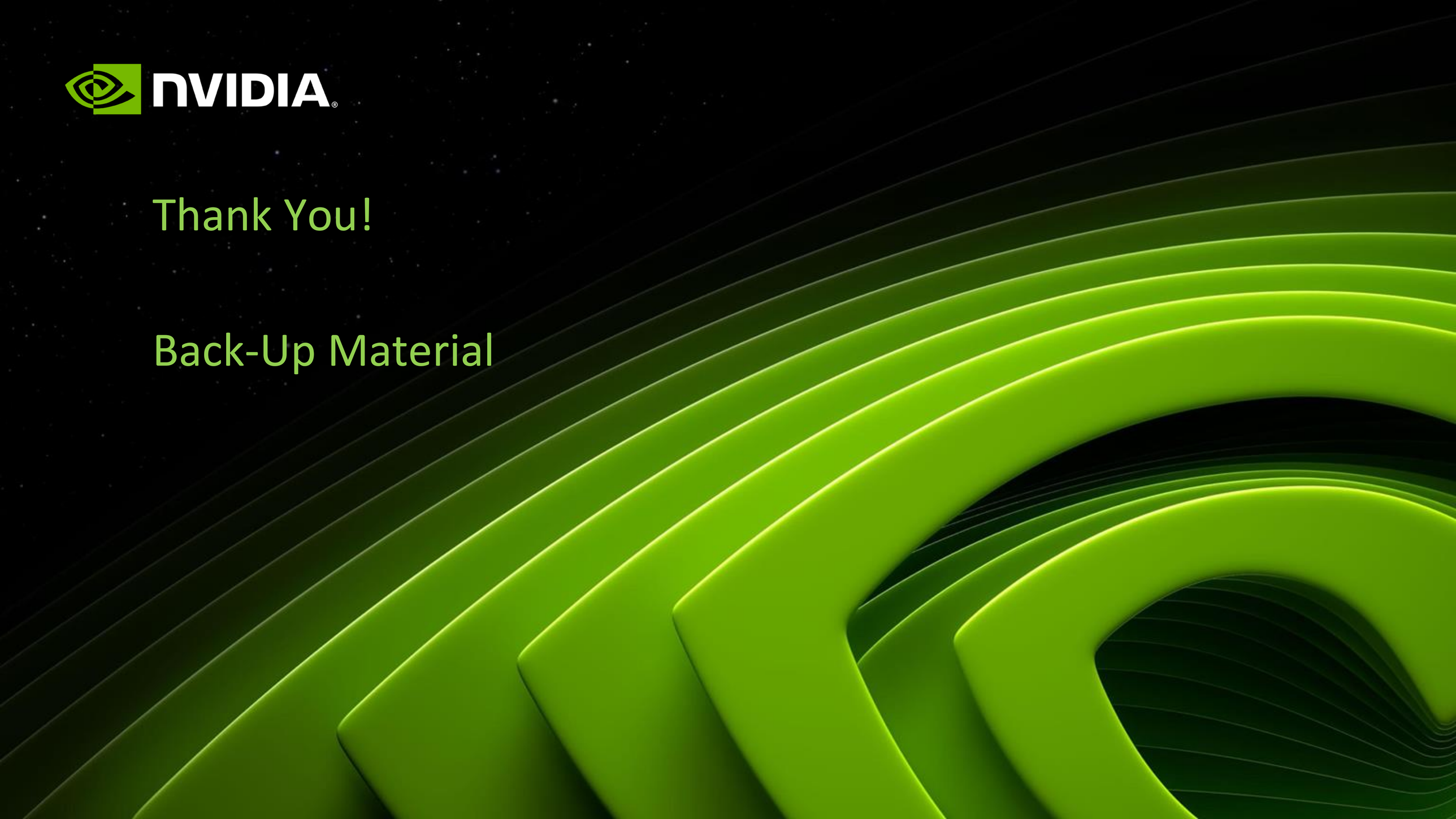
Spectrum-X Video

Networking for AI Whitepaper

Spectrum-X Whitepaper

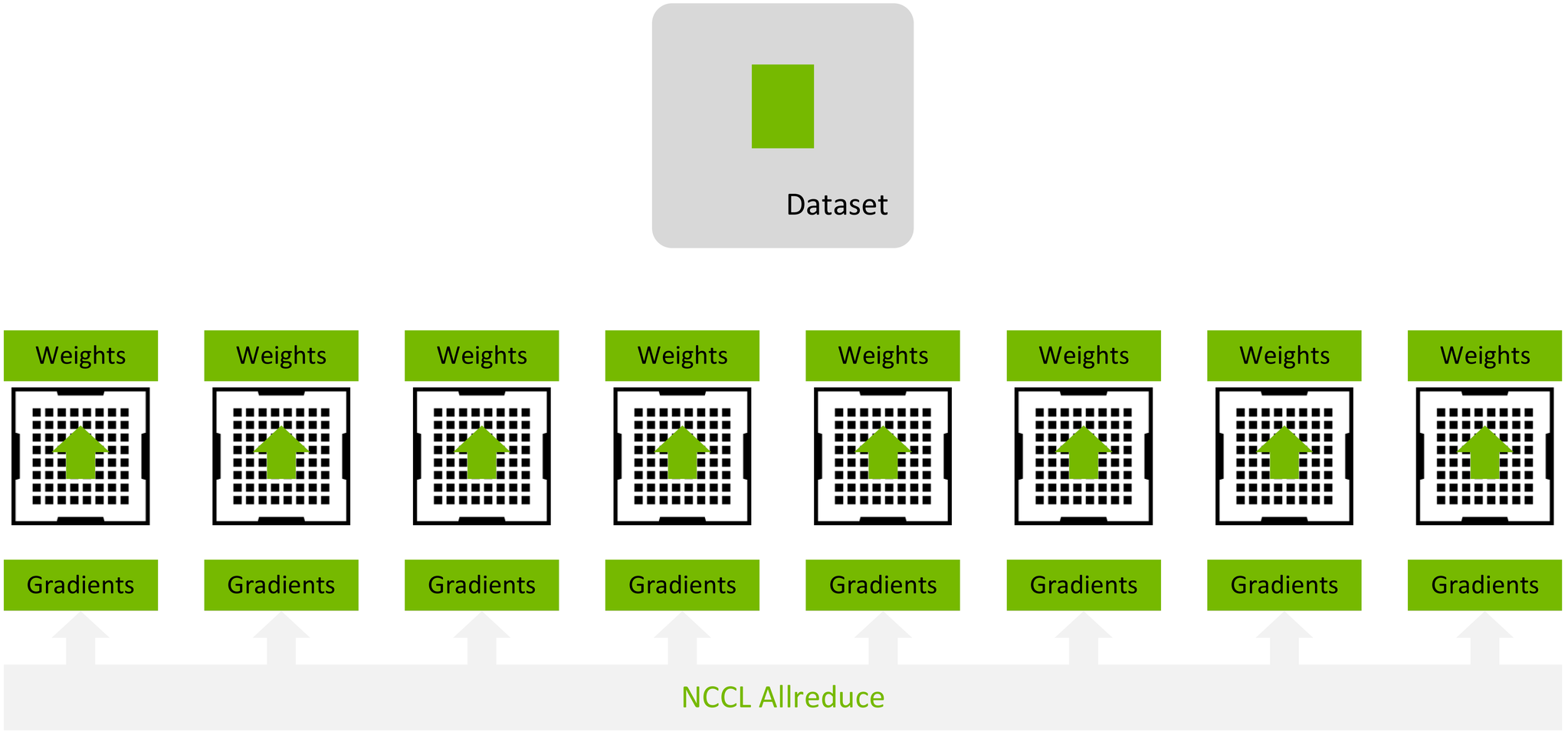
Spectrum-X Webpage

Quantum-X800 Webpage

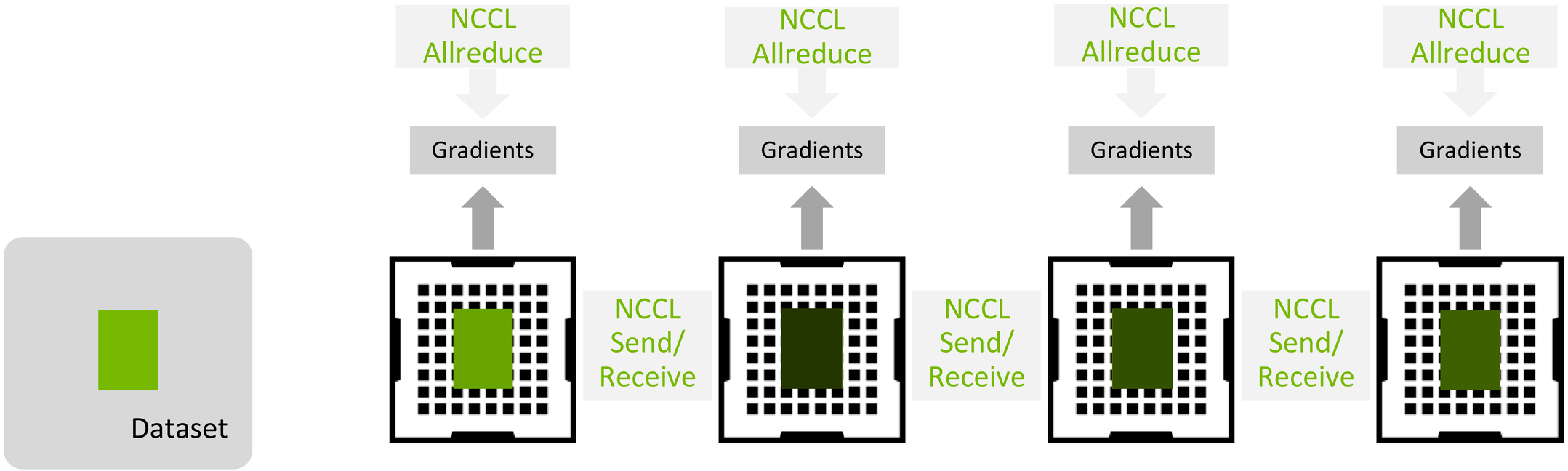

NVIDIA
Thank You!
Back-Up Material

Background on Parallelism with DL Models

Data Parallelism

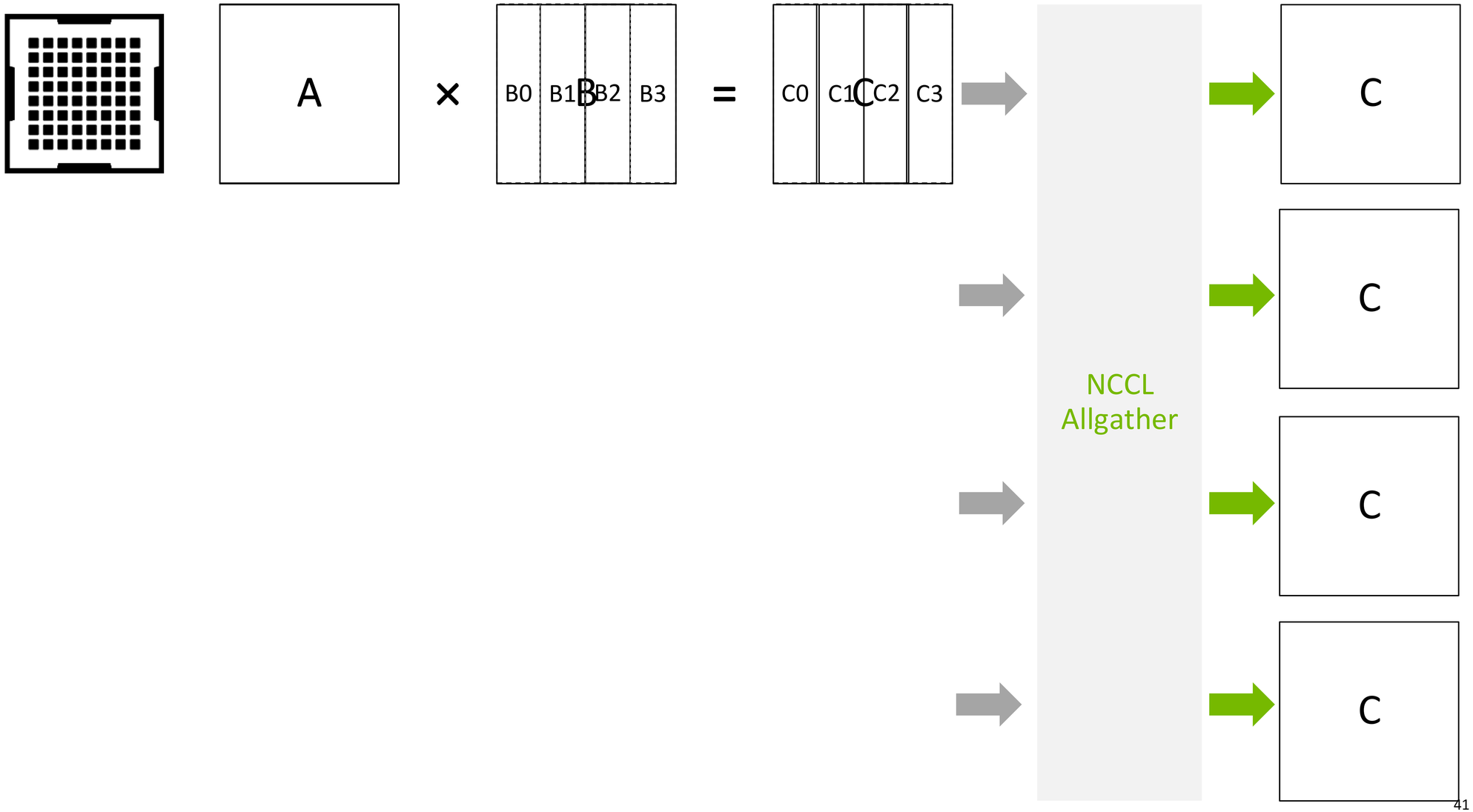


Pipeline Parallelism



Deep Learning Training

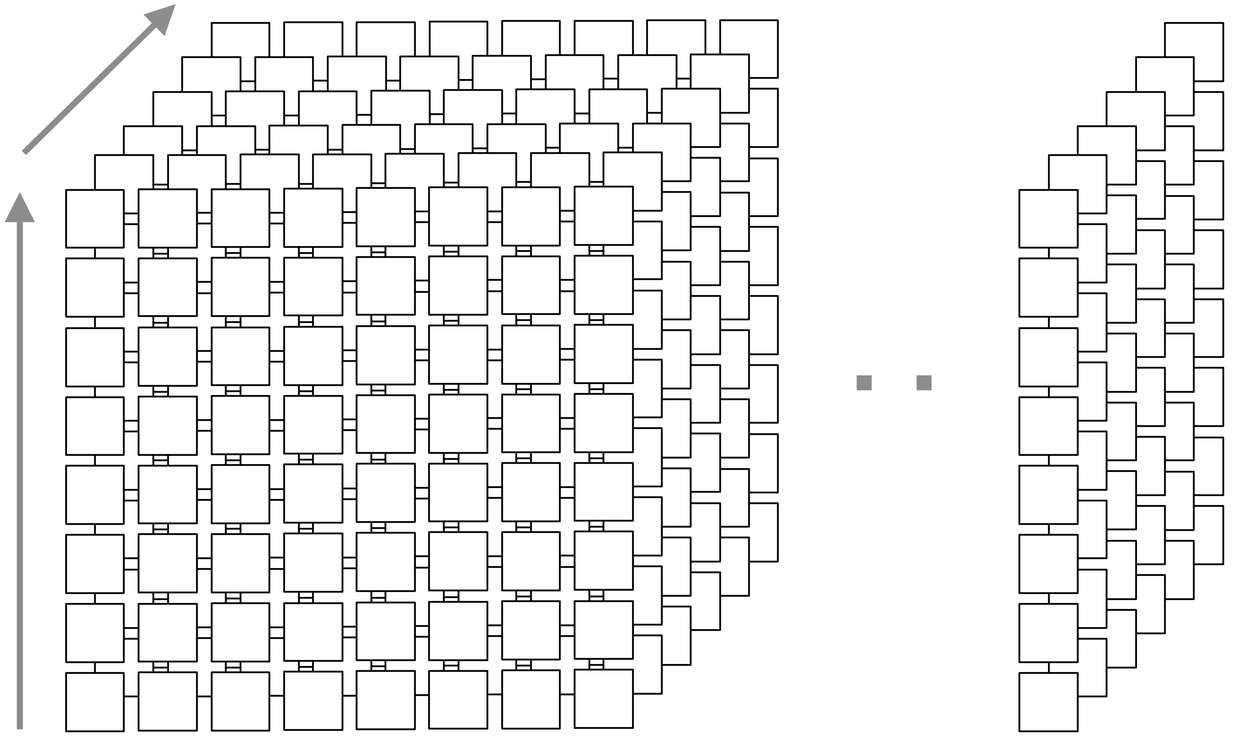
Tensor parallelism



Large scale LLM Training

Pipeline parallelism

ncclSend/
ncclRecv



Tensor parallelism

ncclAllGather

Data parallelism
ncclAllReduce

And also:

MoE (mixture of experts)
ncclSend/ncclRecv (alltoall)

FSDP (fully sharded data parallelism)
ncclAllGather

And other variations …