



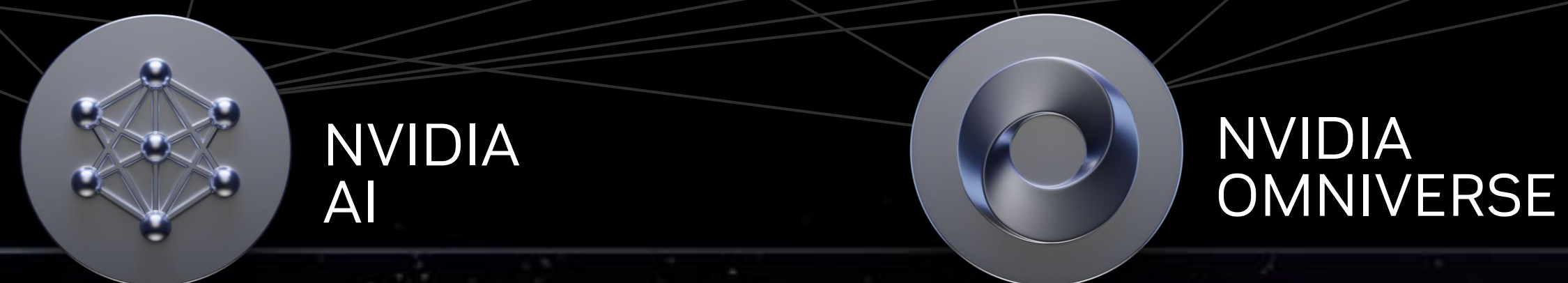
# Entering A New Frontier of AI Networking Innovation

Gilad Shainer | Teratec 2024

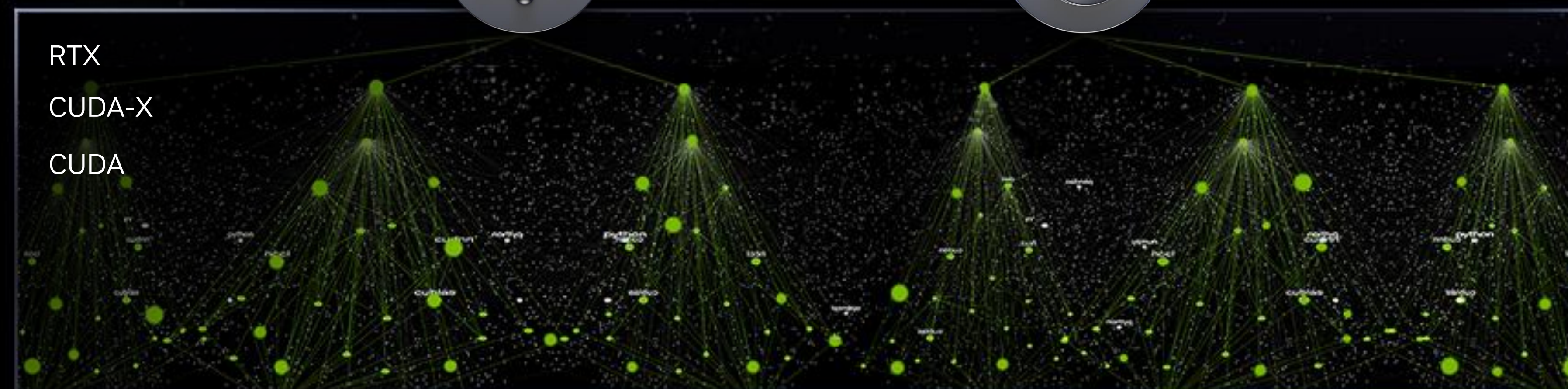
APPLICATION FRAMEWORKS



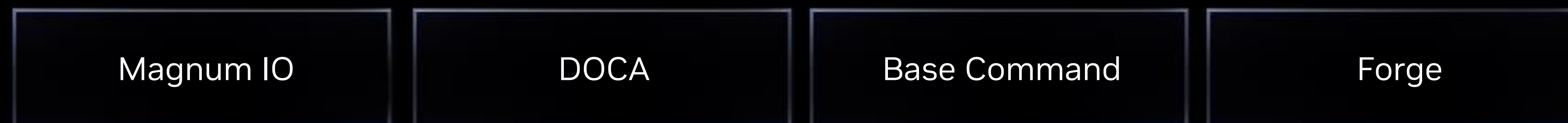
PLATFORM



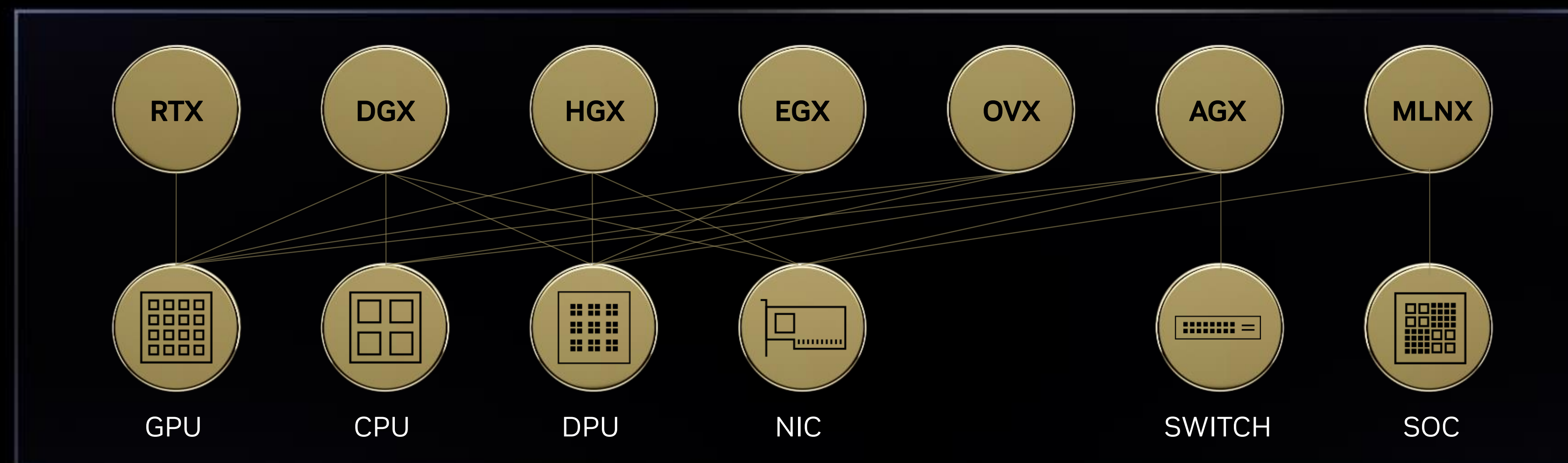
ACCELERATION LIBRARIES

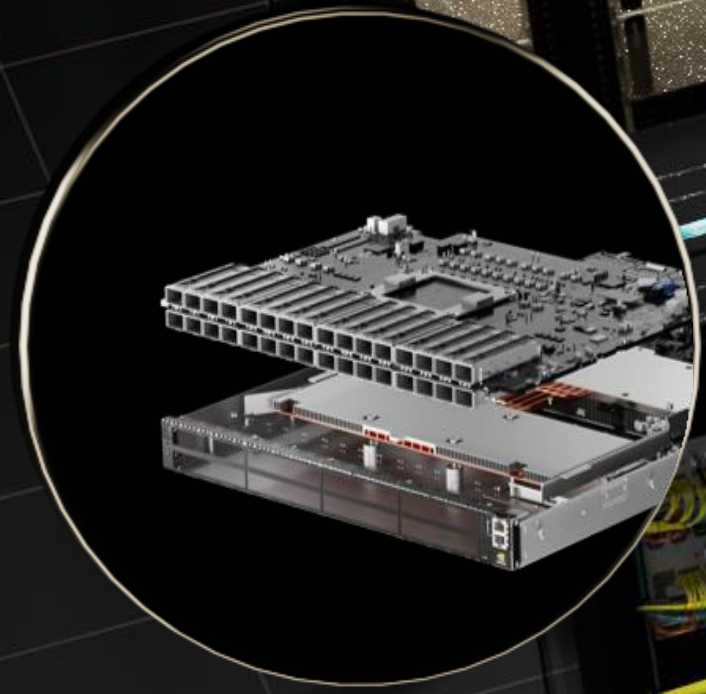


SYSTEM SOFTWARE

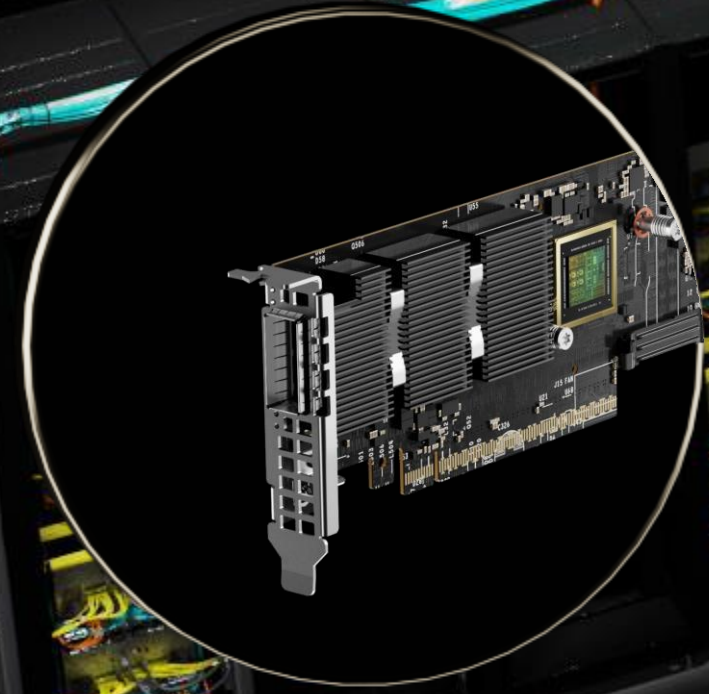


HARDWARE

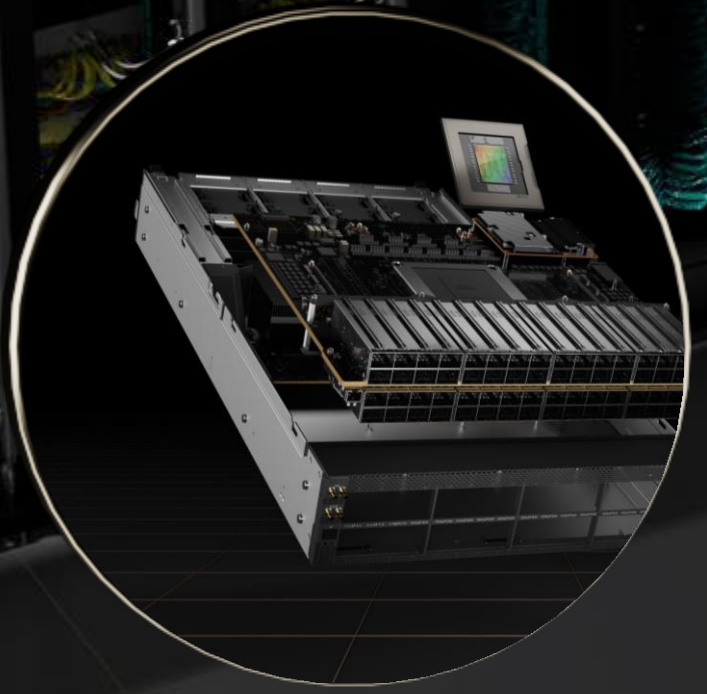




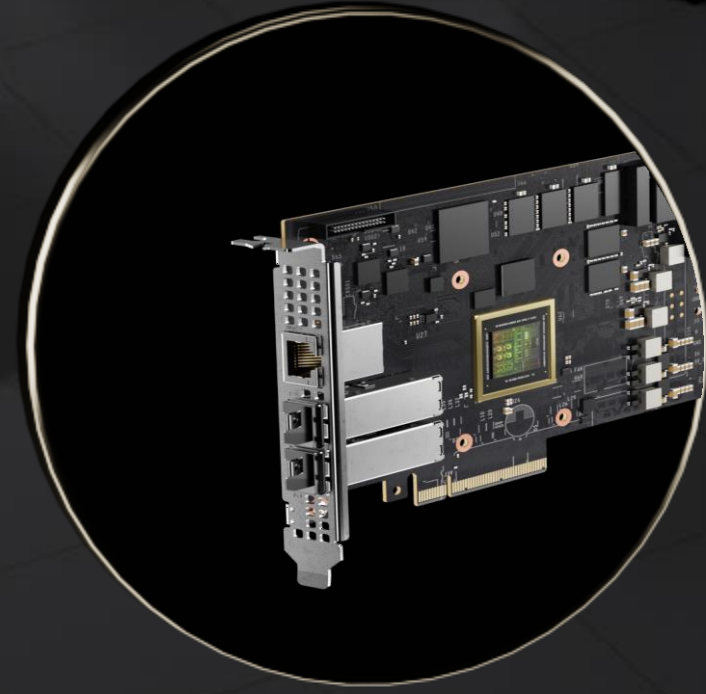
QUANTUM  
INFINIBAND SWITCH



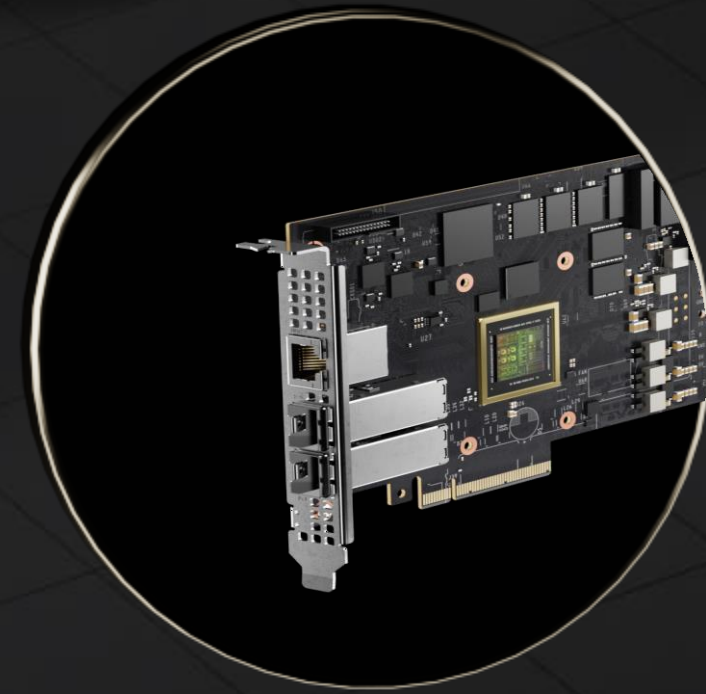
CONNECTX  
SuperNIC



SPECTRUM  
ETHERNET SWITCH



BLUEFIELD  
SuperNIC

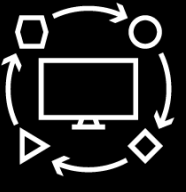

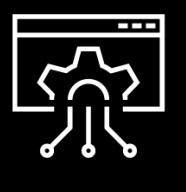
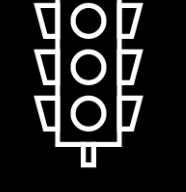


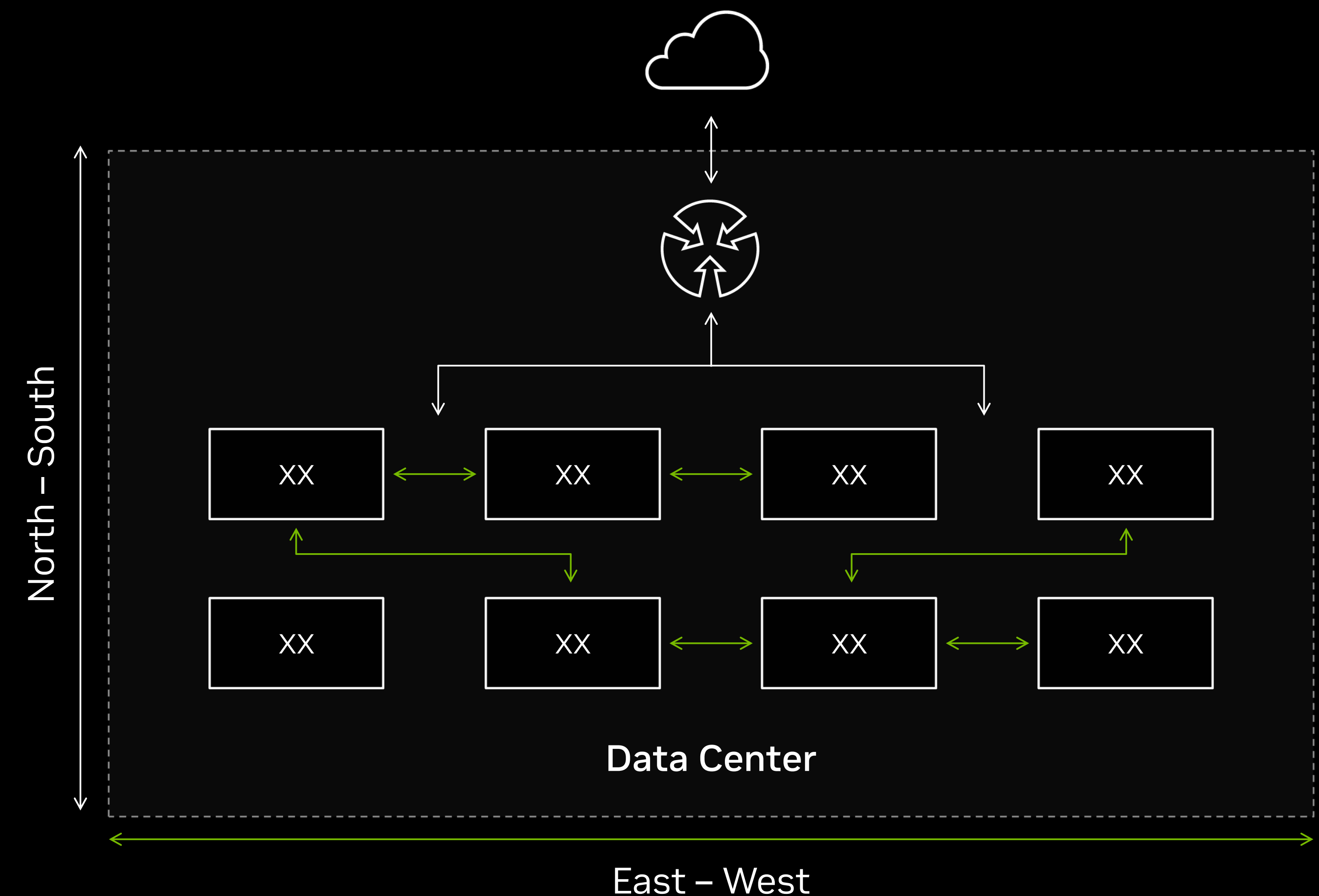
BLUEFIELD  
DPU



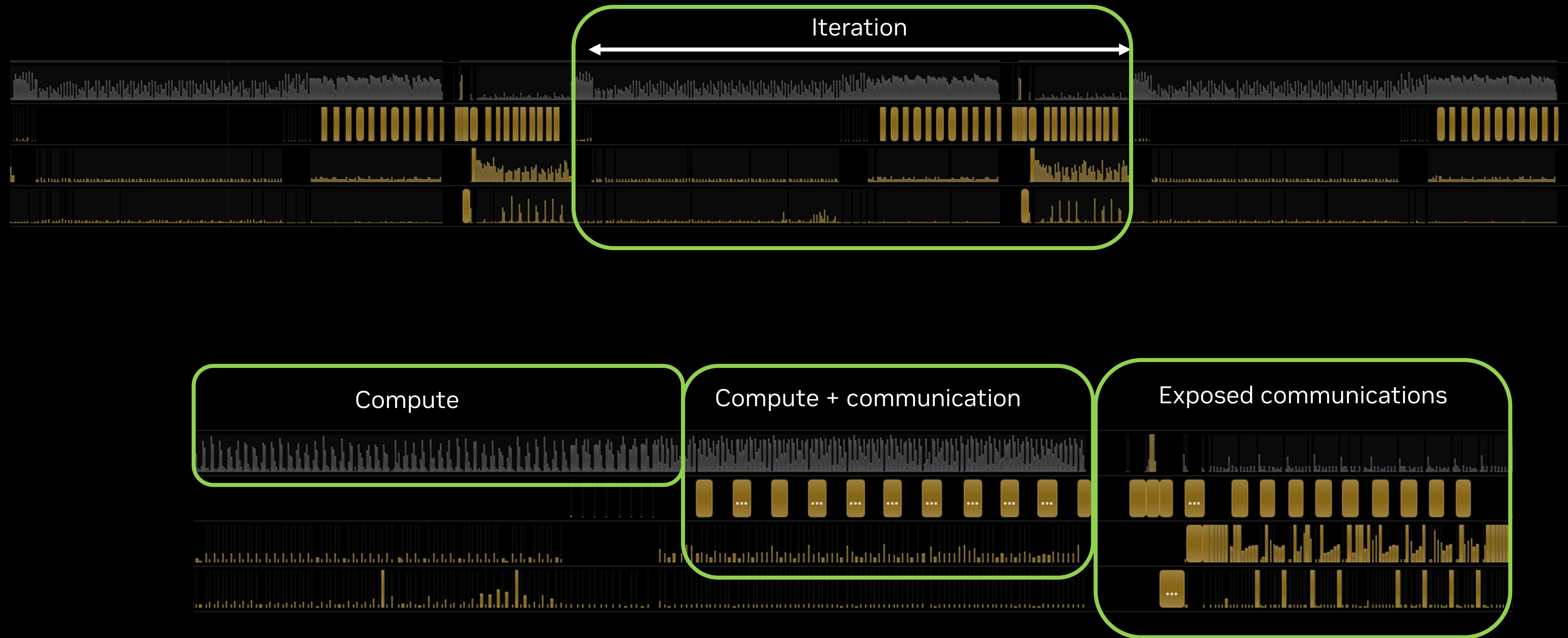
MANAGEMENT  
& TELEMETRY

# The Network Defines the Data Center

Control / User Access Network (North-South)		AI Fabric (East-West)
Loosely Coupled Applications		Distributed Tightly-Coupled Processing
TCP (Low Bandwidth Flows and Utilization)		RoCE (High Bandwidth Flows and Utilization)
High Jitter Tolerance		Low Jitter Tolerance (Long Tail Kills Performance)
Heterogeneous Traffic Average Multi-Pathing		Bursty Network Capacity Predictable Performance



# LLM Compute and Communication Profiling

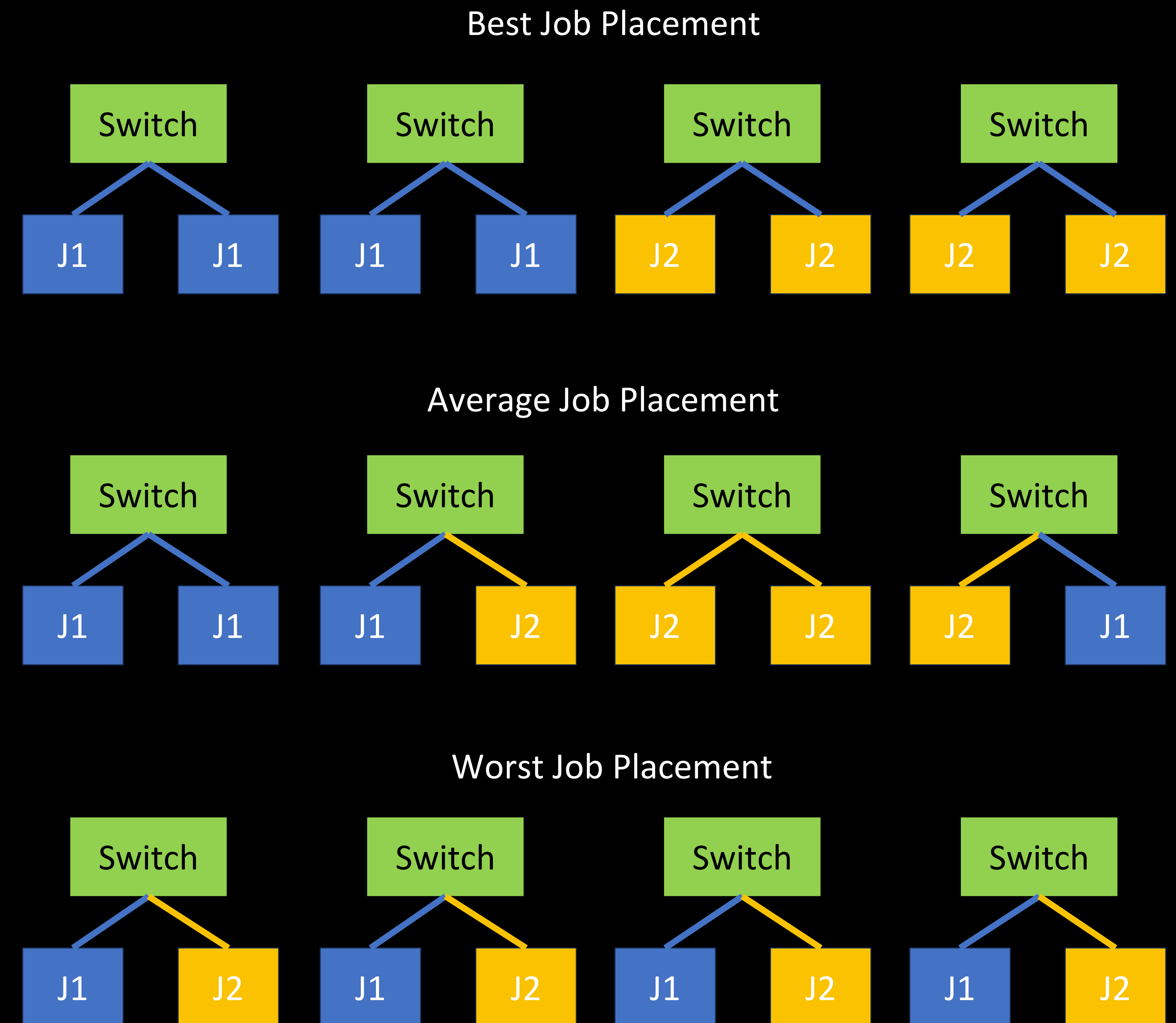
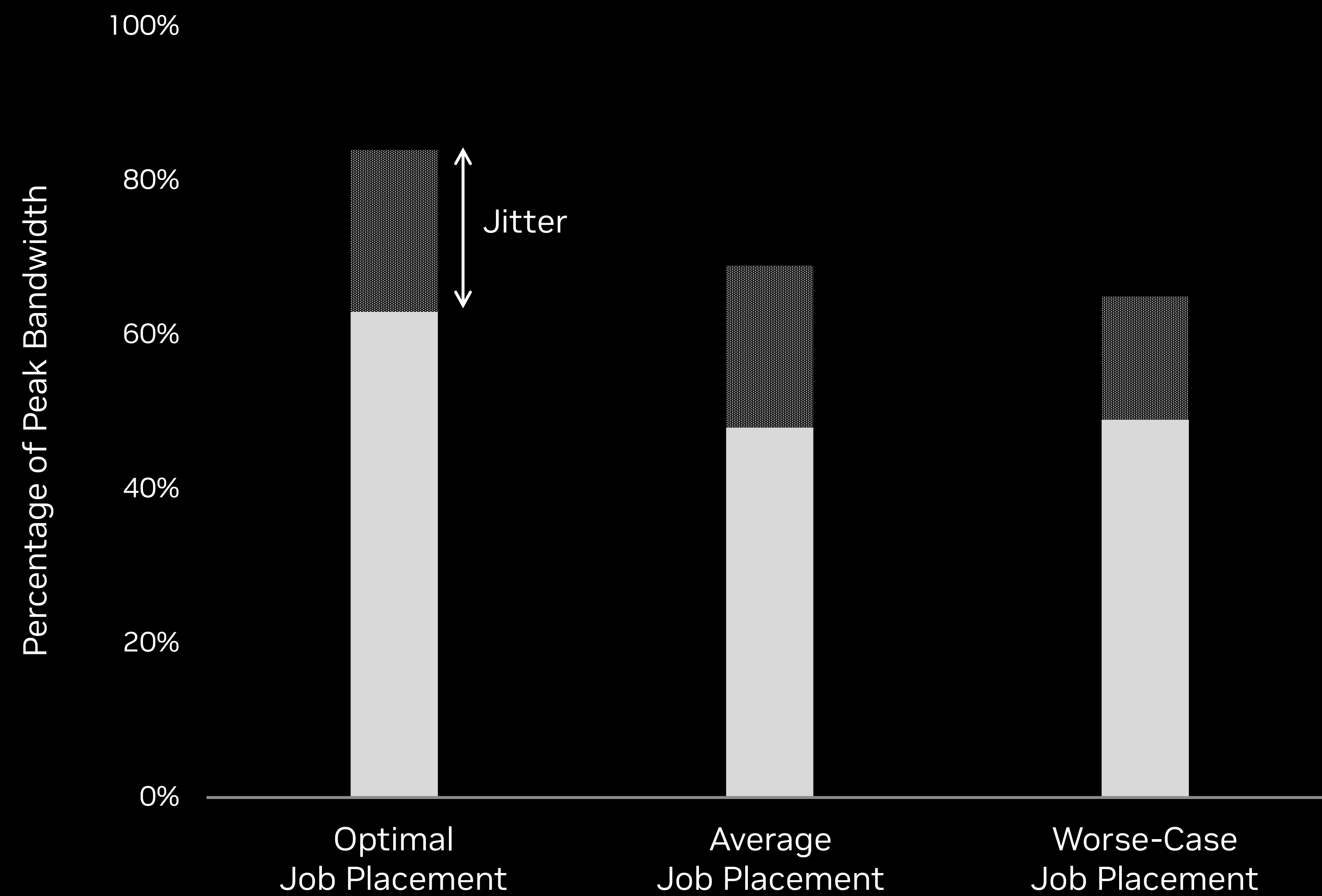


Representative profile from a large scale LLM training run

Communications is bursty in nature, an average bandwidth utilization is not a good network criteria

# The Network Defines the Data Center

## LLM NCCL AllReduce — Traditional Ethernet

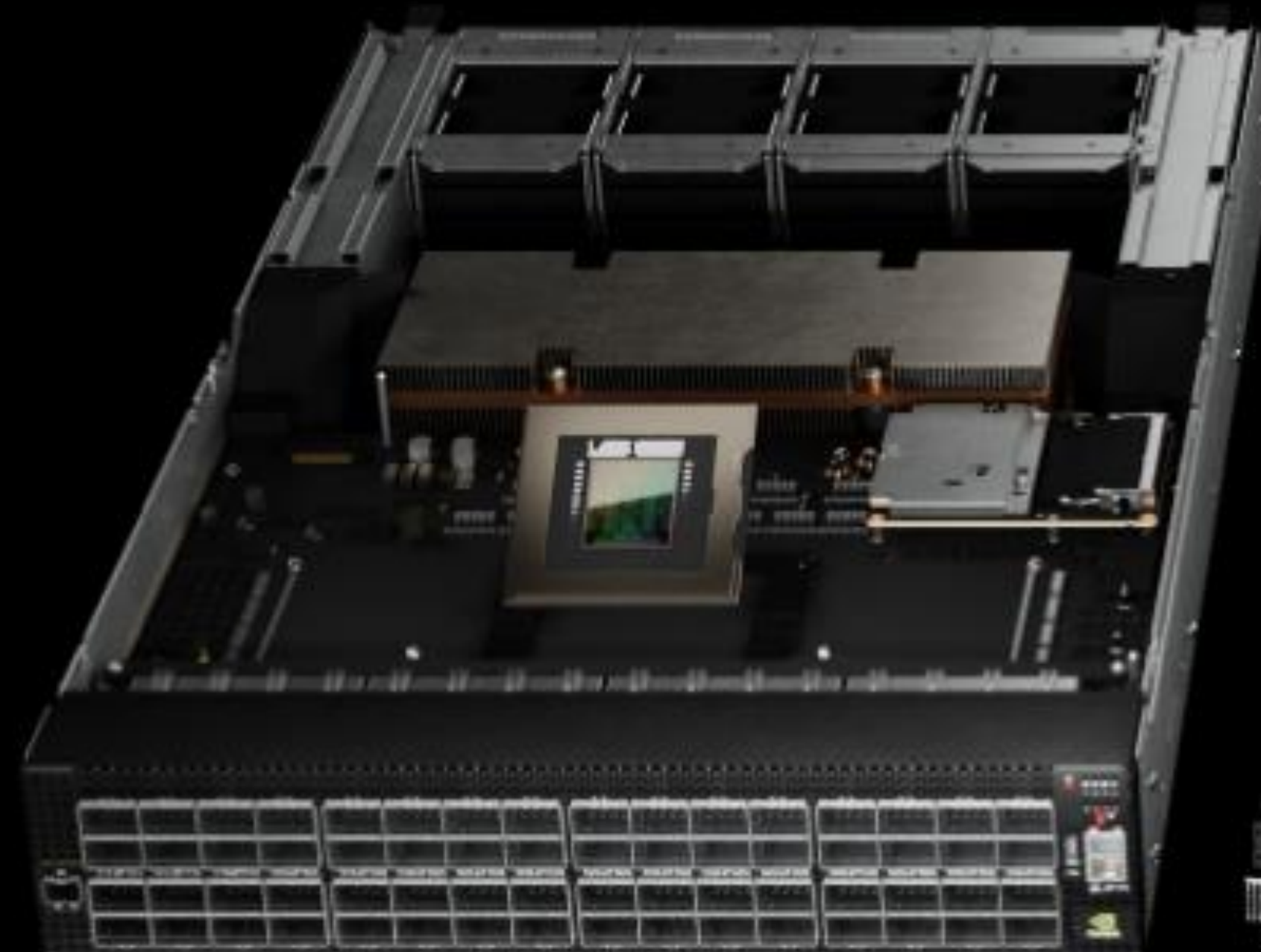


NCCL (NVIDIA Collective Communication Library) is the SDK library for AI communications - connects the GPUs and the network for the AI network operations

# Spectrum-X800 Brings High-Performance AI to Ethernet

AI-optimized networking for every data center

- RoCE Adaptive Routing (local and remote information, a packet granularity)
- Congestion Control (telemetry probes)
- Noise Isolation (multi-jobs or a single large-scale job)
- High frequency telemetry (1000x)



Spectrum-X800



BlueField-3 SuperNIC

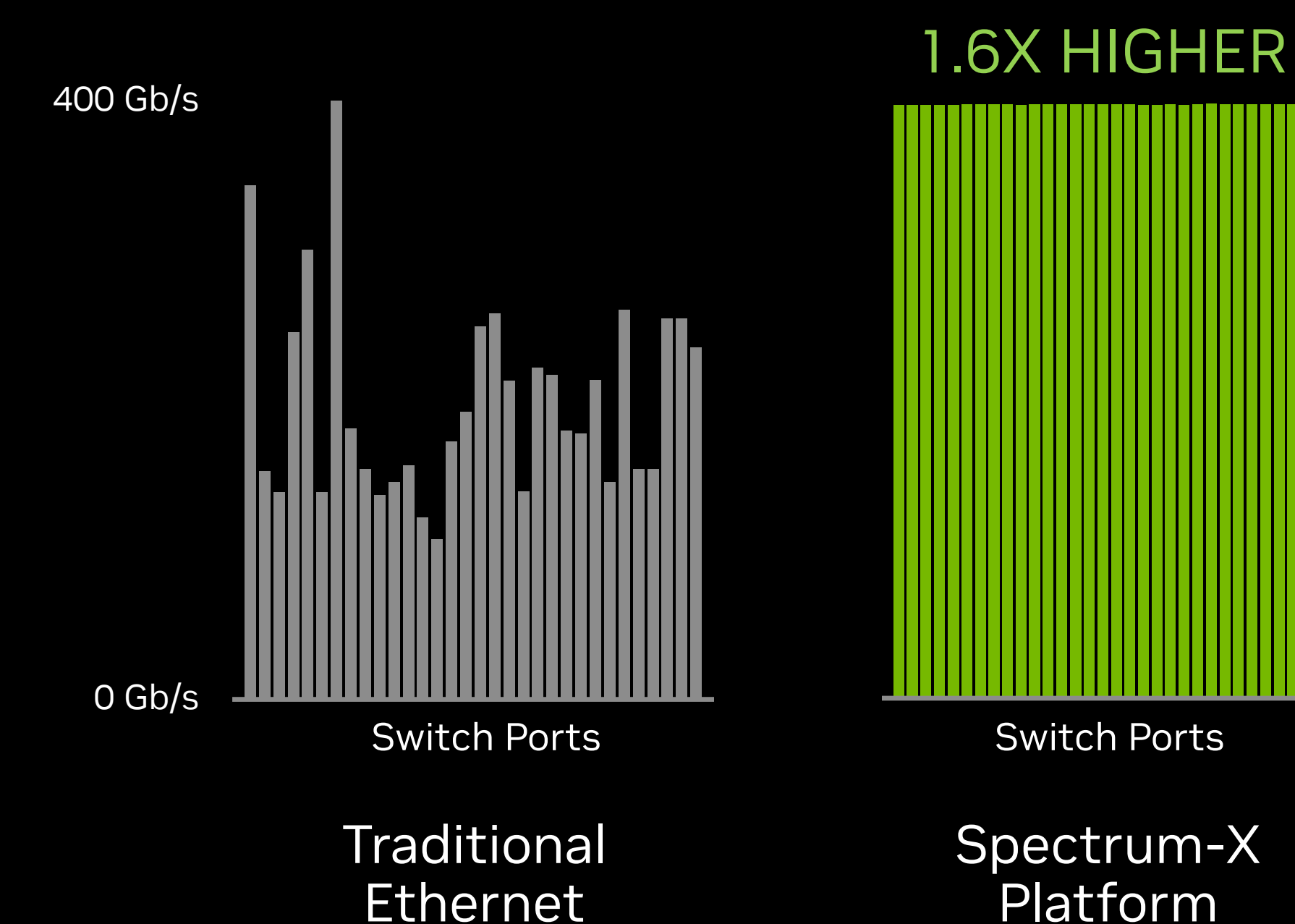
## Spectrum-X800 Switch

- 51.2T bandwidth
- 64 X 800G Ports, 128 x 400G
- Adaptive routing
- Congestion control, noise isolation

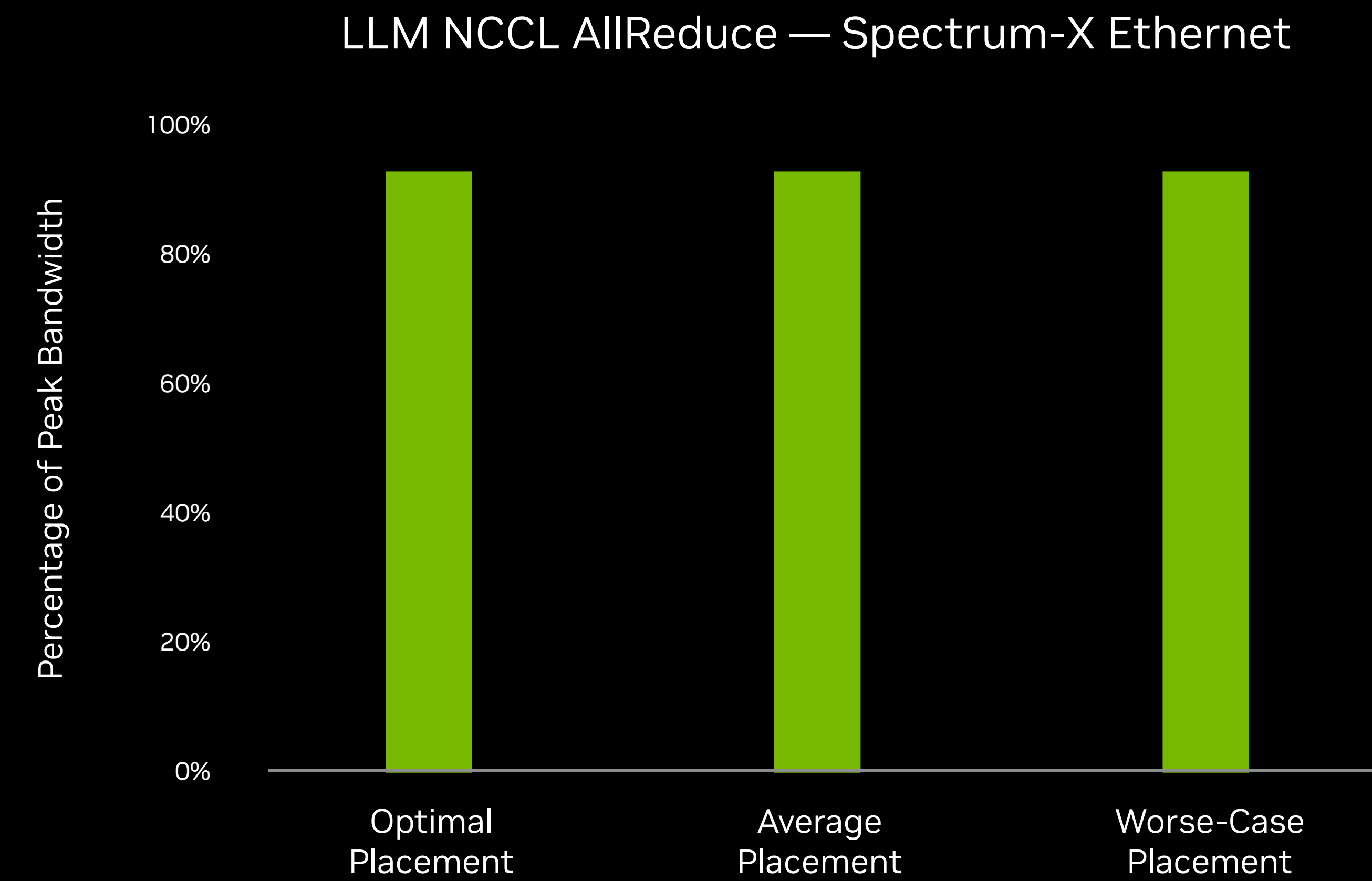
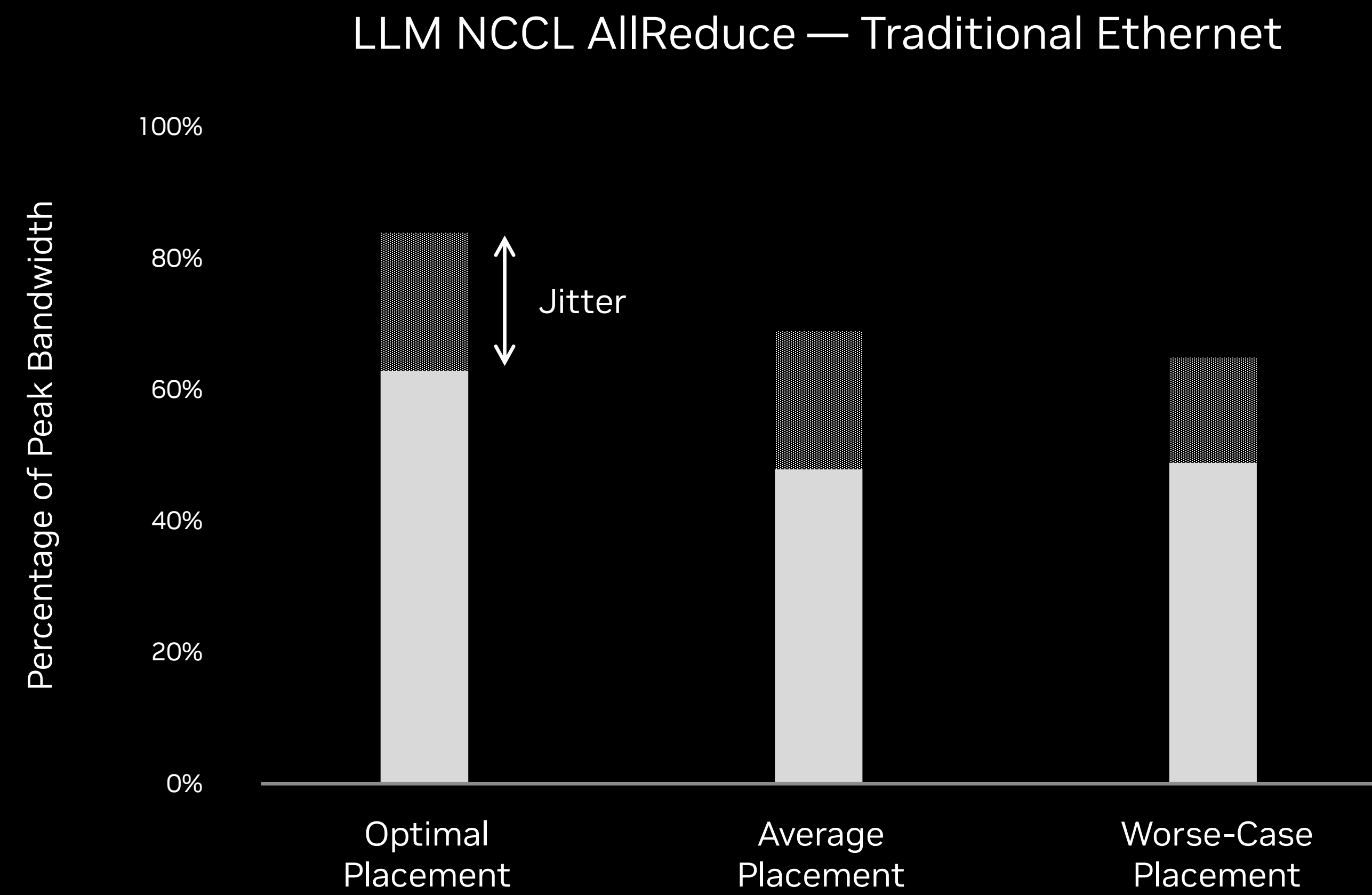
## BlueField-3 SuperNIC

- 16 Arm 64-Bit Cores
- 16 Core / 256 Threads Datapath Accelerator
- DDR memory interface
- ConnectX NIC
- PCIe switch

## Effective Network Bandwidth With and Without Adaptive Routing



# Spectrum-X New Class of Ethernet for AI



- Spectrum-X performance is consistent; Traditional Ethernet shows run-to-run bandwidth variability
- Results in 1.4x higher LLM performance (2K GPUs)



# NVIDIA Spectrum-X Generative AI Cloud

Most powerful supercomputer in Israel



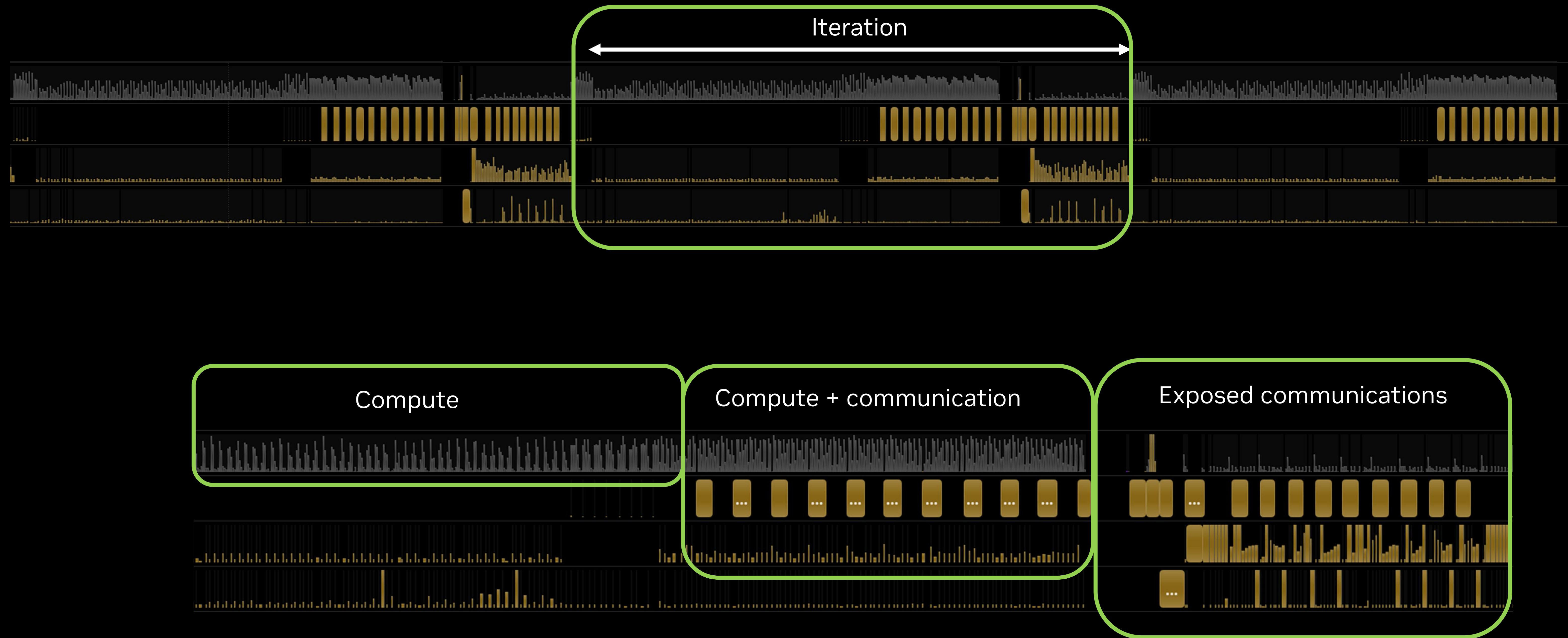
2,048 GPUs

2,560 BlueField-3 SuperNICs

80+ Spectrum-x800 Ethernet switches

Peak AI performance of 8-Exaflops

# LLM Compute and Communication Profiling



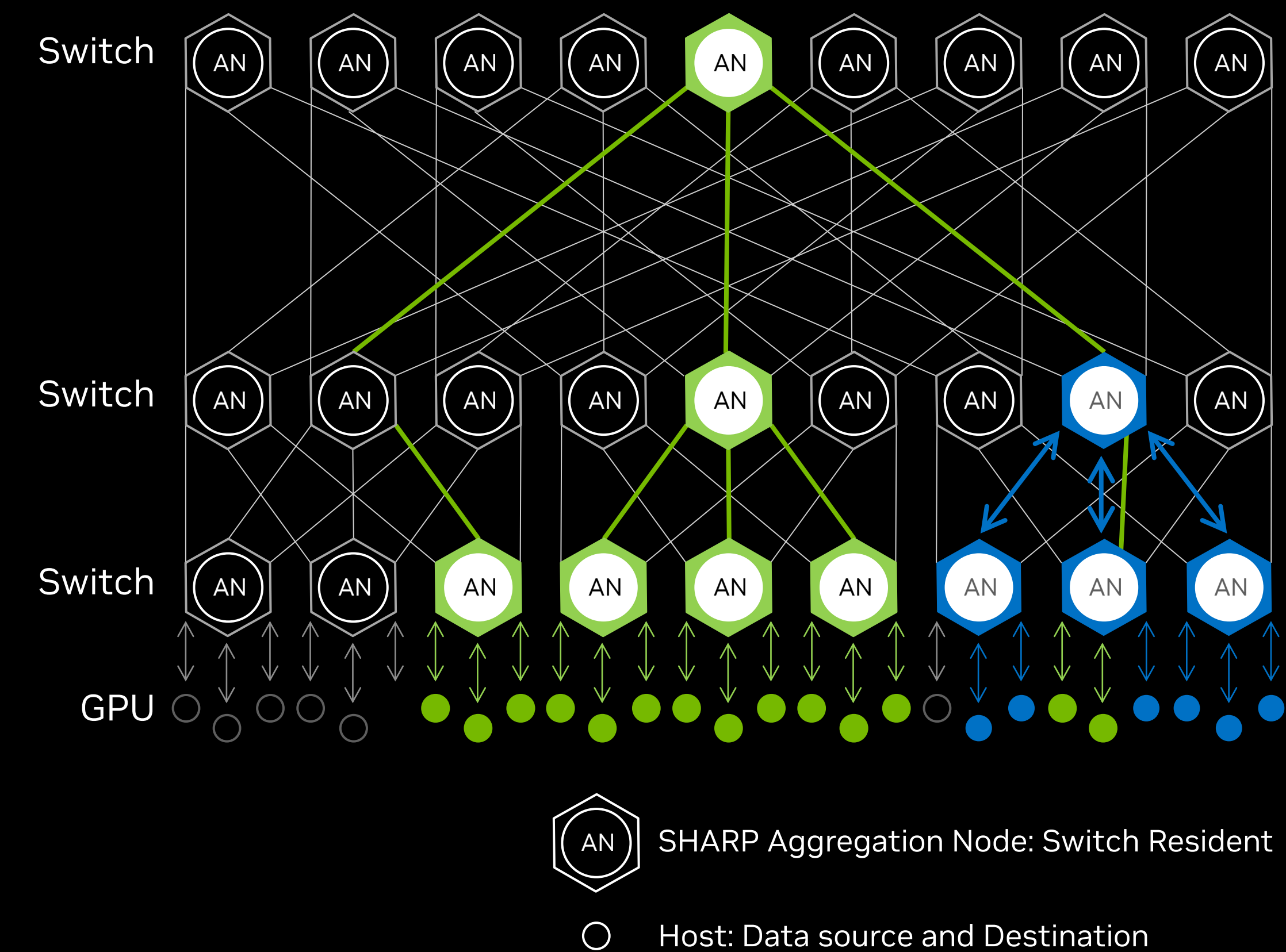
Representative profile from a large scale LLM training run

Communications is bursty in nature, an average bandwidth utilization is not a good network criteria

# NVIDIA SHARP

## Scalable Hierarchical Aggregation and Reduction Protocol Technology

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



# 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

### Quantum-X800 switch

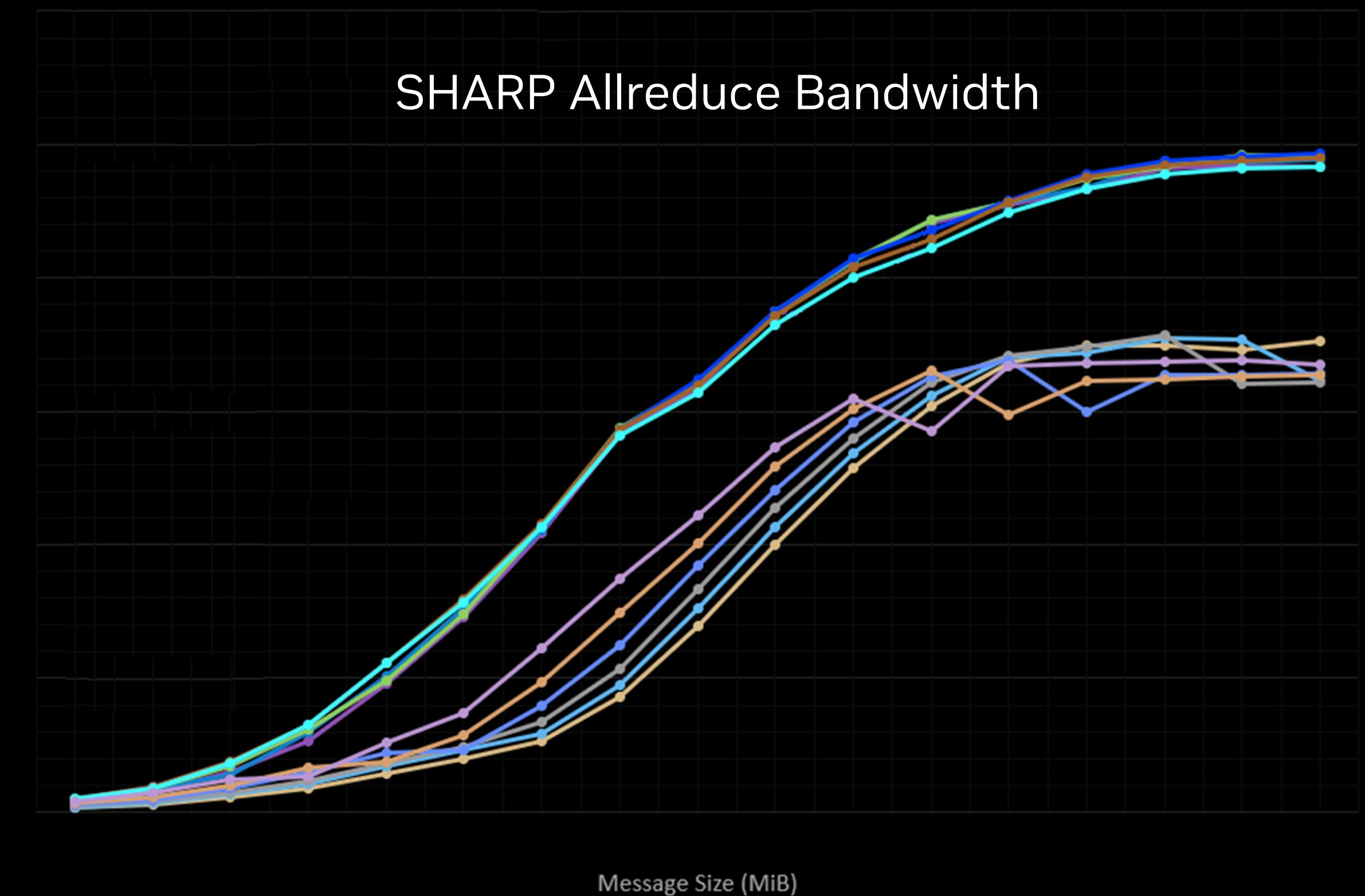
- 144X 800G ports,
- SHARPV4 In-Network Computing
- Adaptive routing
- congestion control, and noise isolation



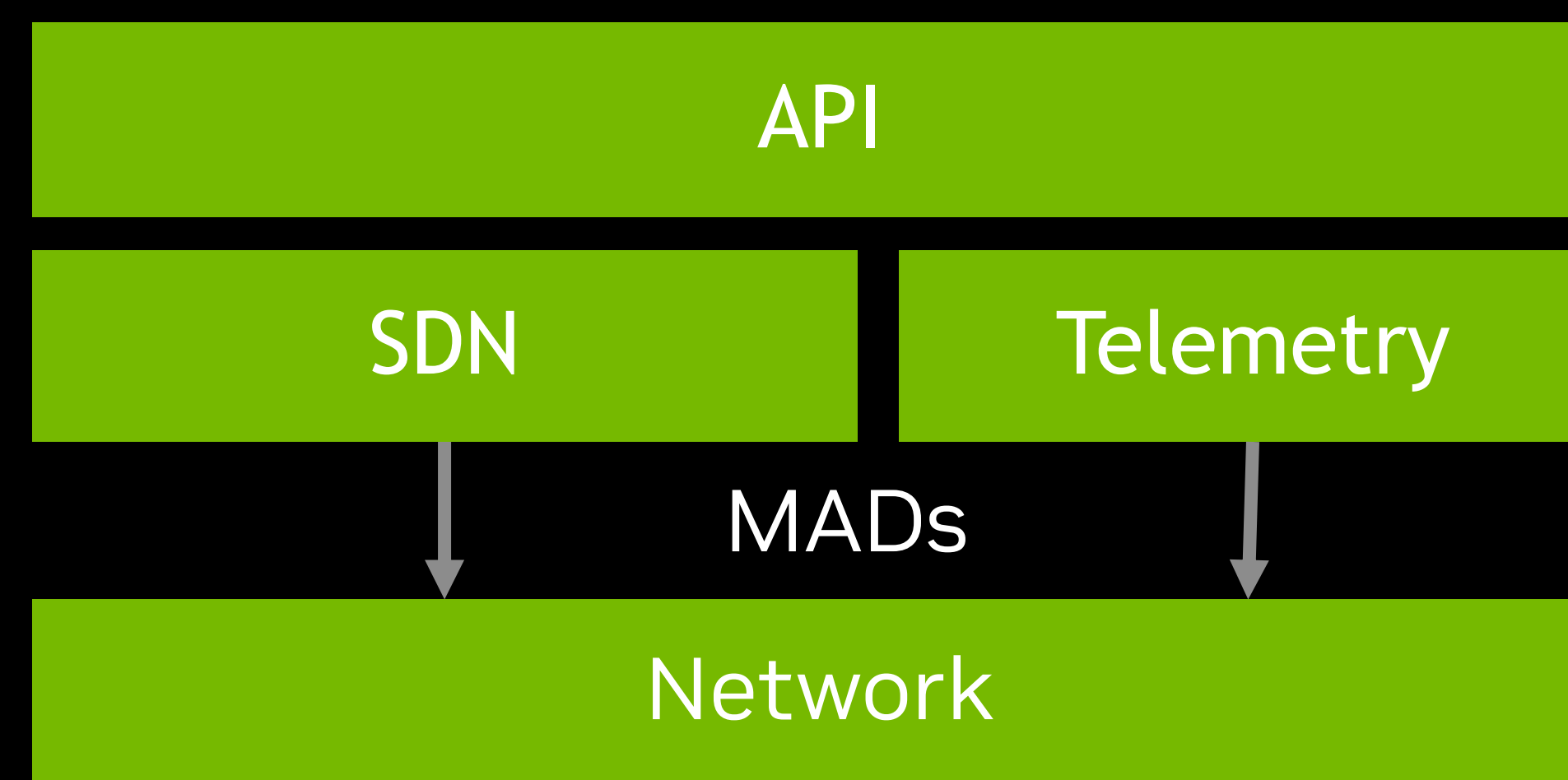
### ConnectX-x800 InfiniBand SuperNIC

- PCIe Gen 6, PCIe switch
- Multi-host

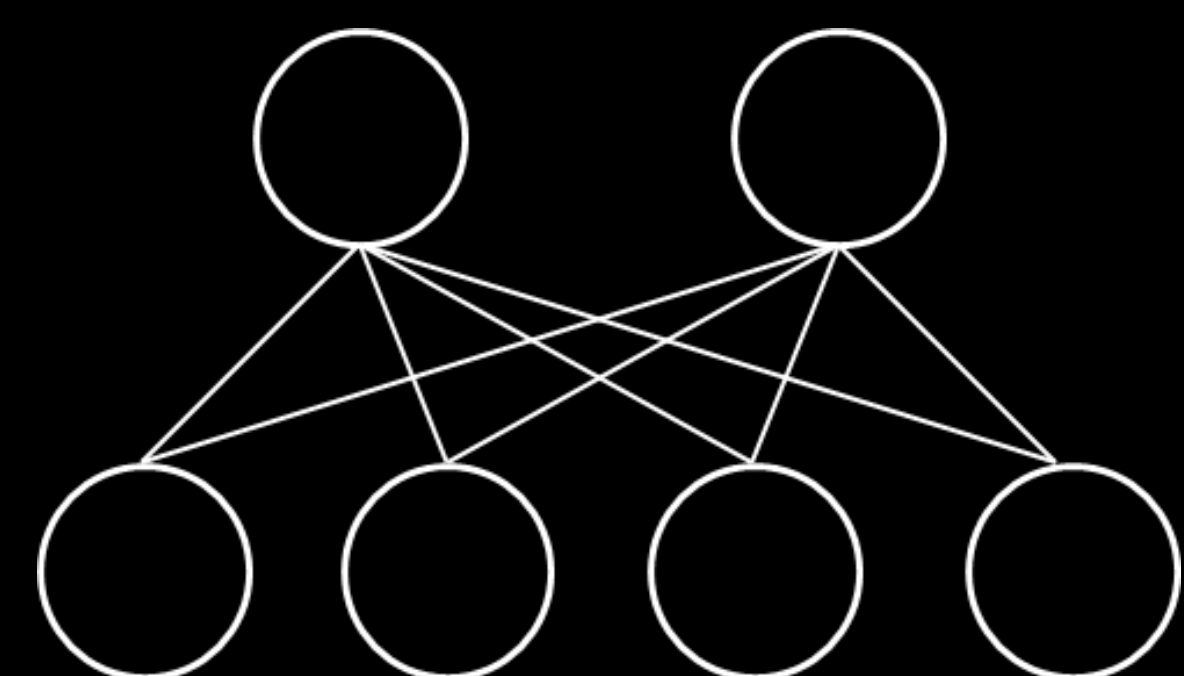
ConnectX-800 SuperNIC



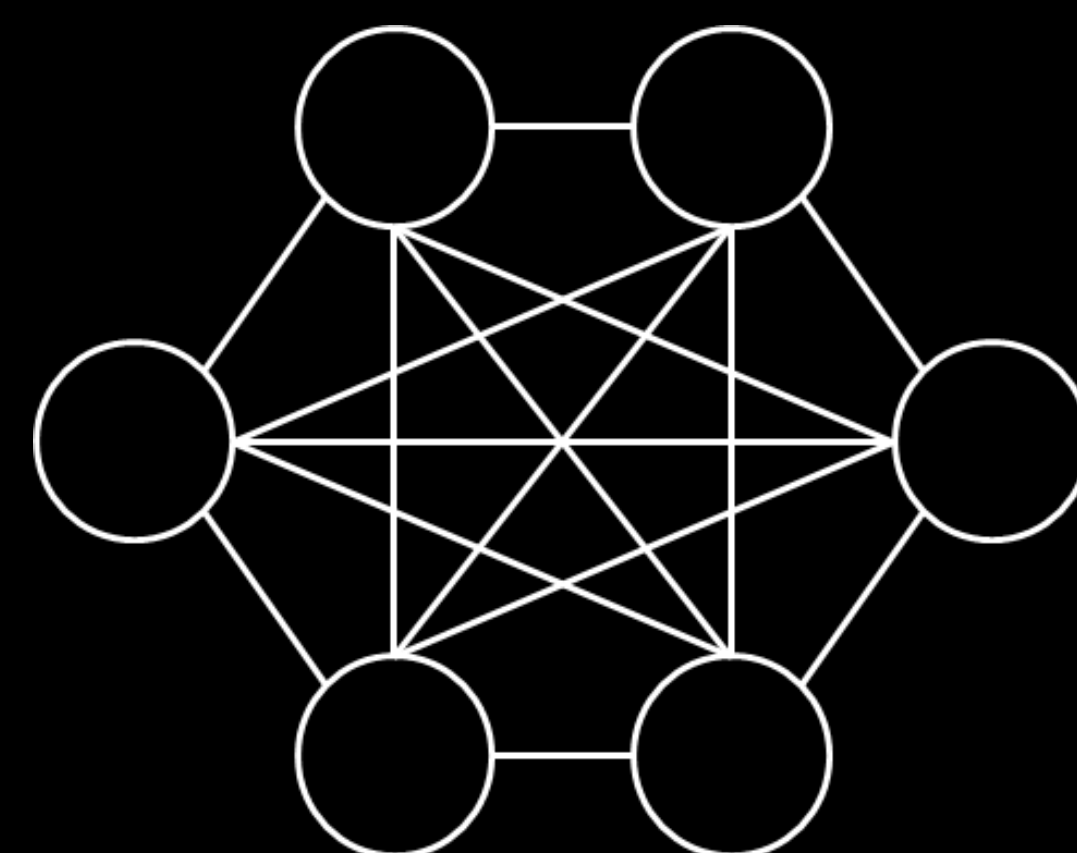
# Exploring Topologies – Creating Routing Algorithms



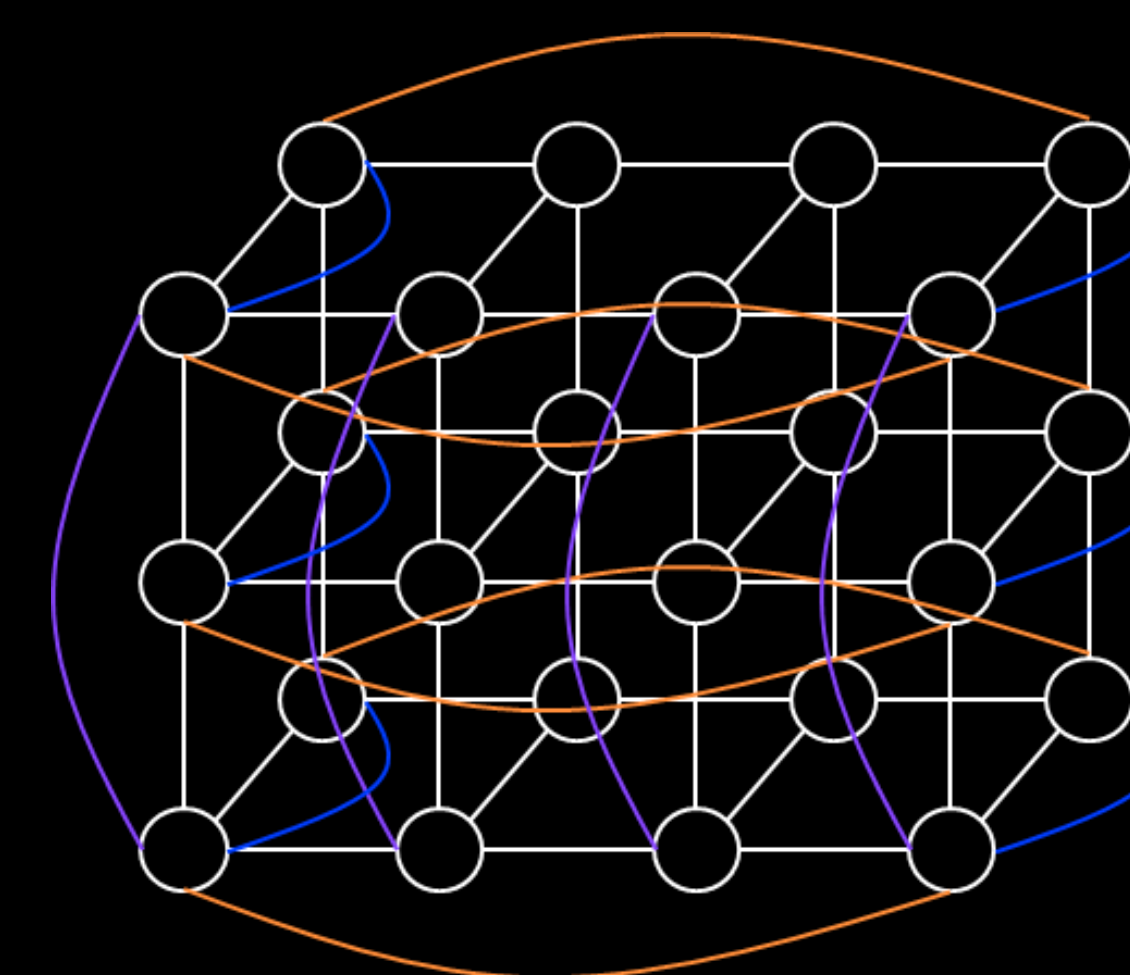
Existing routing algorithms support



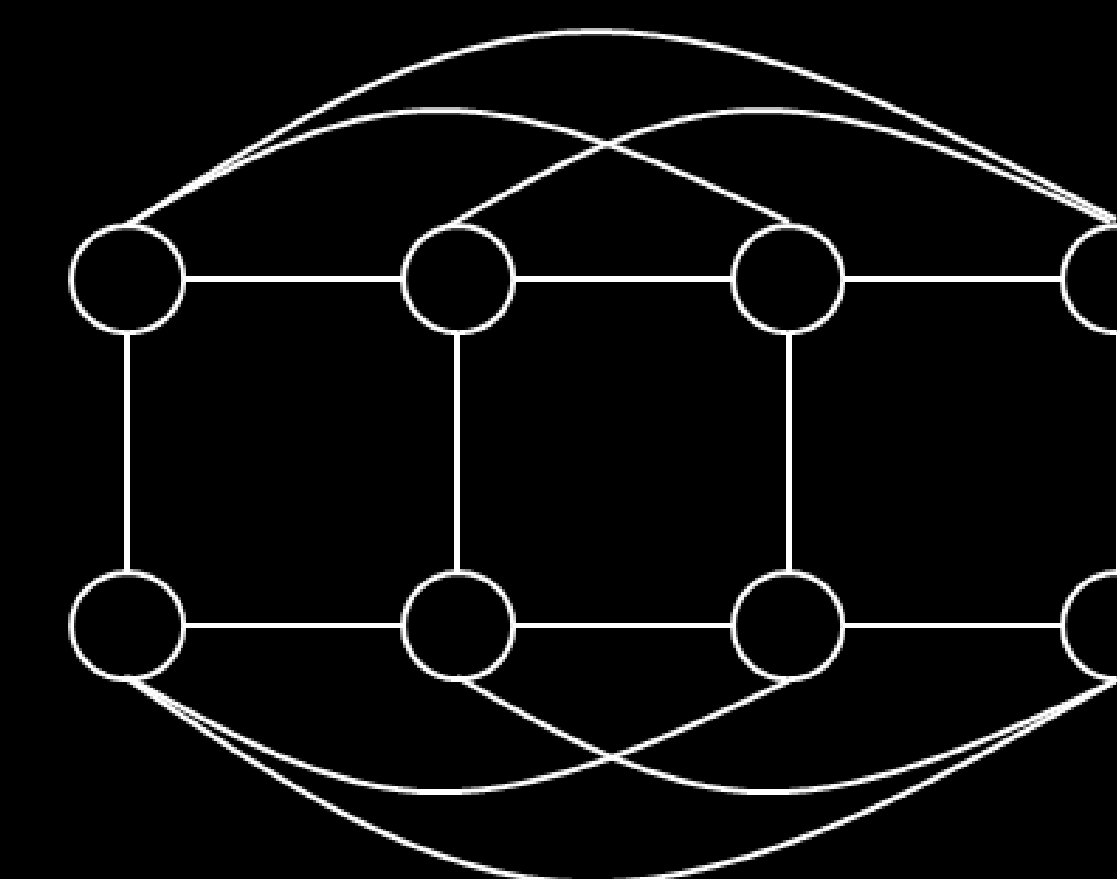
Fat-tree



Dragonfly



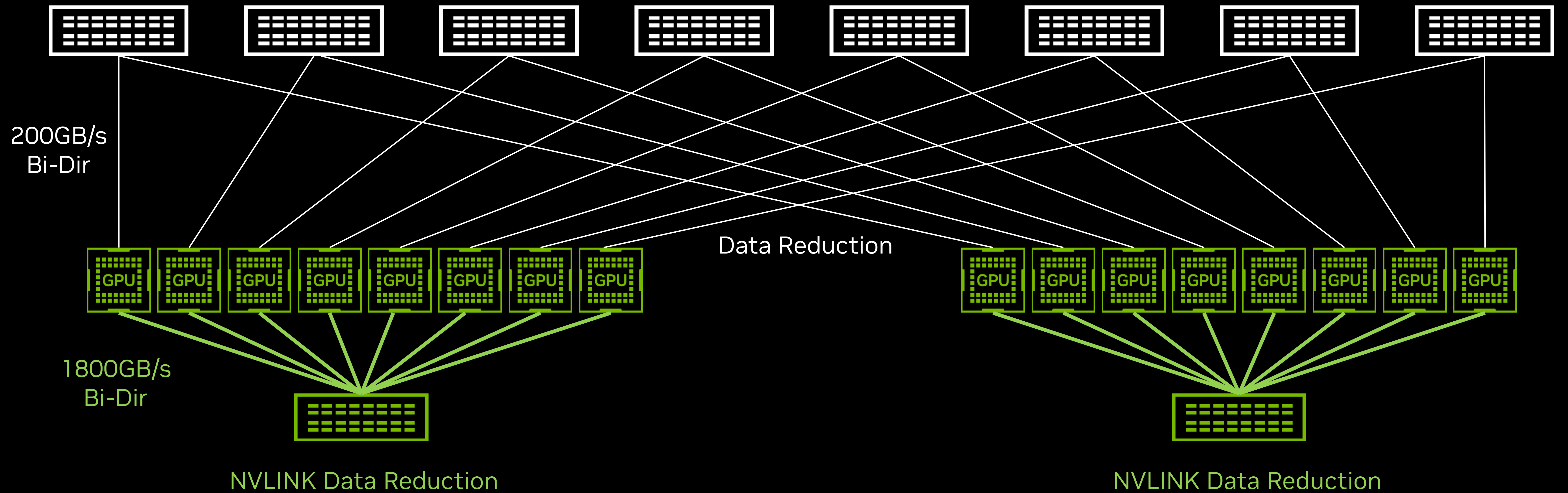
Torus



HyperX

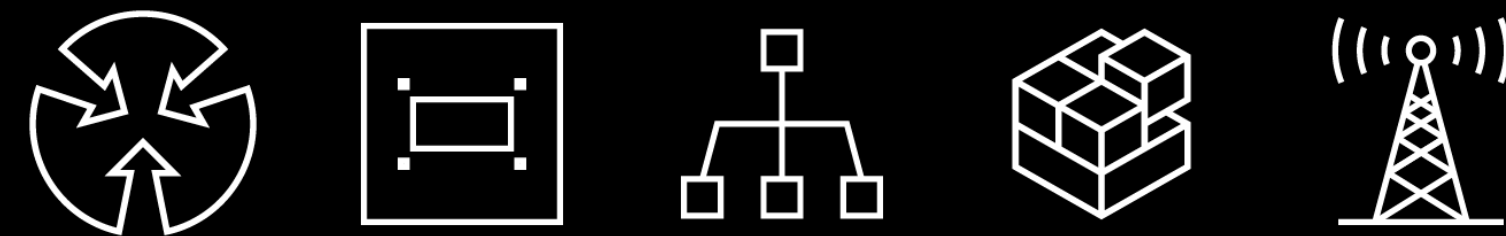
# Rail Optimized Topology

Quantum, Spectrum



# BlueField Data Processing Unit

SOFTWARE DEFINED NETWORKING



SOFTWARE DEFINED SECURITY



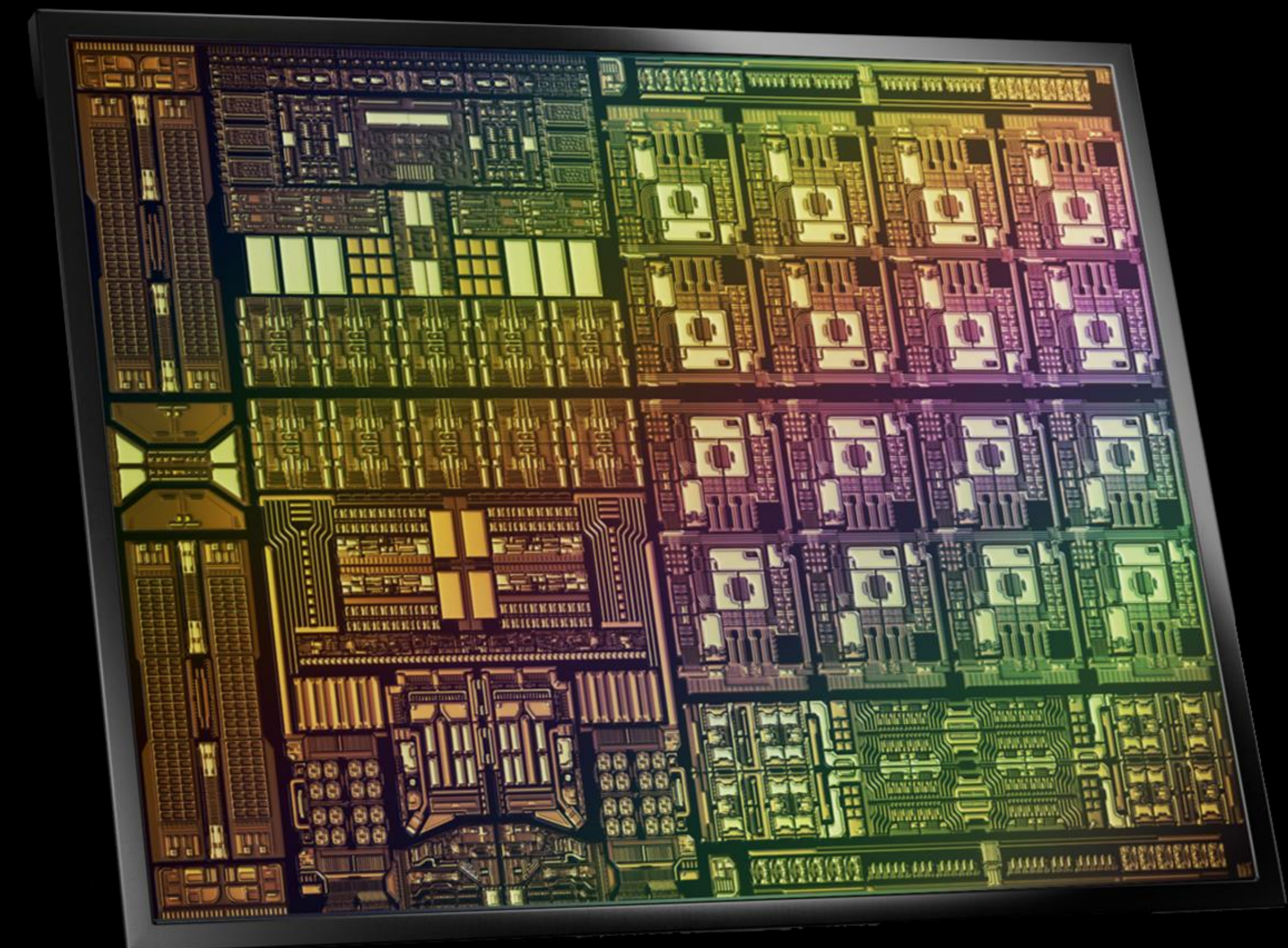
SOFTWARE DEFINED STORAGE



Infrastructure Services



BlueField Infrastructure Compute Platform



## Data Center on a Chip

16 Arm 64-Bit Cores

16 Core / 256 Threads Datapath Accelerator

ConnectX InfiniBand / Ethernet

DDR memory interface

PCIe switch

# Octopus Performance – MPI Time

32 nodes, 100 time steps, Average total time per process in seconds

Function	Host	DPU Offloaded
Application	637.28	455.49
All Communications	245.99	138.15
MPI_Allgatherv	126.38	82.74
MPI_Waitall	57.60	32.02
MPI_Alltoall	37.89	13.87
MPI_Allreduce	7.44	4.96
Other	16.68	4.56

