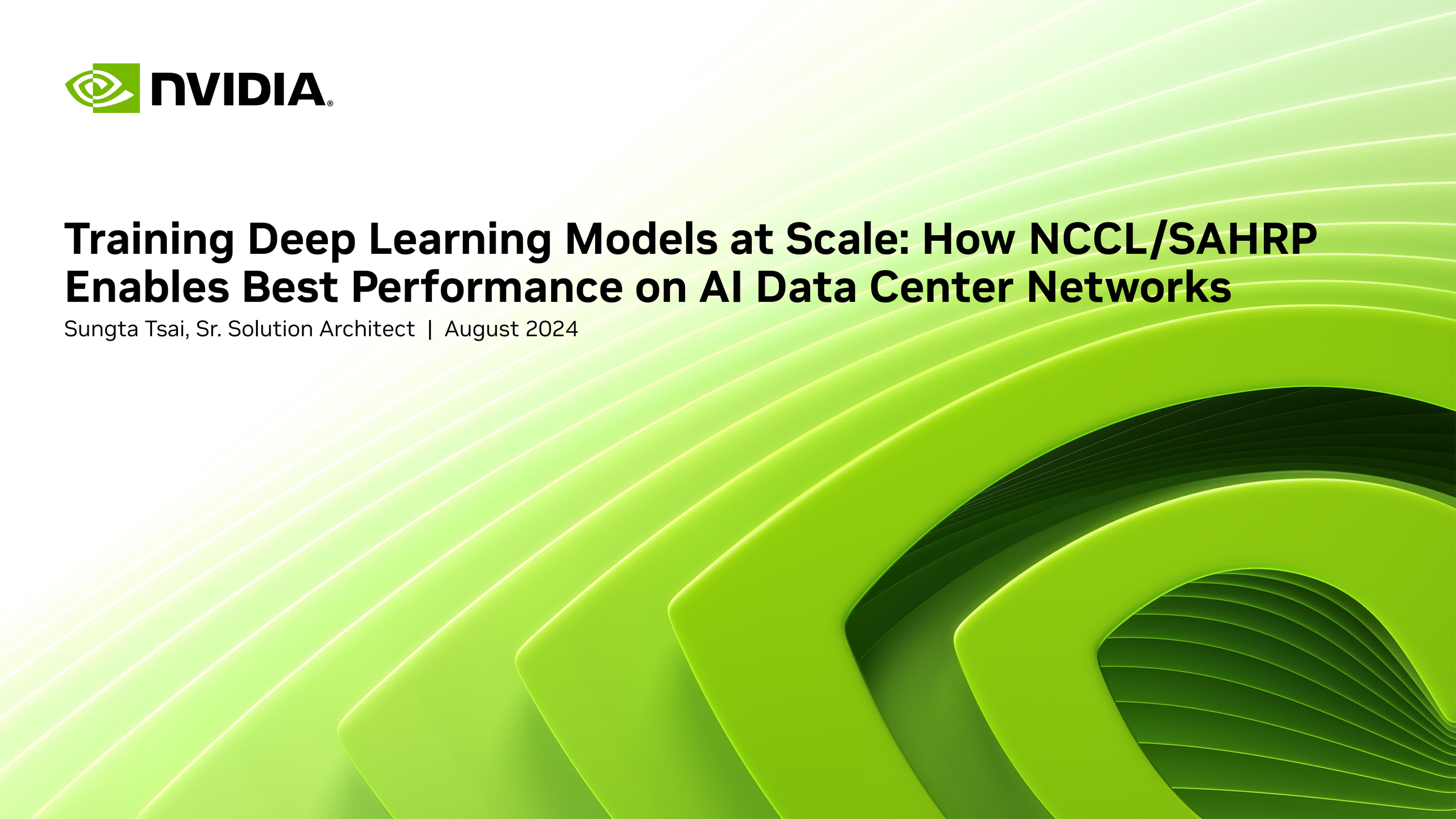
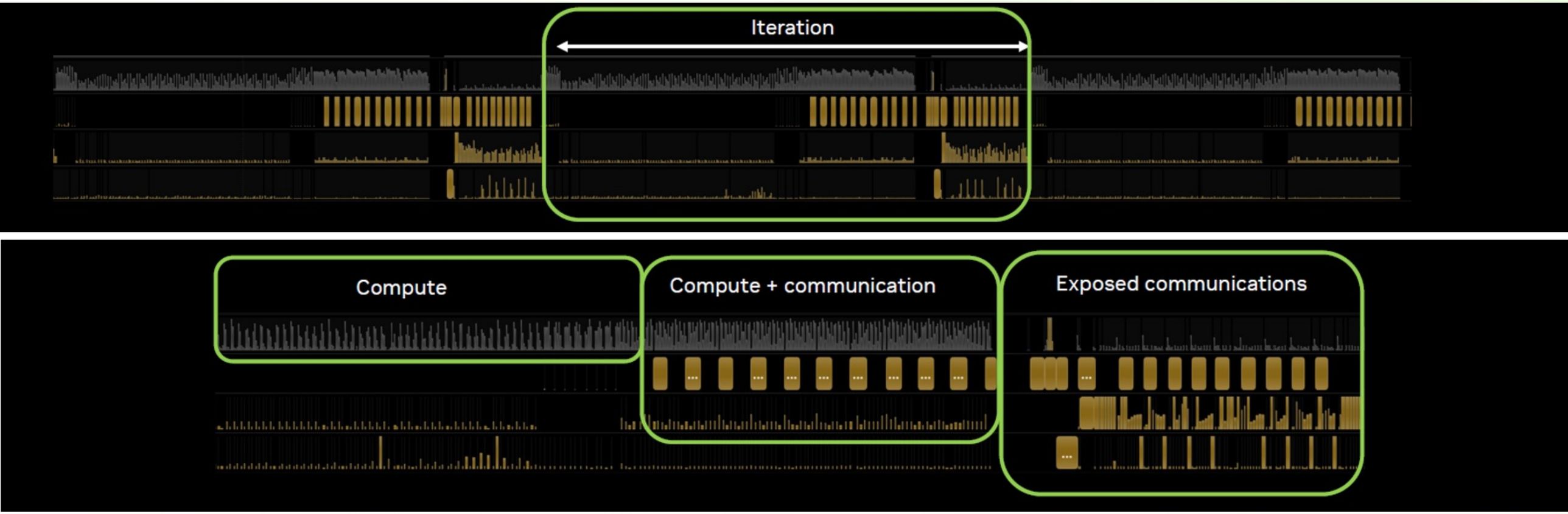
Training Deep Learning Models at Scale: How NCCL/SAHRP Enables Best Performance on AI Data Center Networks

Sungta Tsai, Sr. Solution Architect | August 2024



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

System Designers, to understand how design choices will affect the performance of DL training

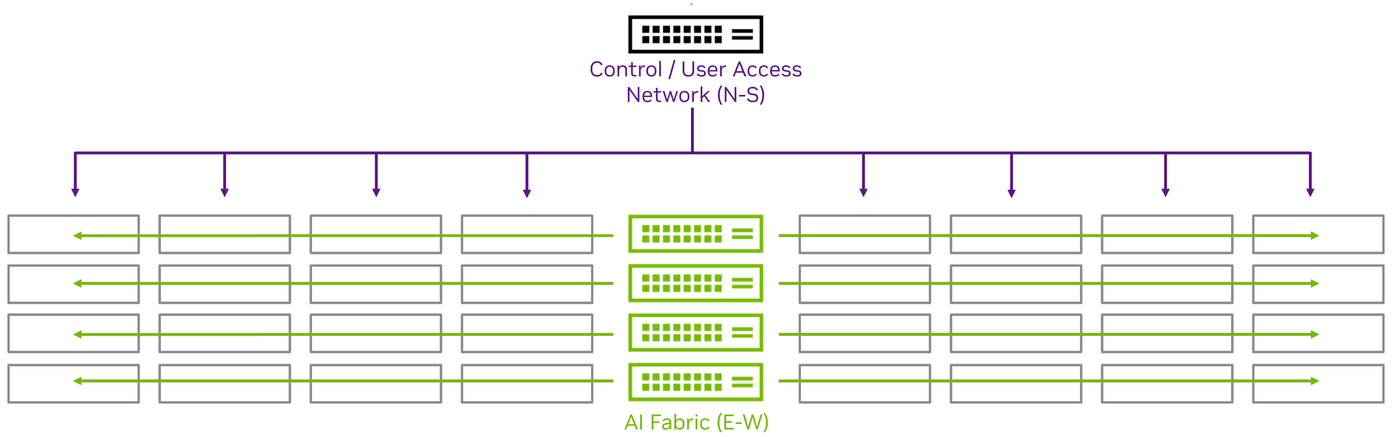
- PCI topology
- NVLink/NVLink switch support (not only for intra-node communication!)
- Network technology and topology

Users, to know what performance and scalability to expect from a given platform.

Developers, if they need inter-GPU communication for their application.

AI Clouds Going Through A Major Change

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 Capacity, Predictive Performance

Hardware & Software Accelerated In-Network Computing

AI Network Considerations

Software Acceleration

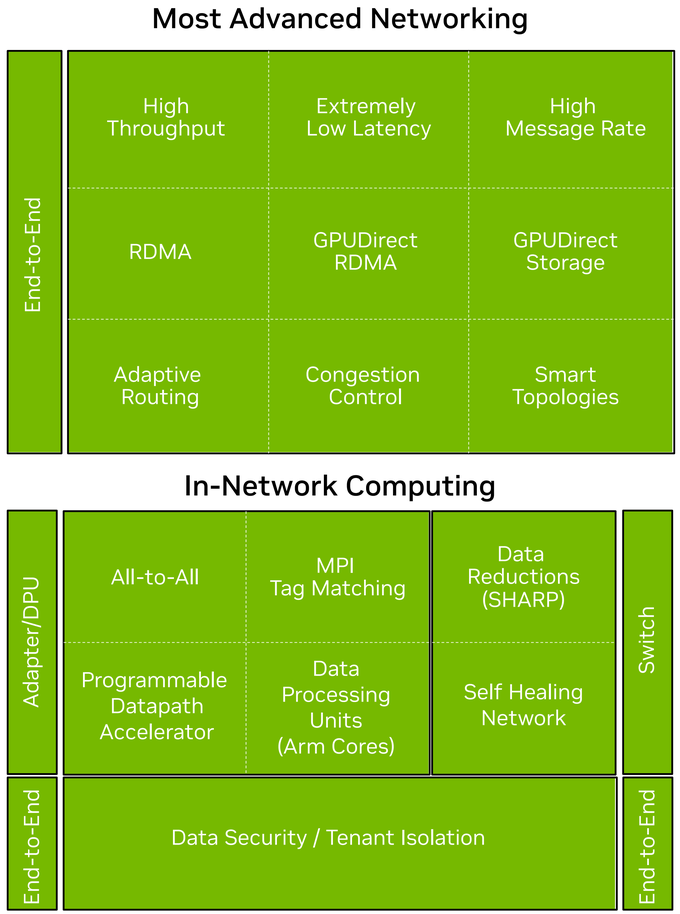
NCCL — NVIDIA Collective Communication Library

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

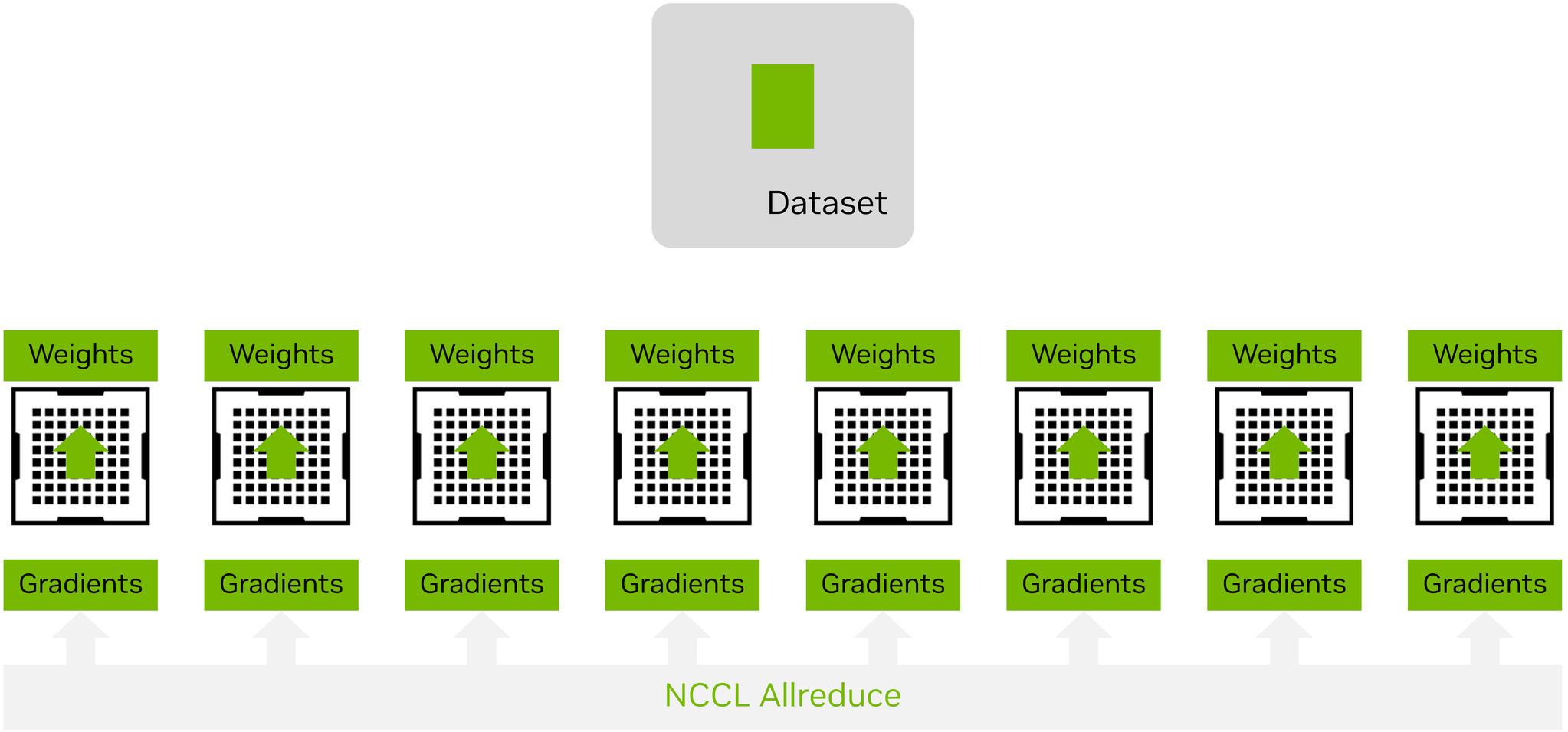
Hardware Acceleration

SHARP — NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol Technology

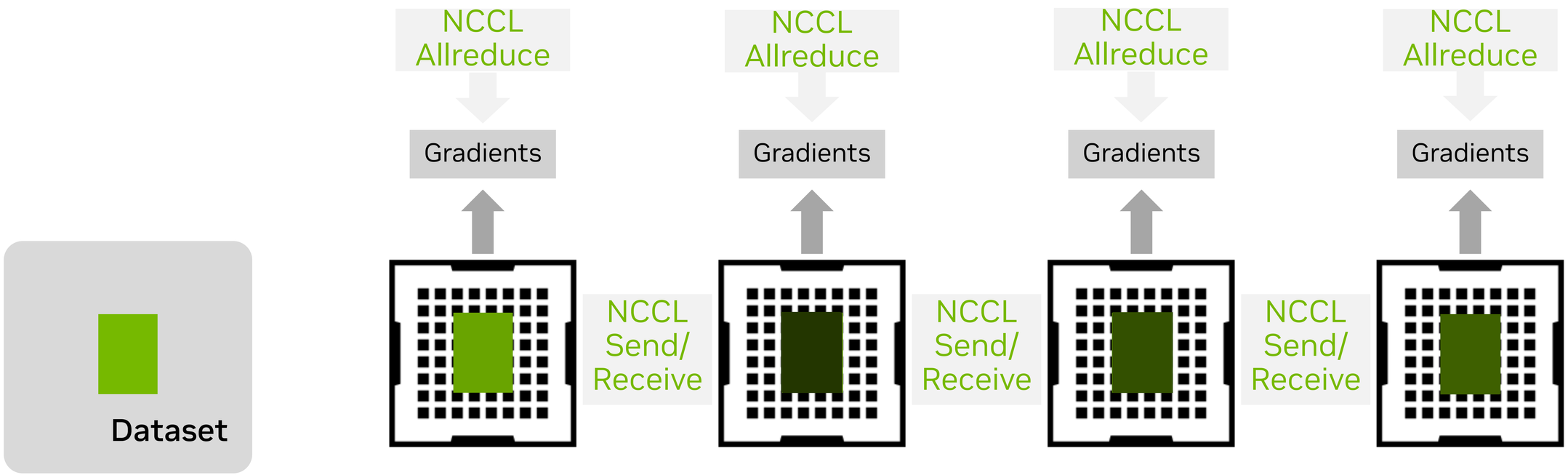
SHARP is part of the InfiniBand and NVLink switch ASICs. It enables the network to perform data reduction operations, an important element of AI workloads. This decreases the amount of data traversing the network and dramatically reduces collective operations time.



Data Parallelism

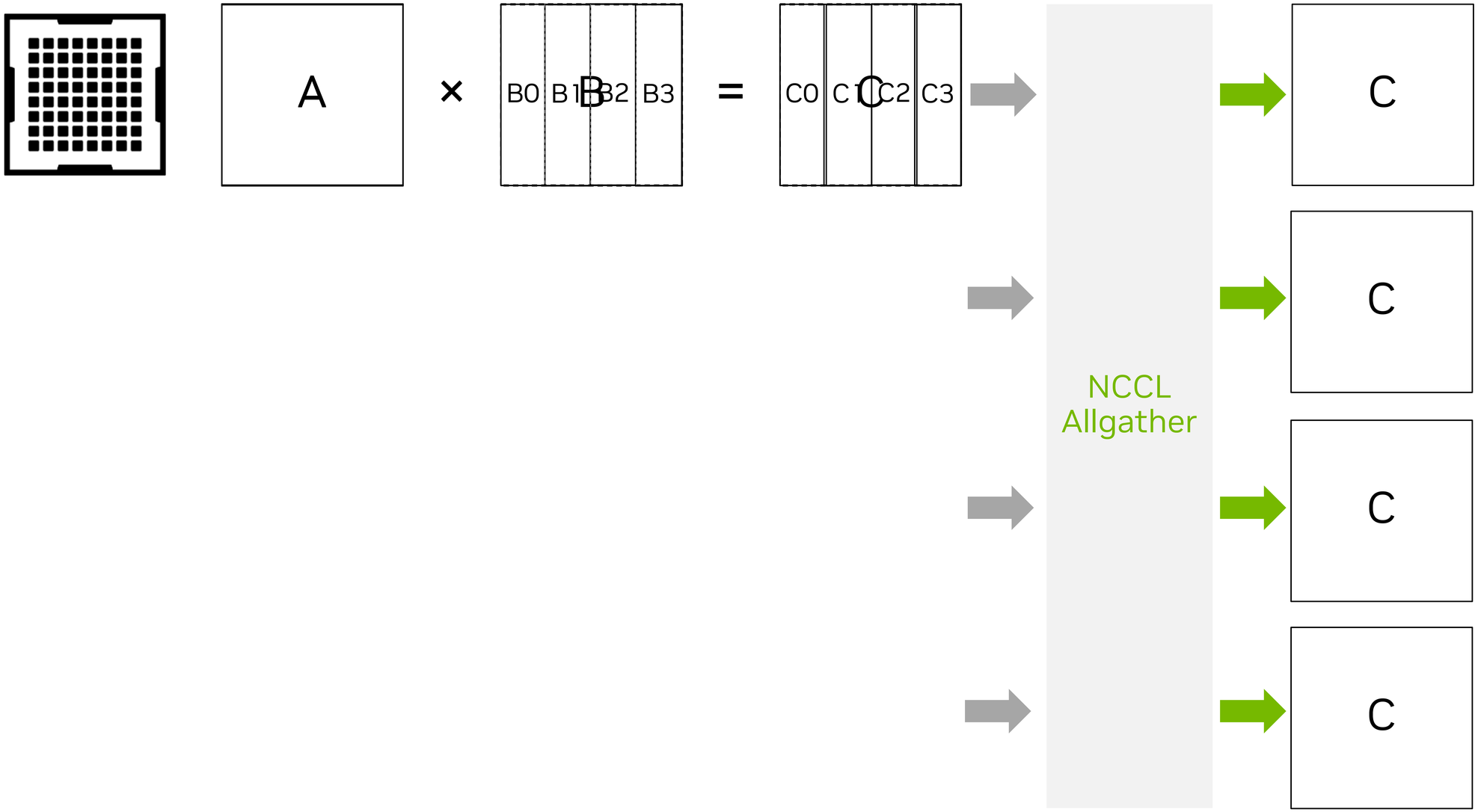


Pipeline Parallelism



Deep Learning Training

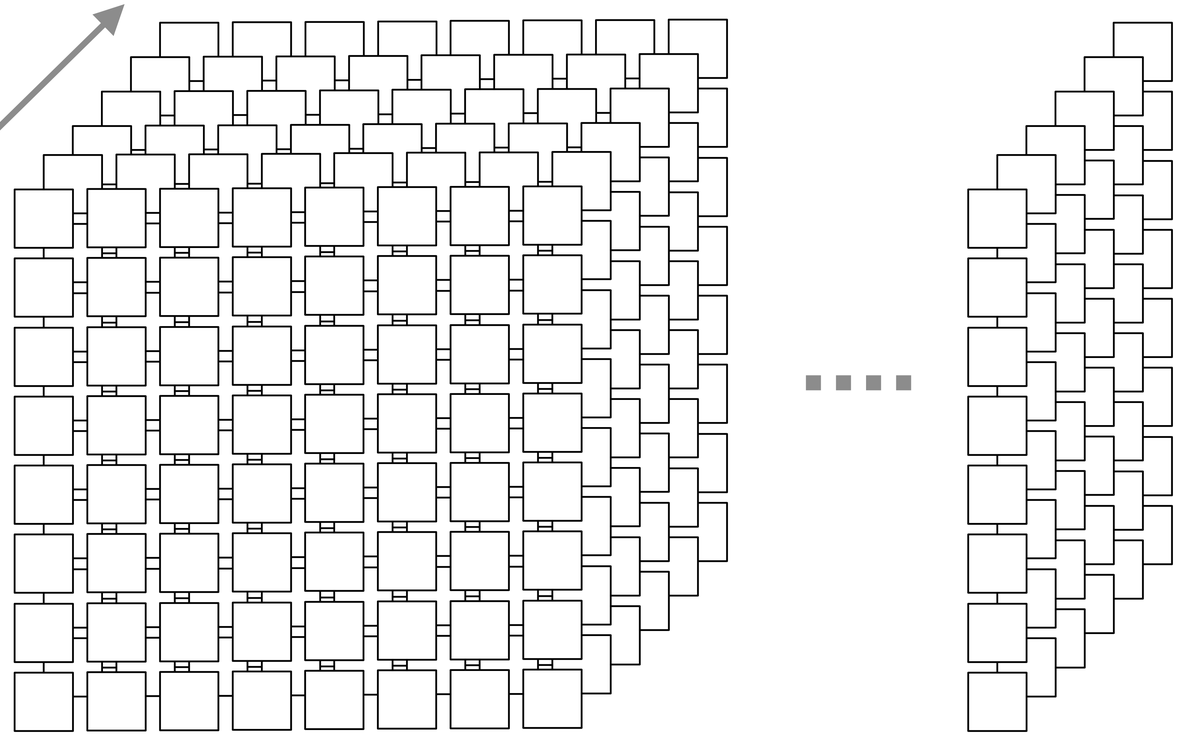
Tensor parallelism



Large scale LLM Training

Pipeline parallelism

ncclSend/
ncclRecv



Tensor parallelism

ncclAllGather

Data parallelism
ncclAllReduce

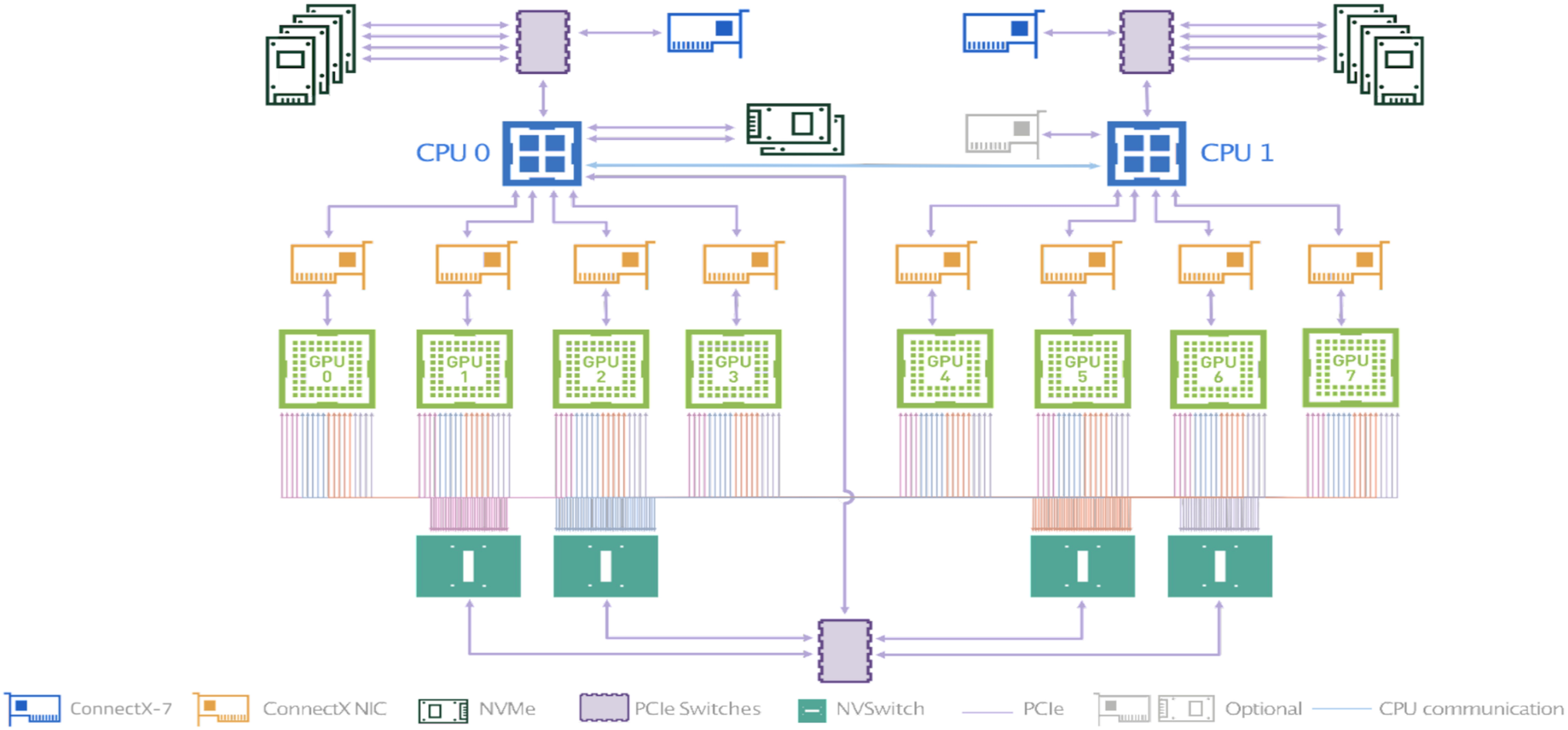
And also:

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

FSDP (fully sharded data parallelism)
ncclAllGather

And other variations …

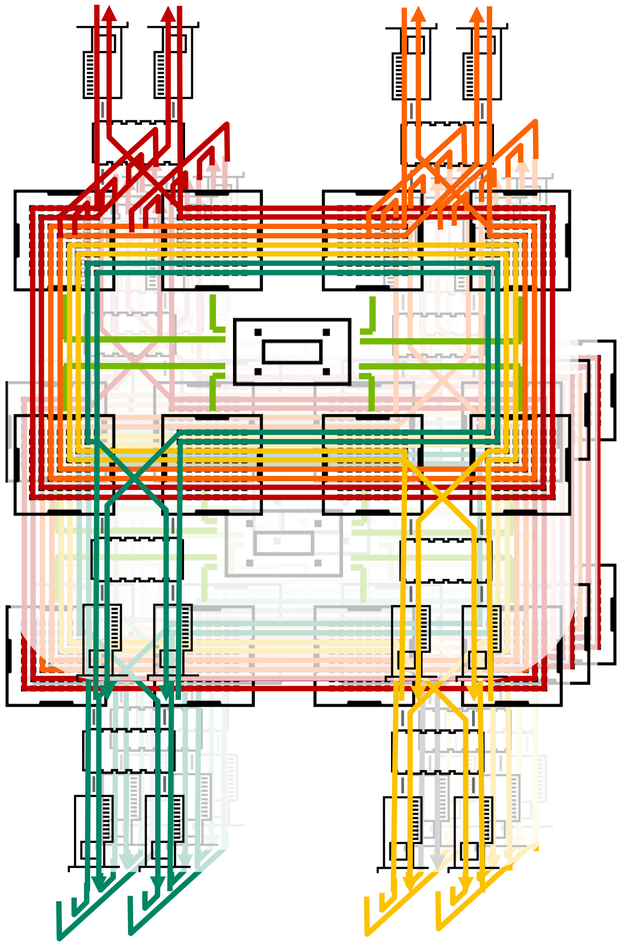
DGX H100 Block Diagram



Hierarchical Fat Tree and Rail-Optimized Design

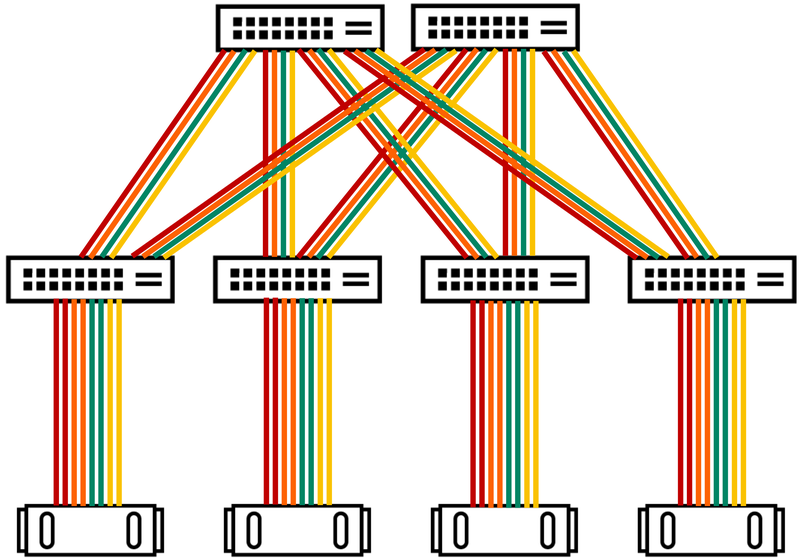
Inter-node Communication

Routing must be perfect to ensure all flows use different links.

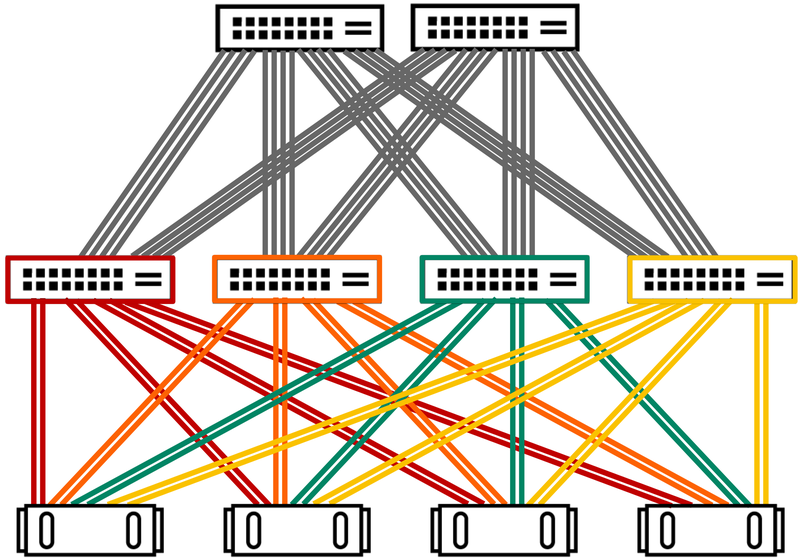


All traffic is local to leaf switches. Routing collisions are impossible.

Classic fabric design

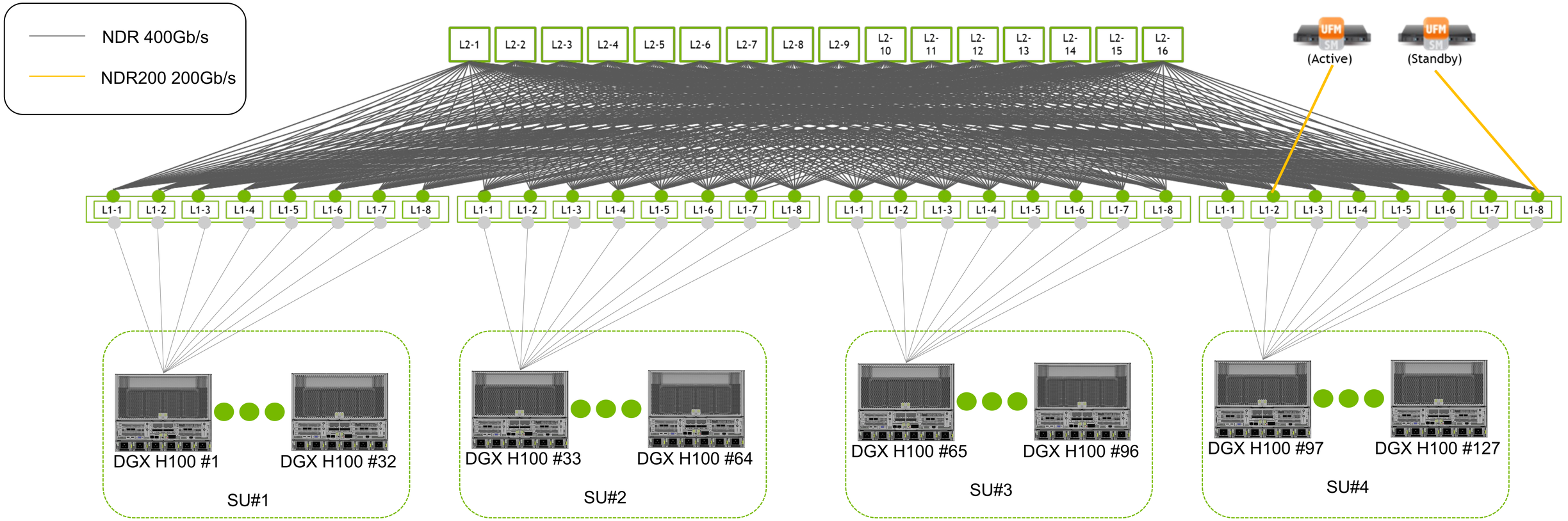


Rail-optimized design



Hierarchical Fat Tree and Rail-Optimized Design

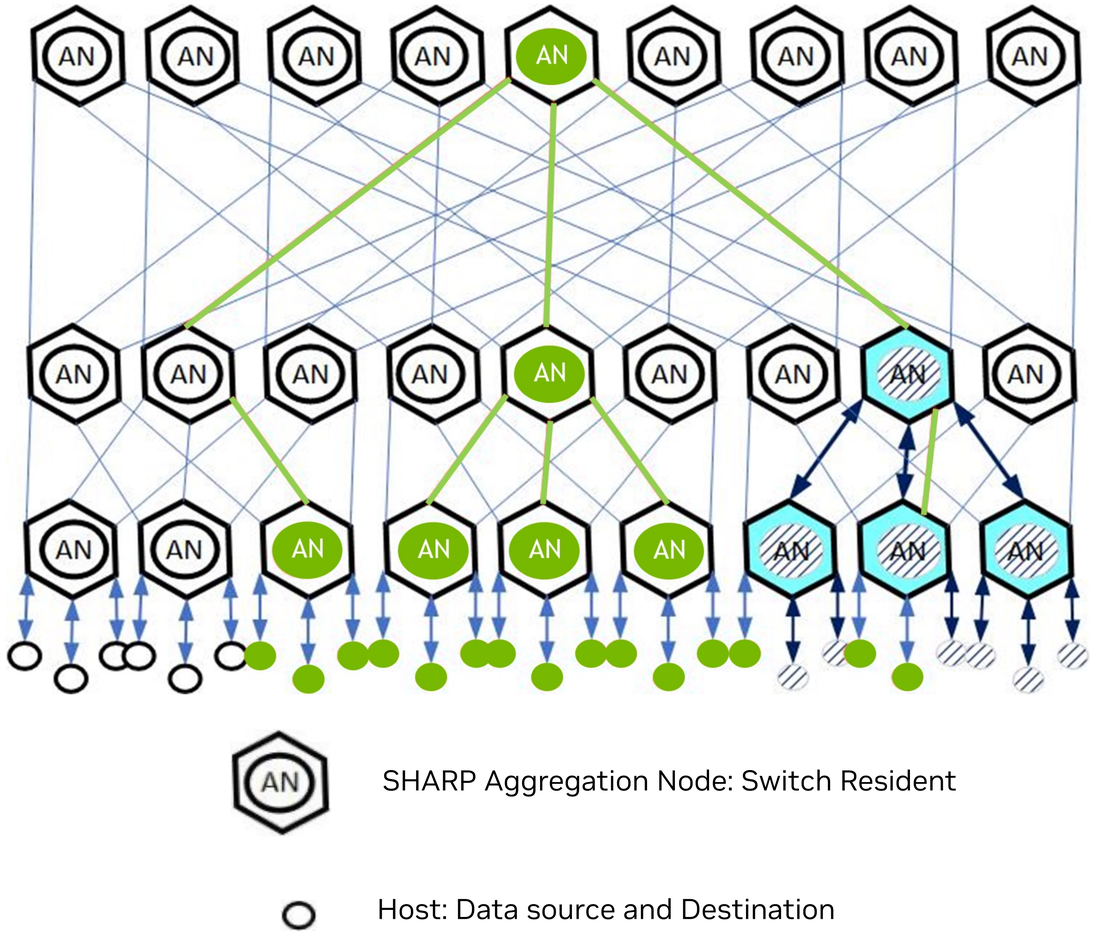

QM9700 Leaf & Spine



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



Scalable Hierarchical Aggregation and Reduction Protocol

SHARP for AllReduce Operation