

# Network Infrastructure for AI: Designing Connectivity for AI Factory & AI Data Center

*NVIDIA (Reference Architecture) and FPT AI Factory in Vietnam & Japan*

---

SPEAKERS

AI Factory Infrastructure Expert

Can Tho | Month 10/2026



# CONTENT OVERVIEW

**01**

## **AI Background & Problems**

The explosion of LLM, Distributed Training and AI Data Center characteristics.

**02**

## **Network Heart: RDMA**

Kernel bypass, GPUDirect RDMA, and Zero-Copy architecture.

**03**

## **Architecture 04 Network Modules**

Backend, Storage, Frontend, and OOB network segments.

Compare IB vs RoCEv2.

**04**

## **Operations & Consultations**

Fiber optic infrastructure, Telemetry and control automation.

Recommendations for AI Data Center in Vietnam



# BACKGROUND & AI PROBLEMS

## The LLM Era: The Rise of the AI Factory

### Explosive model scale

Exceeding the threshold of 1,000 billion parameters, exceeding the processing power of a single server.

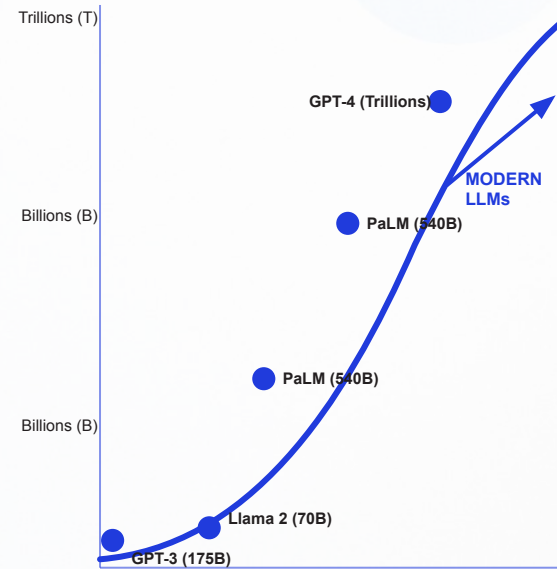
### Migrating to AI Factory

Data Centers are not only storage but artificial intelligence factories.

### Bottlenecks in network

Bandwidth and link latency directly determine GPU performance.

#### EXPONENTIAL GROWTH OF AI PARAMETERS



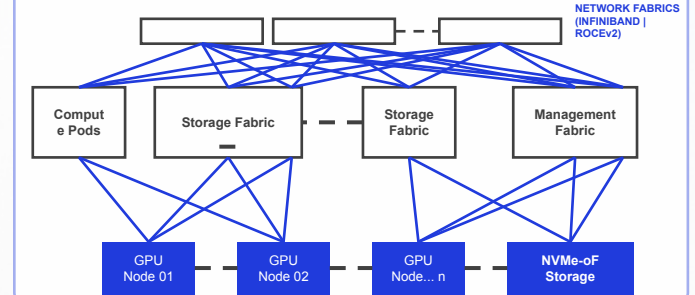
#### THE SHIFT TO AI DATA CENTERS / AI FACTORIES

**High-Density Compute**  
Compute high-density computing in AI data centers.

**Ultra-Low Latency**  
Ultra-low latency for GPU-to-GPU linking.

**Massive Bandwidth**  
Large bandwidth connected to storage infrastructure.

#### AI FACTORY — SPINE-LEAF FABRIC



**~1355 GB**  
data rates

**~8.3B → 1T**  
GPU parameters

**High-Density Compute**  
High-density GPU cluster, high-speed internal linking.

**Ultra-Low Latency**  
Low-latency fabric for distributed training.

**Massive Bandwidth**  
Large bandwidth for datasets and checkpoints.

# BACKGROUND & AI PROBLEMS

## Distributed Training: The Supercluster Synchronization Challenge

### Distributed Data & Models

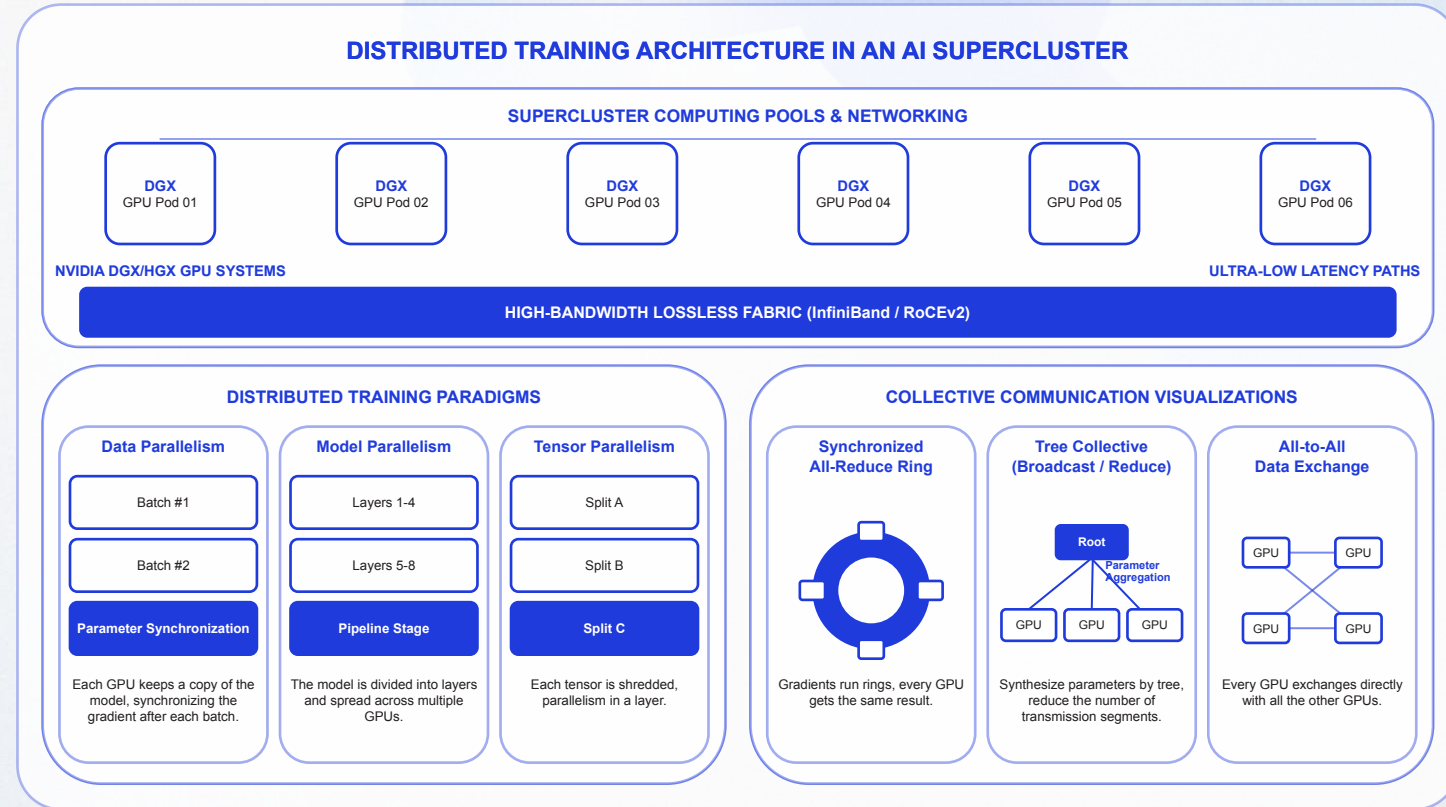
Data Parallelism and Model Parallelism across thousands of GPUs.

### Jump synchronization mechanism

Gradient weights must be updated synchronously after each iteration

### Network congestion causes GPU Idle

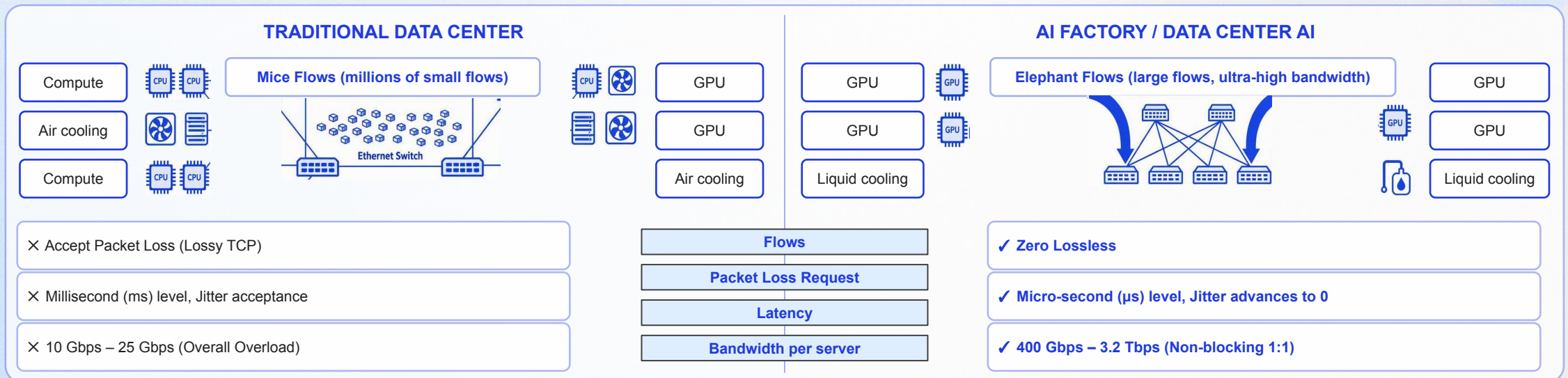
The fastest GPU always has to wait for the slowest GPU to complete the transmission.



# BACKGROUND & AI PROBLEMS

## Traditional Data Center vs AI Data Center

| NETWORK CHARACTERISTICS | TRADITIONAL DATA CENTER                        | AI FACTORY / DATA CENTER AI                         |
|-------------------------|------------------------------------------------|-----------------------------------------------------|
| Flows                   | Mice Flows (millions of small flows scattered) | Elephant Flows (few flows, extreme intensity)       |
| Packet Loss Request     | Accept Random Packet Loss (Lossy TCP)          | Zero Lossless                                       |
| Latency                 | Millisecond level (ms), Jitter acceptance      | Micro-second ( $\mu$ s) level, Jitter advances to 0 |
| Bandwidth per server    | 10 Gbps – 25 Gbps (Overall Overload)           | 400 Gbps – 3.2 Tbps (Non-blocking 1:1)              |





# HEART OF AI NETWORKS - RDMA TECHNOLOGY

## The bottleneck of the traditional TCP/IP stack

### Continuous Context Switching

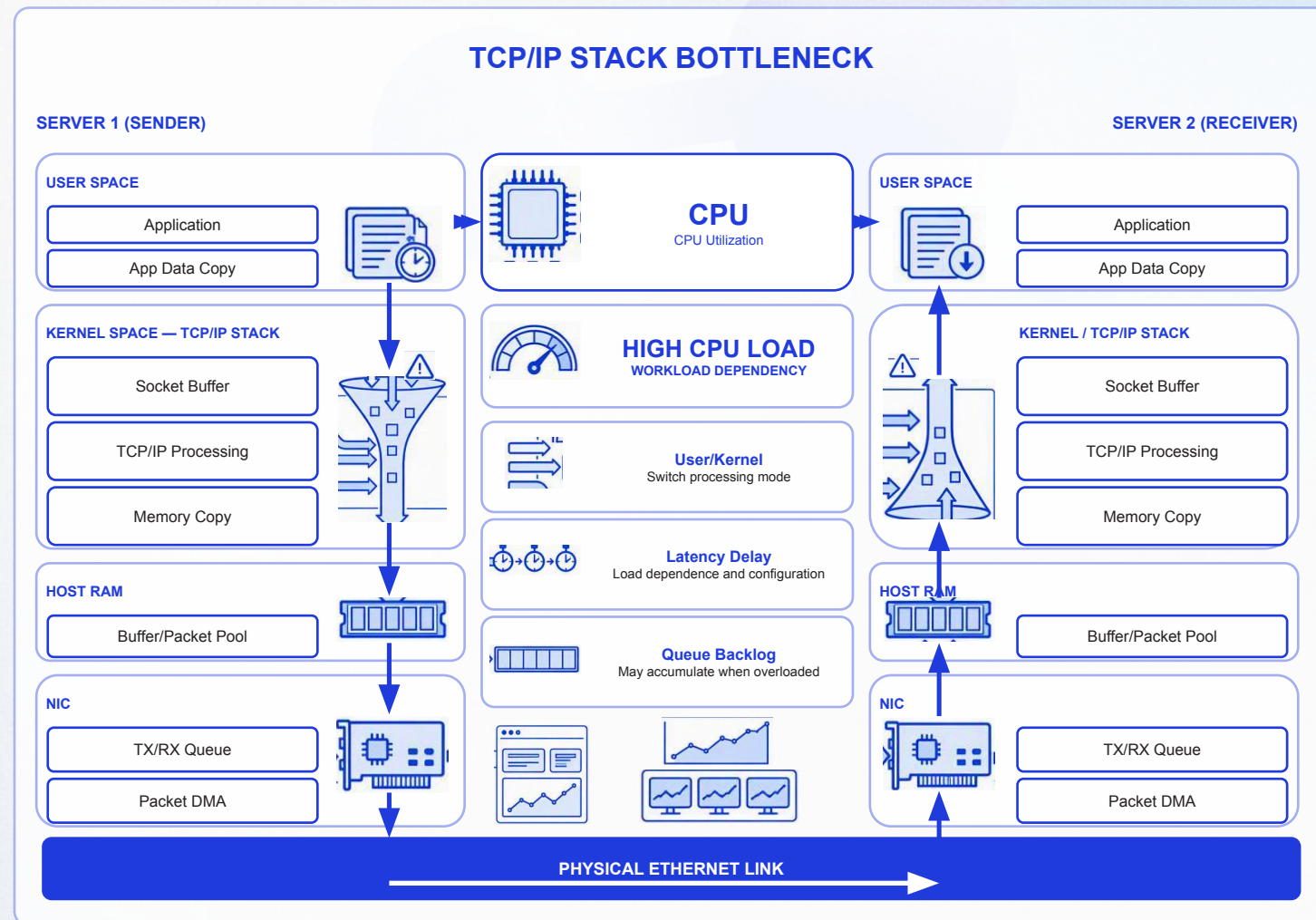
Frequent context switching between user space and kernel space creates a CPU bottleneck

### Multi-Tier Memory Replication

Continuous buffer and host memory copies cause latency spikes

### CPU Interrupts

It is not possible to handle 400Gbps/800Gbps traffic per node.



Simplified data flow diagram. Workload, configuration, and offload dependent performance.



# HEART OF AI NETWORKS - RDMA TECHNOLOGY

## RDMA: Kernel Bypass & Zero-Copy Architecture

### Remote Memory Access

Read/write other server RAM directly via NIC hardware.

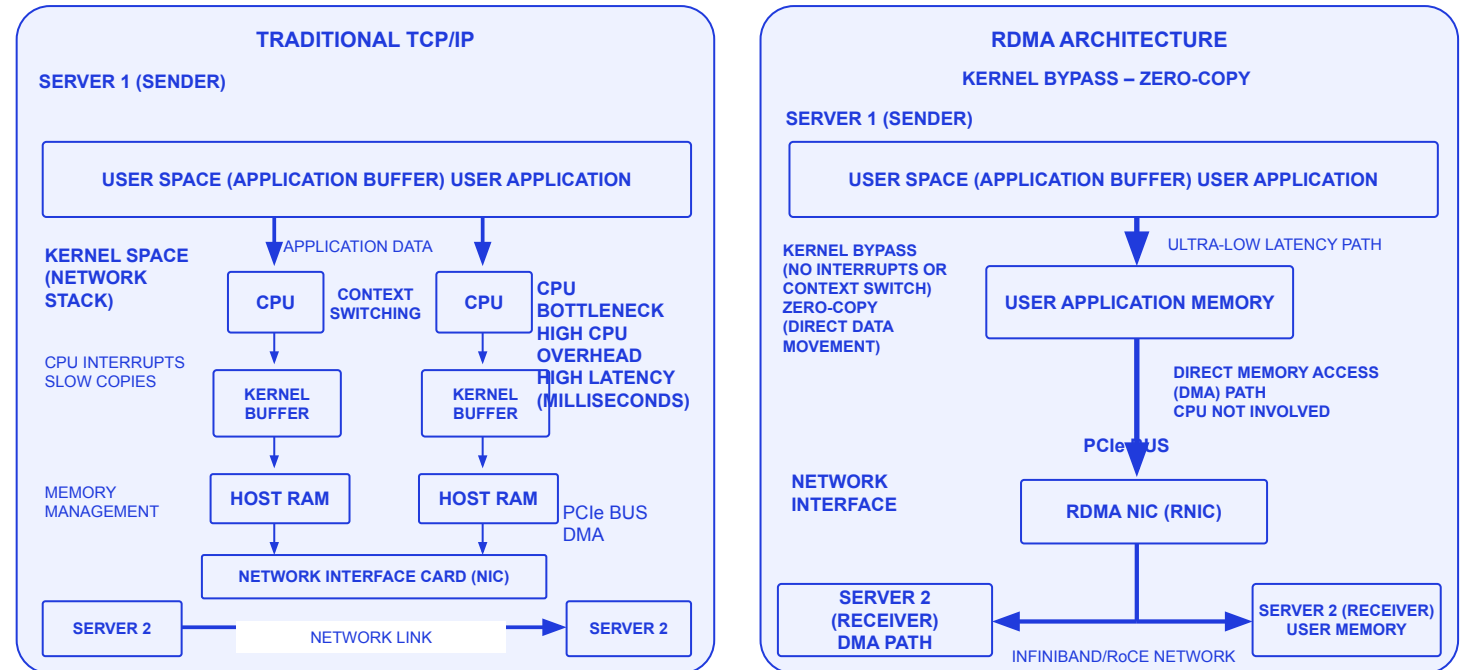
### GPUDirect RDMA

Data goes straight from the NIC to the GPU VRAM via PCIe/NVLink.

### Completely frees up CPU

0% CPU overhead for data transport on the network.

### RDMA: KERNEL BYPASS AND ZERO-COPY ARCHITECTURE



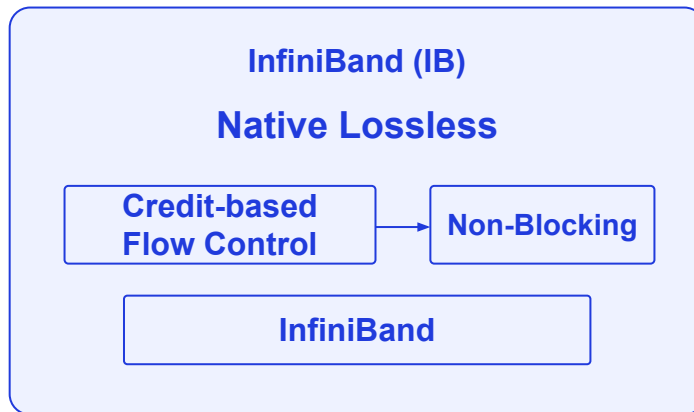
TRADITIONAL TCP/IP: 11 ms / 1 ms RDMA: 0 ms / MICROSEC ULTRA-LOW LATENCY



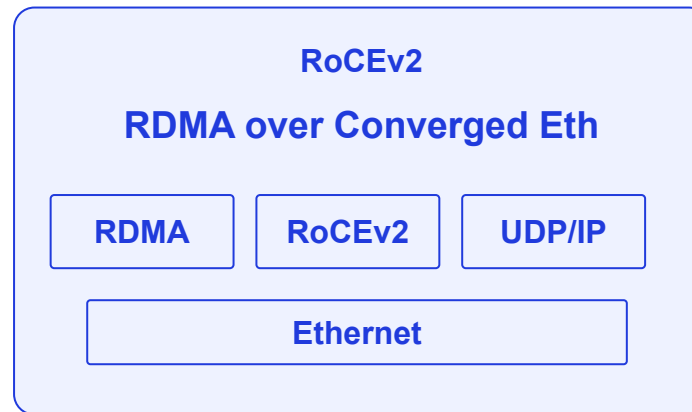
# HEART OF AI NETWORKS - RDMA TECHNOLOGY

## HEART OF AI NETWORKS - RDMA TECHNOLOGY

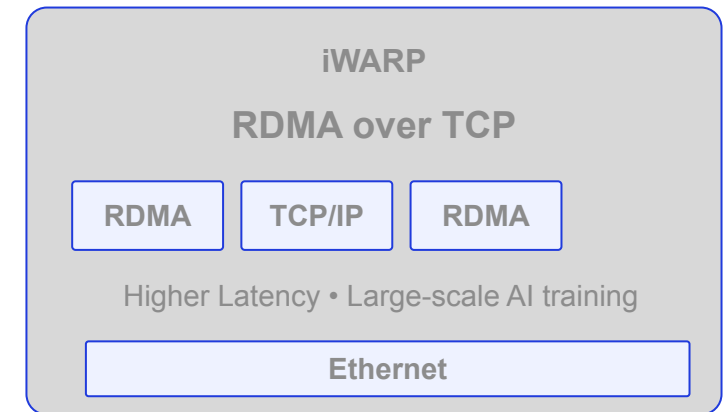
Three common RDMA realization protocols



Native HPC protocol with hardware Credit-based Flow Control for congestion avoidance



RDMA encapsulation on UDP/IP. Cost-effective, take advantage of the extensive Ethernet ecosystem.



Runs on standard TCP. Easy to deploy but lag high, rarely used for large-scale AI Training.



# AI NETWORK MASTER ARCHITECTURE

## Overview 04 Network Segments

### 1. Backend Fabric (GPU Compute)

Inter-node GPU connectivity (East-West traffic)

### 2. High-Performance Storage Fabric

Continuous dataset ingestion and checkpointing

### 3. Frontend Fabric (North-South)

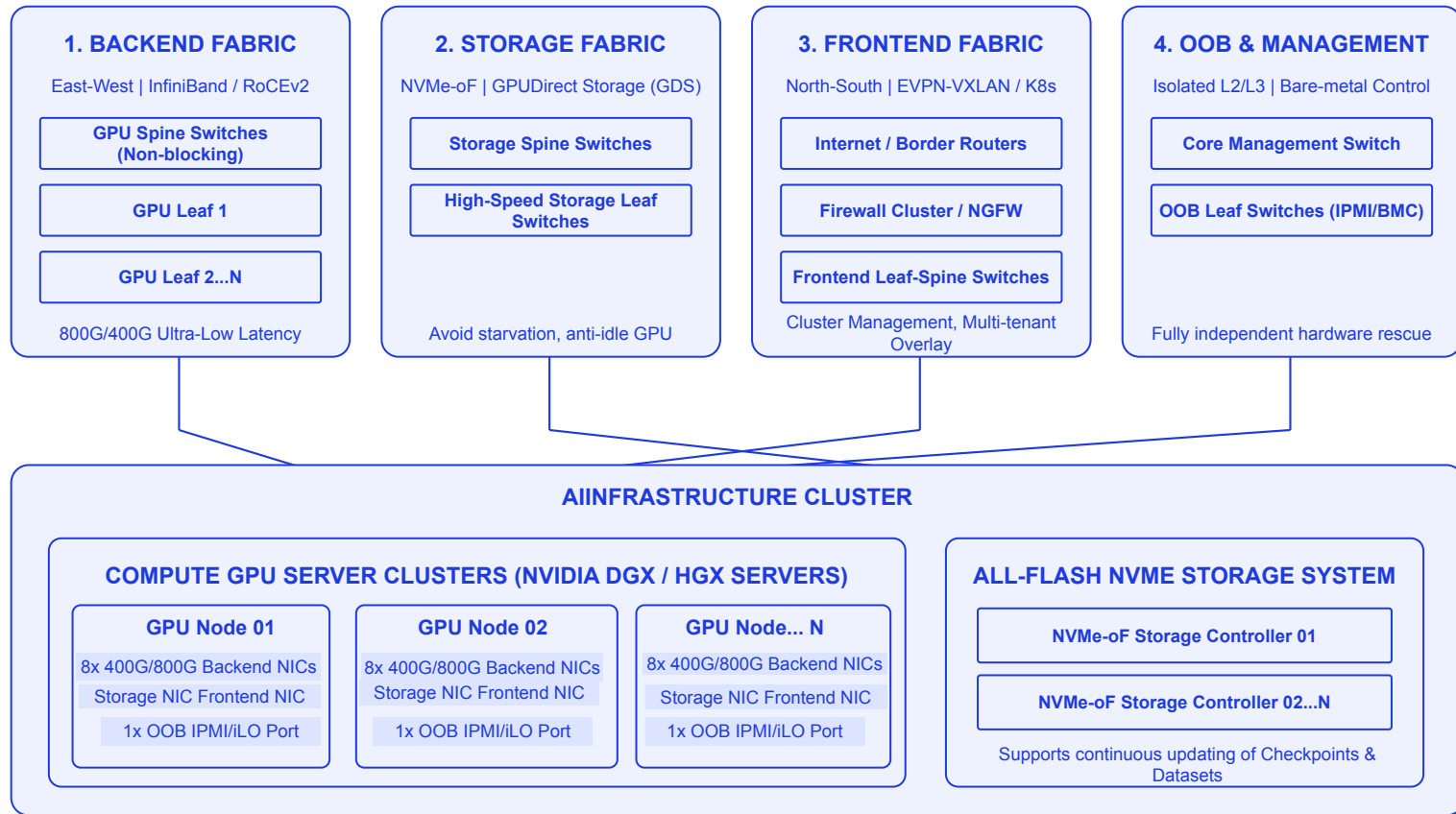
Kubernetes/Slurm/Cloud administration and data ingestion for end-user services

### 4. Out-of-Band (OOB) Management

Independently managed network via IPMI/BMC.

## OVERALL ARCHITECTURE 4 NETWORK MODULES (AI DATA CENTER FABRICS)

Independent sharding design optimized for GPU & NVMe Storage distributed computing





# AI NETWORK MASTER ARCHITECTURE

## Segment 01: Backend Fabric - The Heart of GPU Computing

### Overwhelming East-West Traffic

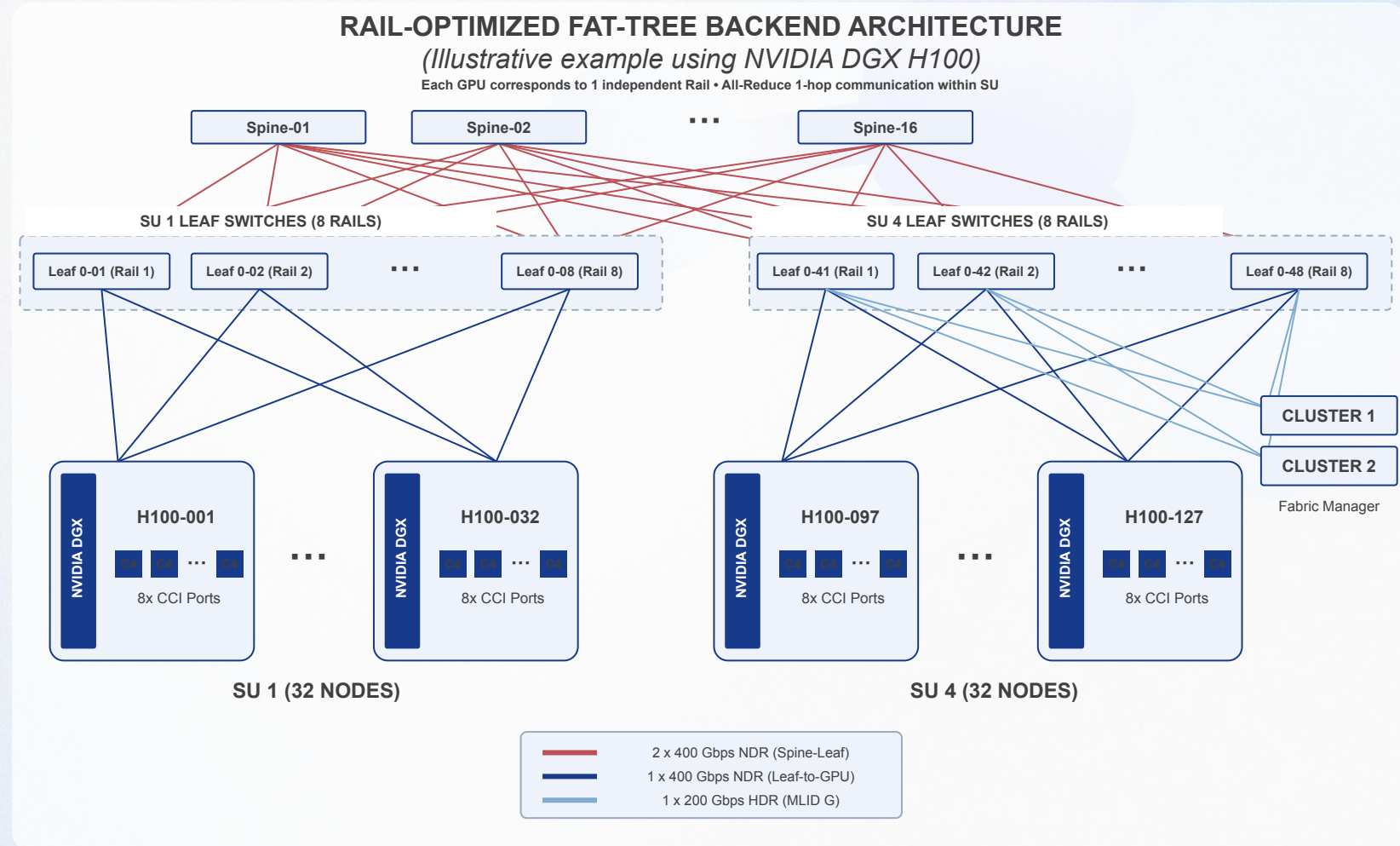
Accounting for more than 85% of the total exchange traffic during the training process.

### Huge bandwidth

From 400 Gbps (NDR) to 800 Gbps (XDR) per port.

### 1:1 Non-Blocking Architecture

Non-blocking architecture to eliminate uplink/downlink bottlenecks



# OVERALL ARCHITECTURE OF AI DATA CENTER NETWORK

## Rail-Optimized & Fat-Tree Spine-Leaf Design

### Fat-Tree Spine-Leaf Model

Consistent hop count between nodes to eliminate latency variance

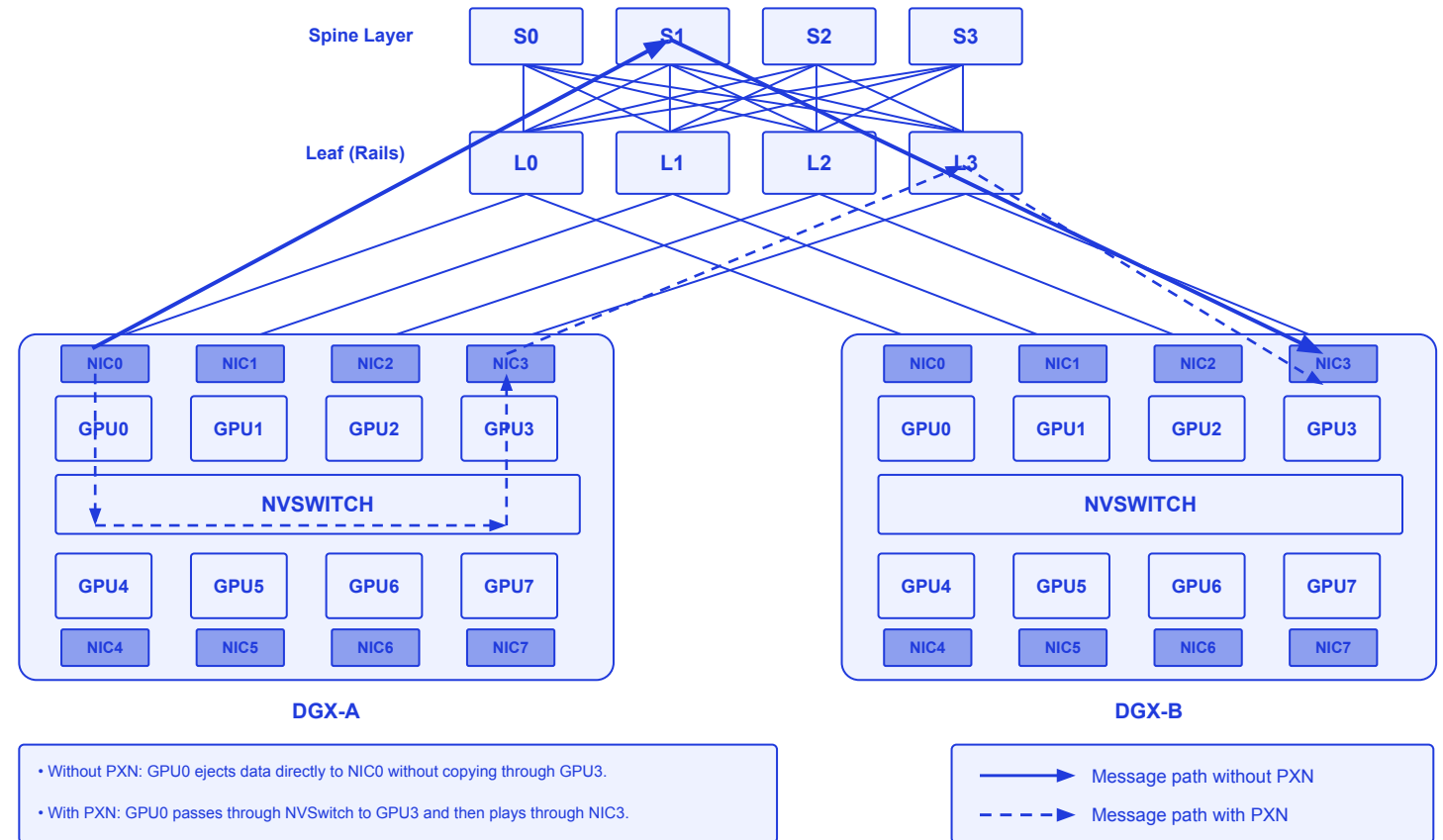
### Rail Structure

GPUs with the same ID (GPU 0...7) are connected to an independent cluster of Switches.

### All-Reduce 1-Hop Communication

There is no need to go through the Spine layer when transferring messages between GPUs and the same rail.

NVIDIA RAIL-OPTIMIZED NETWORK TOPOLOGY  
Rail-Optimized Architecture & PXN Accelerated Routing (PCIe/NVLink Cross-NIC)



# OVERALL ARCHITECTURE OF AI DATA CENTER NETWORK

## Option 1: InfiniBand (NVIDIA Quantum)

### Native Lossless with Credit Flow

The recipient grants a buffer credit before the sender is entitled to transmit the message.

### Quantum-2 & Quantum-X Generation

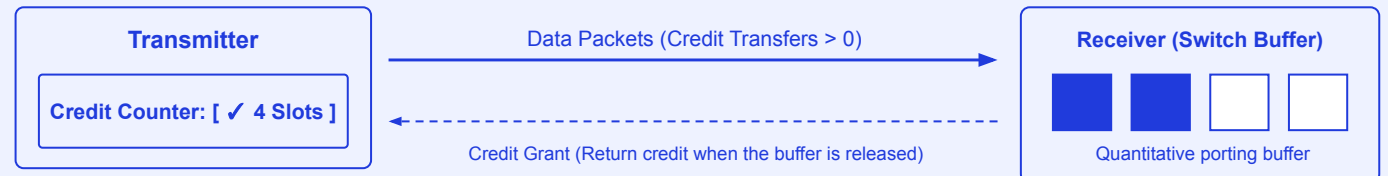
400 Gbps NDR & 800 Gbps XDR with nano-second (ns) switching latency.

### Closed-loop optimization

Nvidia's perfect sync ecosystem from GPUs, ConnectX NICs to Quantum Switches.

## DEDICATED AI NETWORK ARCHITECTURE School 1: NVIDIA InfiniBand Architecture

### 1. Native Lossless: Credit-based flow control



### 2. Quantum-2 & Quantum-X generation

#### NVIDIA Quantum-2 (NDR)

400 Gbps bandwidth per port (64 NDR Ports)

~100–130 ns

#### NVIDIA Quantum-X (XDR)

800 Gbps bandwidth per Blackwell generation port

< 100 ns

### 3. Closed optimization ecosystem

#### NVIDIA HGX/DGX GPU

H100 / H200 / B200/ B300  
Tensor Core

#### ConnectX-7 / ConnectX-8 SuperNIC

GPUDirect RDMA Engine

#### Quantum Switches & NVIDIA UFM

SHARP & Adaptive Routing

Perfect synchronization of hardware - firmware - drivers, no intermediate lag



# AI DC NETWORK MASTER ARCHITECTURE

## InfiniBand: Adaptive Routing & SHARP Technology

### Adaptive Routing (AR)

Dynamically route each packet, finding the least congested route in real time.

### NVIDIA SHARP Protocol

Push the All-Reduce calculation directly on the Switch.

### 50% reduction in network traffic

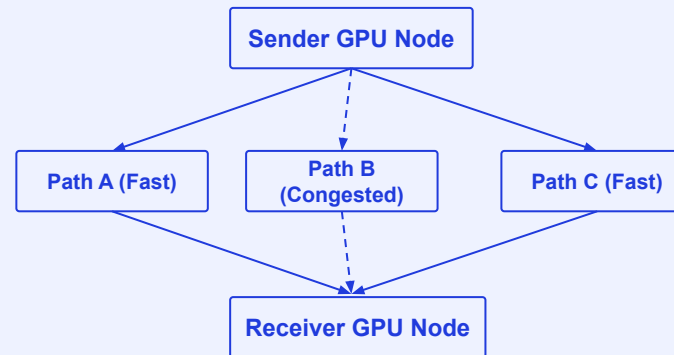
Enables deduplication of intermediate data moving between GPUs.

### DEDICATED AI NETWORK ARCHITECTURE

## InfiniBand: Adaptive Routing & SHARP Technology

### 1. Adaptive Routing (AR)

Real-time dynamic load balancing (Per-packet)



- **Suppression of ECMP hash conflicts**

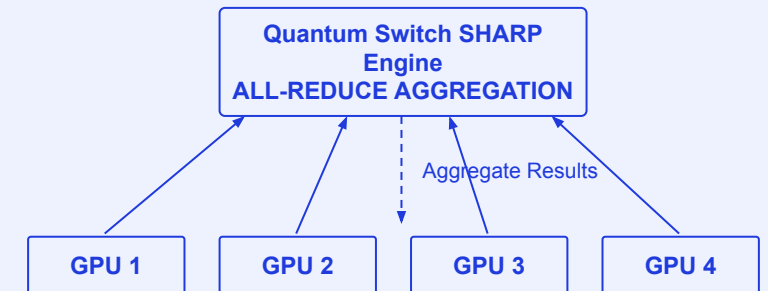
Packets automatically avoid traffic jams to optimize 100% of network capacity.

- **Packet Re-ordering**

The ConnectX hardware rearranges packets at the destination itself.

### 2. NVIDIA SHARP Protocol

In-Network Computing: Compute on Switch



**-50%**

### 50% REDUCTION IN NETWORK TRAFFIC

Completely eliminate GPU-mediated transfer steps.

- **Free up GPU computing resources:**

GPUs just need to focus on training without the public gathering of gradients.



# AI DC NETWORK MASTER ARCHITECTURE

## Option 2: Ethernet / RoCEv2

### Avoid Vendor Lock-in

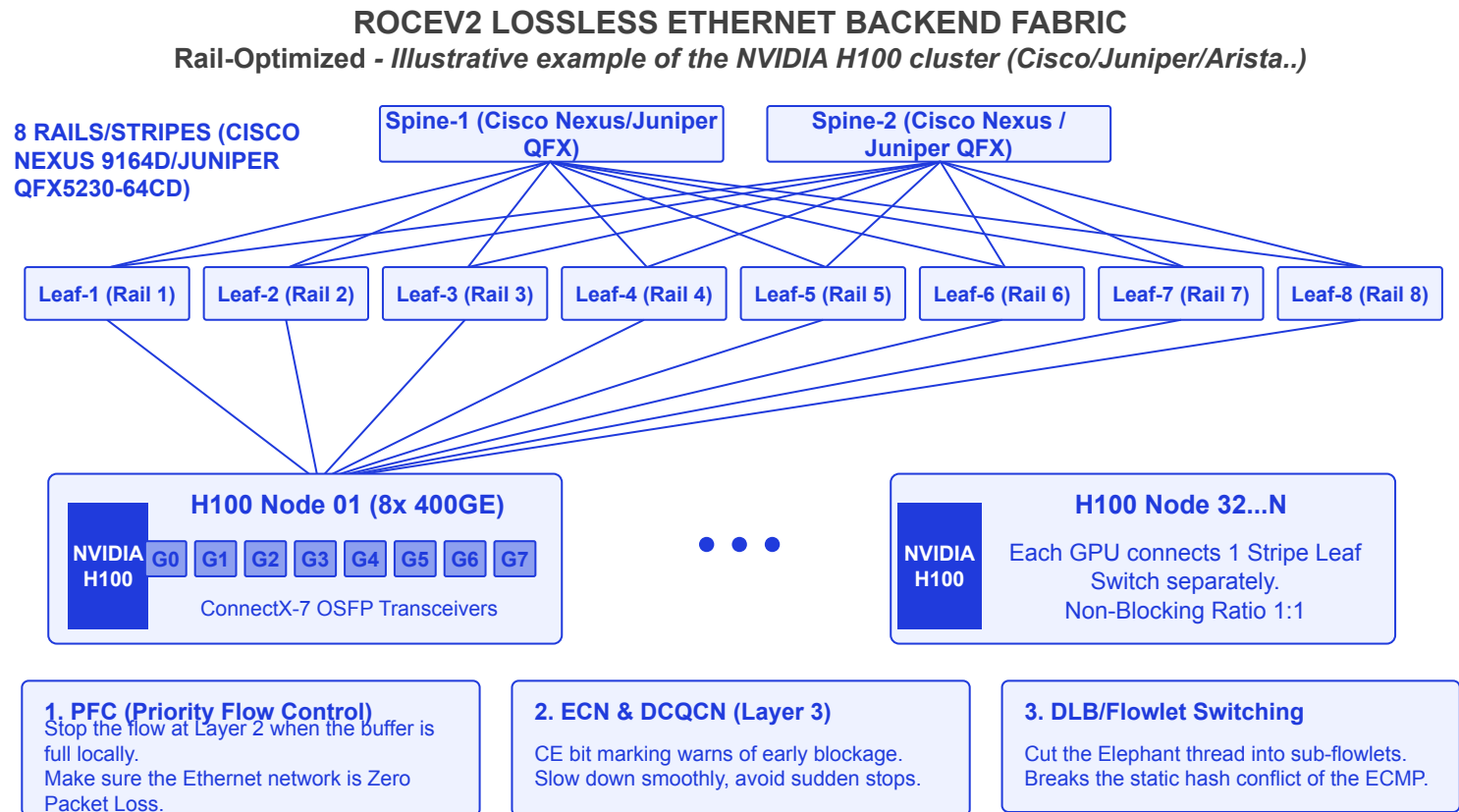
Diversified suppliers of optical equipment and components standard.

### Speed to catch up with InfiniBand

The 51.2 Tbps generation of Ethernet switches supports 400GE and 800GE ports.

### Mandatory Lossless Ethernet

Adopt PFC (Priority Flow Control) flow control mechanism.





# AI DC NETWORK MASTER ARCHITECTURE

## Congestion Control: ECN, DCQCN & DLB Load Balancing

### ECN (Explicit Congestion Notification)

The CE bit marking warns early when the buffer begins to accumulate.

### DCQCN Combined Algorithm

Smooth speed adjustment remotely, holding the PFC as the final stopper.

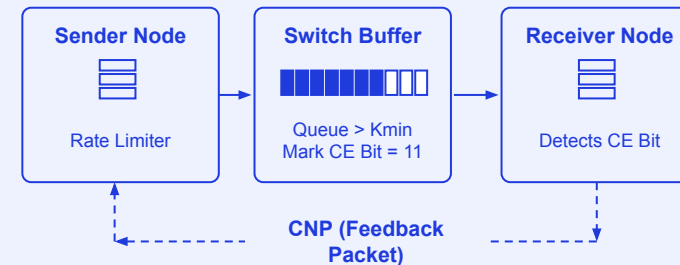
### Dynamic Load Balancing (DLB/Flowlet)

Load sharing mechanism for Elephant flows.

## ROCEV2 CONGESTION CONTROL & LOAD BALANCING ECN, DCQCN & DLB Load Balancing (Cisco & Juniper)

### 1. Congestion Control: ECN & DCQCN

Smooth rhythm rate throttling loop, PFC prevention.



#### Step 1 & 2: ECN Threshold Trigger

Switches (Nexus/QFX) that cross the Kmin threshold will mark the CE bit. Do not drop the packet, do not emit a PFC signal too early.

#### Step 3: CNP Generation & Rate Backoff

The Receiver creates a CNP packet that sends it back to the source network card. Sender reduces the transfer rate to free up the switch queue.

#### PFC (Priority Flow Control) is the last stop

Only when the buffer hits the emergency ceiling will PAUSE be activated. A well-functioning ECN suppresses PFC Deadlock sending.

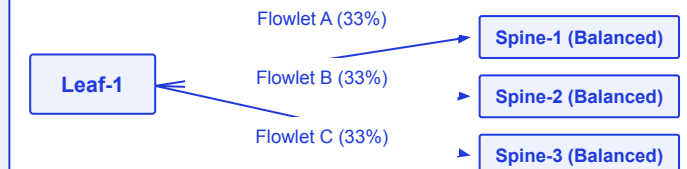
### 2. Load Balancer: ECMP vs DLB/Flowlet

Thoroughly resolve Hash Collision conflicts on AI traffic.

#### Traditional ECMP (Static by 5-tuple Hash)



#### DLB/Flowlet Switching (Cisco Nexus & Juniper QFX)



#### Advantages of deployment on actual infrastructure:

- Cisco Flowlet Load Balancing (FLB): Threading micro-gaps, splitting elephant threads.
- Juniper Dynamic Load Balancing: Reduces queue congestion in real-time.

Achieve 95% > line utilization efficiency equivalent to InfiniBand Adaptive Routing.



# OVERALL ARCHITECTURE OF AI CENTER NETWORK

## Technology Comparison: InfiniBand vs Ethernet (RoCEv2)

### NVIDIA INFINIBAND (QUANTUM)

HPC & AI Optimum Standards

#### 1. Native Lossless (Credit-Based)

Allocate hardware credits before transmission.

✓ Never drop the package, no complicated configuration required.

#### 2. Ultra-low Switch latency ~100 - 130 ns (even <100ns)

Switch Quantum-2 (NDR 400G) & Quantum-X (800G).

✓ Ultra-uniform latency (Zero Jitter) for All-Reduce.

#### 3. Adaptive Routing & SHARP In-Network

Dynamic load balancing each packet, automatically avoiding clogged queues.

✓ SHARP reduces traffic between GPUs by 50%.

#### 4. Centralized Administration with NVIDIA UFM

Built-in Subnet Manager, automatically detects cable loss.

✓ Comprehensive optimization for DGX SuperPOD clusters.

**PROS:** Absolute performance, plug-and-play configuration.

**CONS:** Exclusive 1 brand (Vendor Lock-in), high cost.

Perfect for: The ultimate computational AI supercluster, HPC hub.

VS

### ETHERNET/ROCEV2

Open Ecosystem (Cisco, Juniper, Arista, etc.)

#### 1. Lossless via PFC+ECN/DCQCN

PFC stops Layer 2 queues; ECN/DCQCN slows down Layer 3.

△ Requires strict configuration and fine-tuning of parameters.

#### 2. Switch Latency: > 400 ns

Switch 51.2T (Cisco Nexus 9364D, Juniper QFX5240).

✓ Bandwidth equivalent to 400GE/800GE but slightly more latency.

#### 3. DLB/Flowlet & Packet Spraying

Split Elephant Flows overcomes the limitations of ECMP.

✓ Spectrum-X/DLB achieves 90% > line utilization.

#### 4. No Vendor Lock-in & Cost Optimization

Leverage open supply chains: Cisco, Juniper, Arista, Dell.

✓ Freedom of choice of industry-standard optical transceiver.

**ADVANTAGES:** Open ecosystem, TCO budget optimization, company flexibility.

**CONS:** Complex operation, requires intensive SDN tools.

Suitable: Cloud-scale AI clusters (Cloud Hyperscalers, Meta, Tesla, etc.).



# AI DC NETWORK MASTER ARCHITECTURE

North-South Traffic

## Segment 02: Frontend Fabric - Communication Administration & Orchestration

### North-South Traffic

Connect AI server clusters to the Internet, Public Cloud and users.

### Kubernetes & Slurm Platform

Pushing Dataset data, loading Docker Container Images into the cluster.

### Standard TCP/IP protocol

Even though BGP/OSPF routing does not require RDMA.

### Multi-Tenancy Isolation

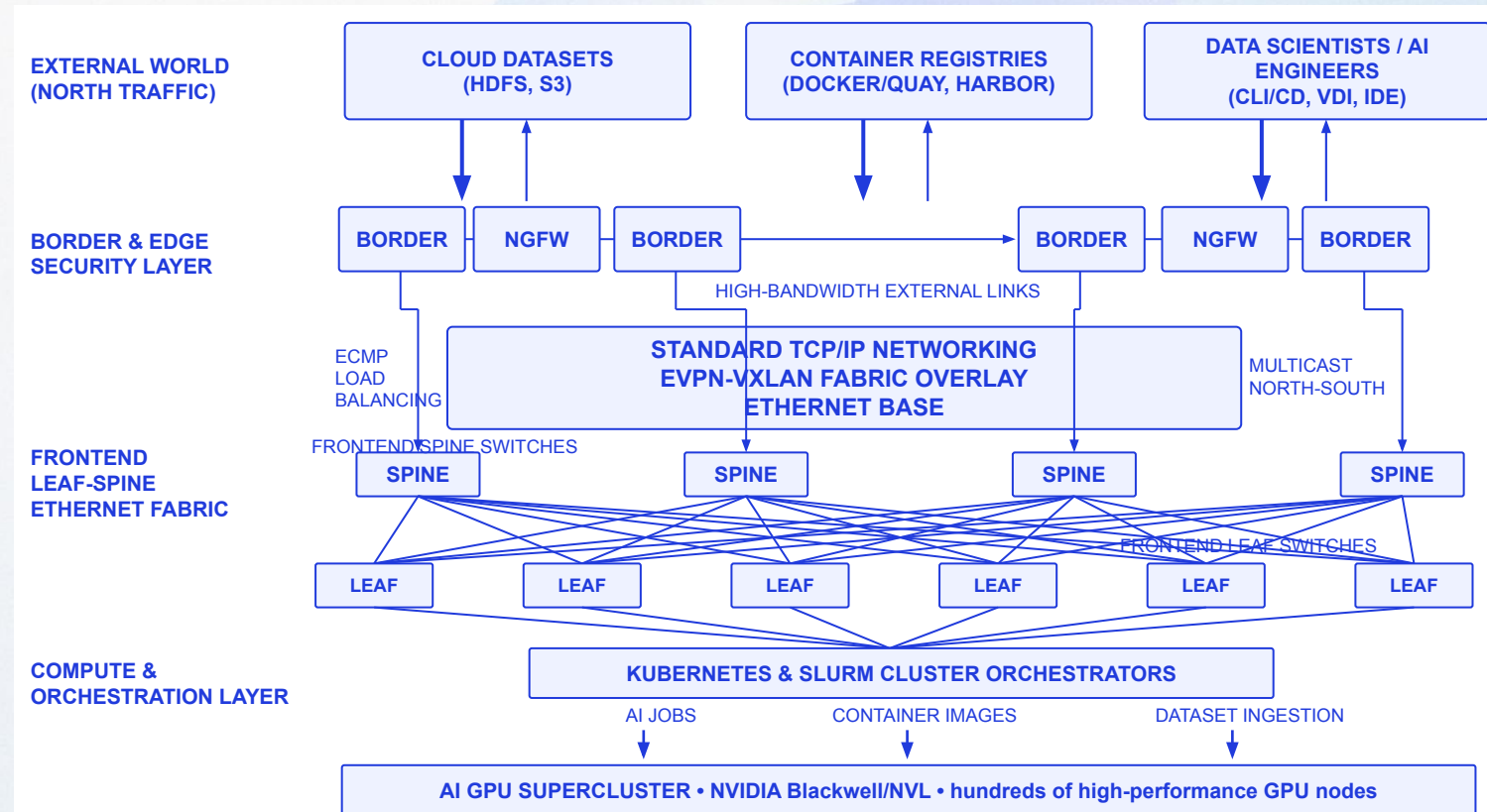
VXLAN and BGP EVPN completely separate VRF between projects.

### Built-in next-generation firewall (NGFW)

Protect training data and model weights against the risk of leakage.

### Zero Trust Policy

Strict control of each flow in/out of the center data.





# AI DC NETWORK MASTER ARCHITECTURE

## Segment 03: Storage Fabric – Starvation GPU Processing

### Starvation

The GPU calculates a hundred times faster than the speed of reading data from the hard drive.

### Save checkpoints in parallel

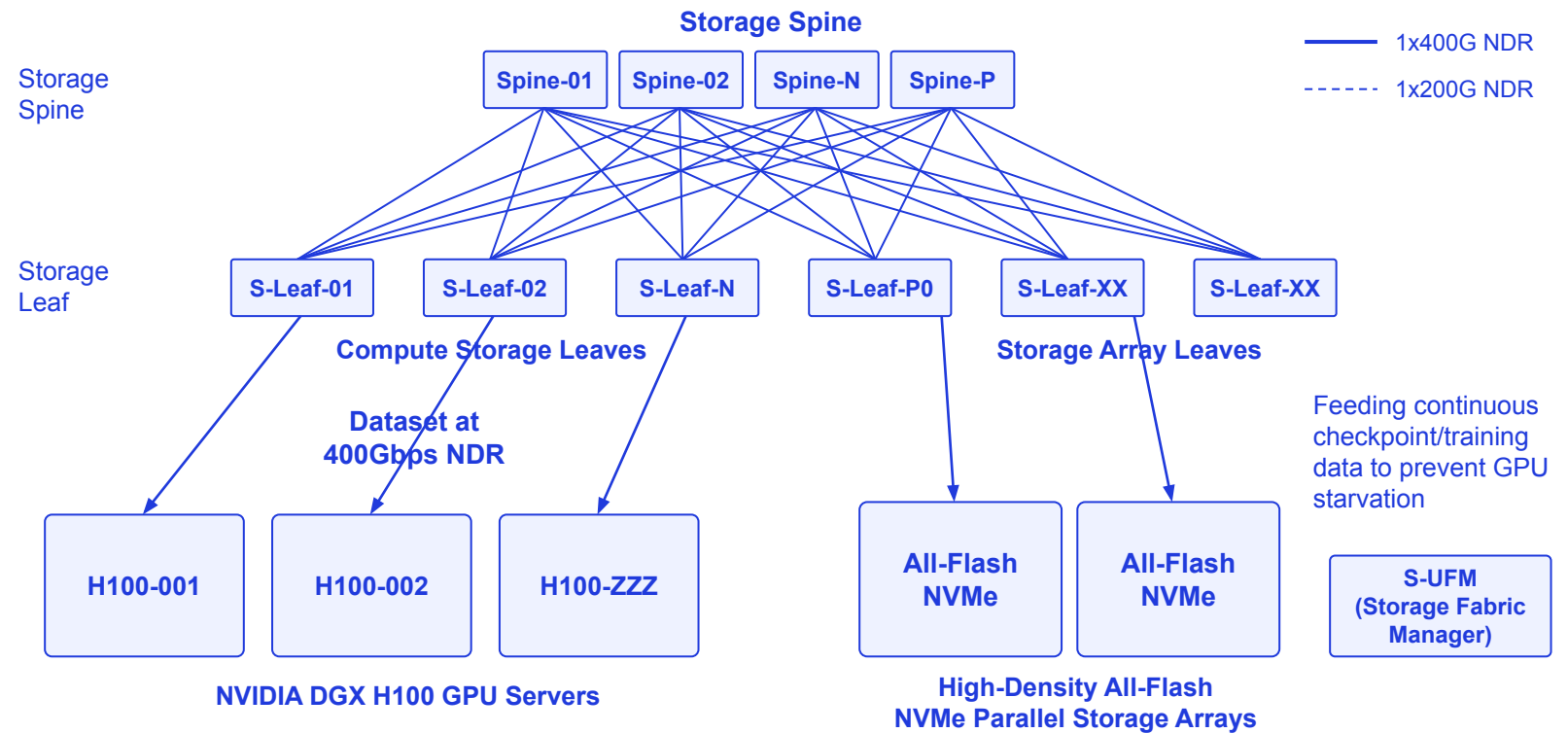
Parallel hundreds of GB/s of instantaneous write bandwidth when capturing model status.

### All-Flash NVMe Infrastructure

Parallel storage system - High Performance Storage (VAST Data, DDN, WEKA..).

### NVIDIA DGX H100/H200/B200/B300 High-Performance Storage Fabric

Multi-Topology. Eliminates GPU Starvation





# AI DC NETWORK MASTER ARCHITECTURE

## GPUDirect Storage (GDS) & NVMe-oF Technology

### NVIDIA GPUDirect

**Storage (GDS)**  
Enables direct data transfer between NVMe Storage and GPU VRAM over the network.

### NVMe over Fabrics

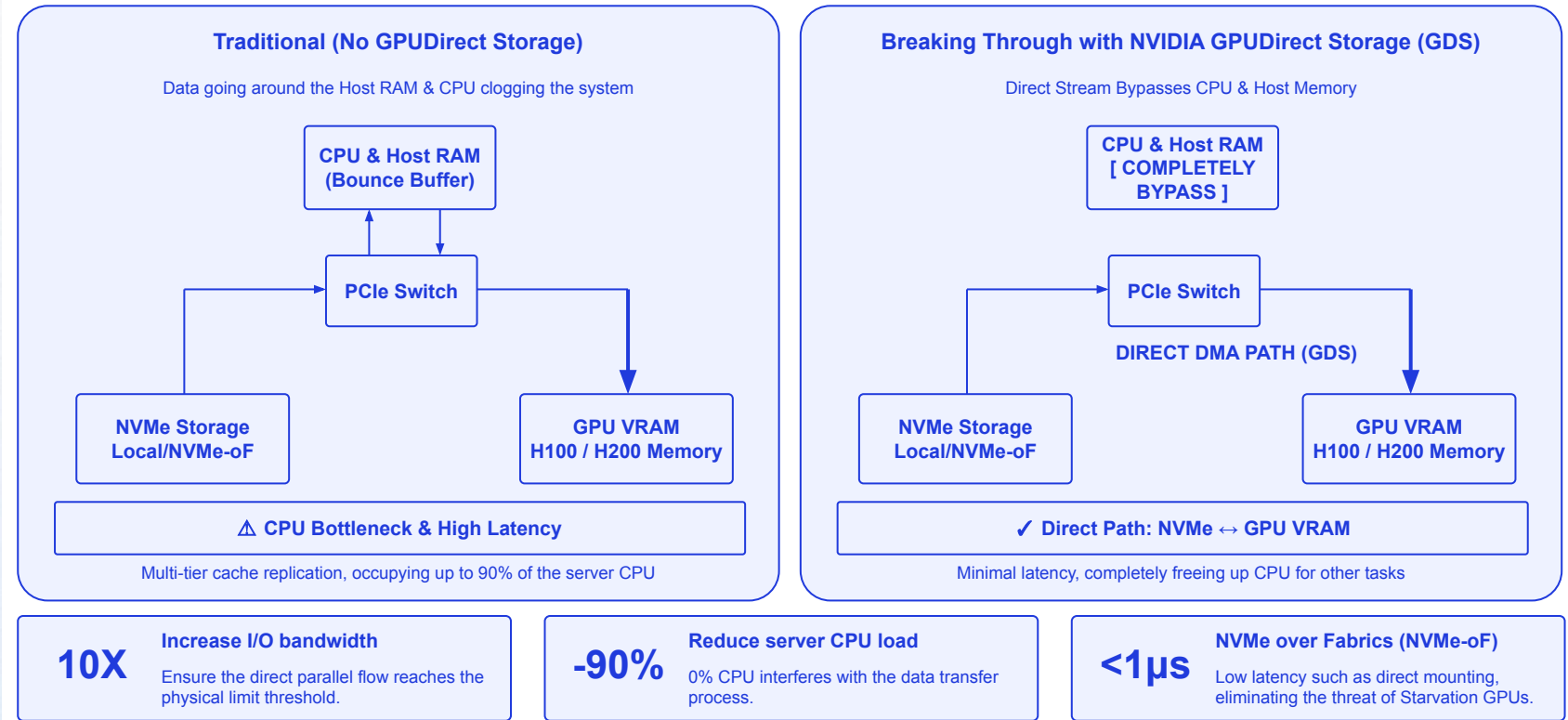
**(NVMe-oF)**  
Enables NVMe over Ethernet (E) or InfiniBand with low latency as internally mounted.

### 10x increase in I/O

**Bandwidth**  
Ensures 10x increase in CPU load and elimination of latency of latency load data.

## GPU Direct Storage (GDS) & NVMe-oF Technology

### STORAGE FABRIC ARCHITECTURE





# AI DC NETWORK MASTER ARCHITECTURE

## Segment 04: Out-of-Band (OOB) Networks & System Administration

### Complete physical isolation

100% isolation from compute and storage networks

### Hardware Level

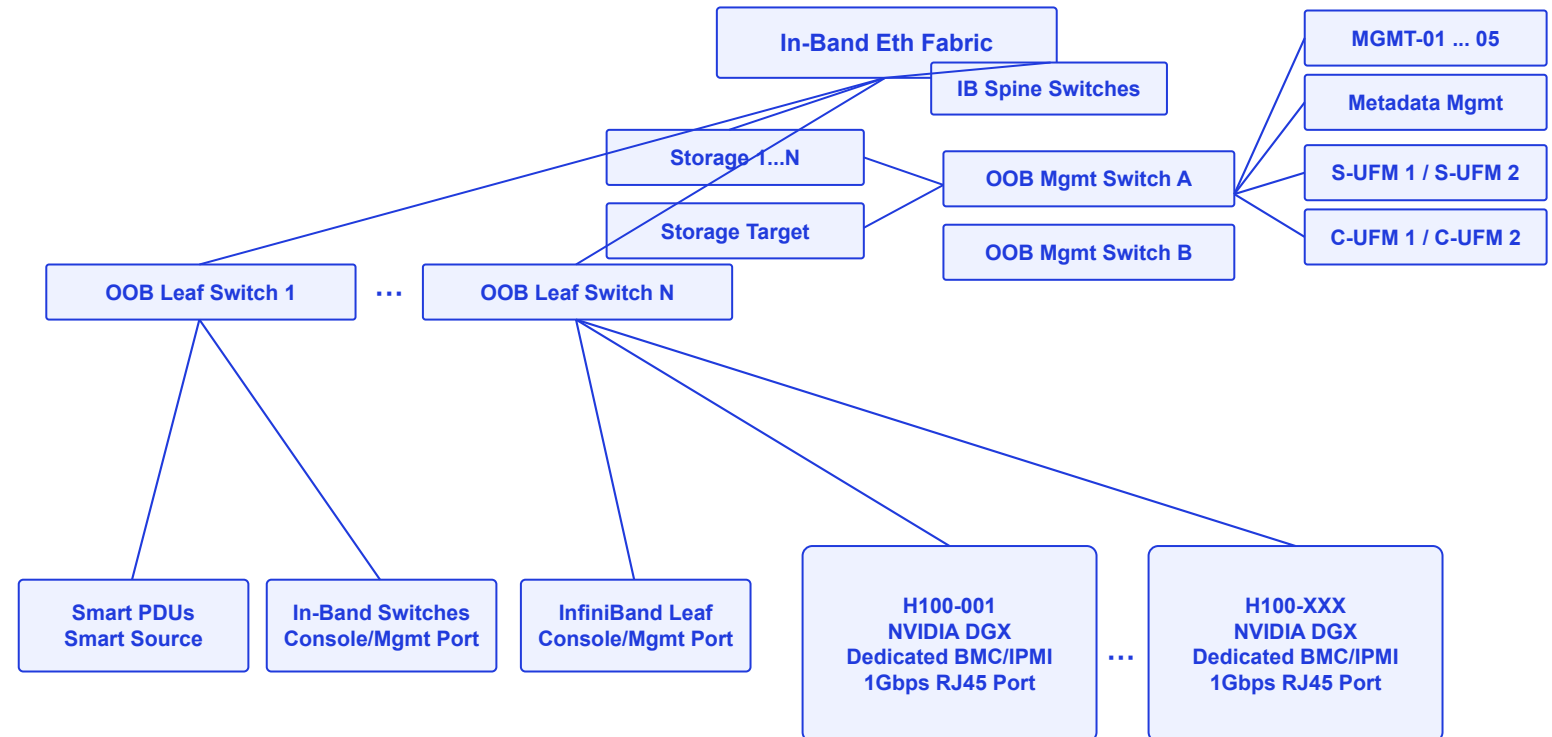
Control BMC, IPMI, iDRAC, Console switches, and smart PDUs.

### Rescue and load the operating system

Ensure control when the main data network cluster is isolated.

### Rescue Network Infrastructure & Independent OOB System Administration

#### OUT-OF-BAND MANAGEMENT NETWORK



#### ✓ Ensure absolute rescue control:

- Remote Power Cycle via PDU/BMC
- Bare-Metal Provisioning, loading new operating system over OOB network
- Operates 100% independently when the backend or frontend crashes

— 1 x 1 Gbps RJ45: OOB (Isolated) Admin Network Cable

- - - - - 100G/200G/400G: In-Band Data Fabric (Data Traffic)



# AUTOMATIC OPERATION AND ADMINISTRATION

## Challenges of operating AI factories

### Huge scale of fiber optic cables

Tens of thousands of DAC/AOC/MPO cables, the risk of plugging in the wrong port is very high.

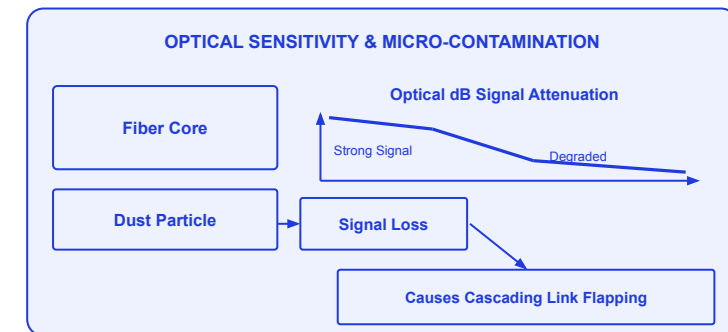
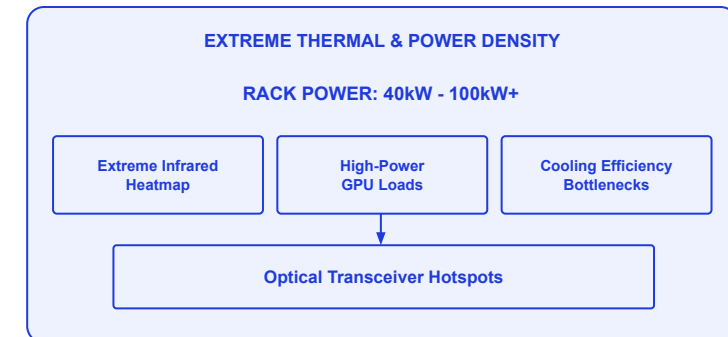
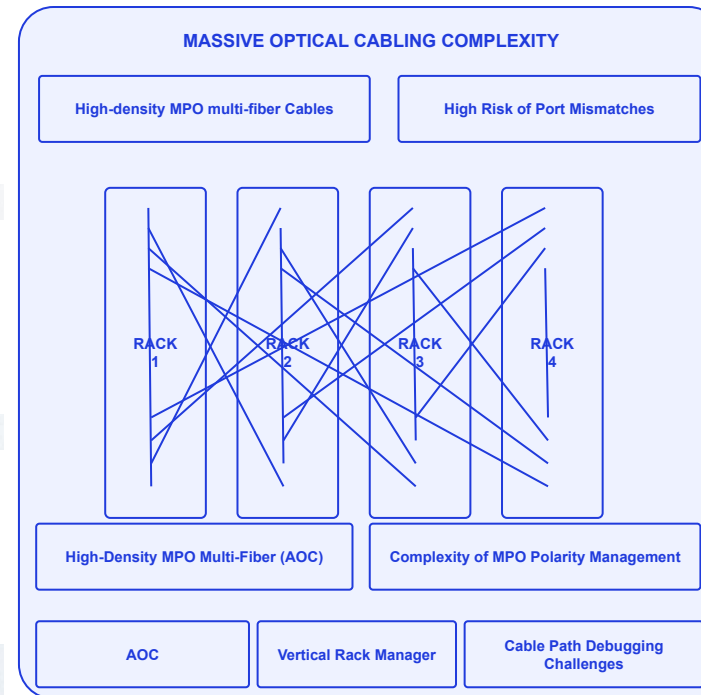
### Heat Dissipation & Capacity Challenges

400G/800G optics emit a lot of heat, requiring precise cooling of each rack.

### High optical sensitivity

Micro-dust on the connector causes signal loss and continuous port drops.

### CRITICAL OPERATIONAL CHALLENGES OF AI FACTORIES & AI DATA CENTERS



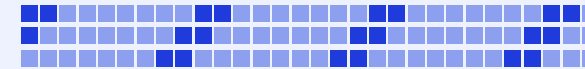


# AUTOMATED OPERATION & ADMINISTRATION

## AI Network Monitoring: Telemetry & Deep Visibility

### AI NETWORK MONITORING: TELEMETRY & DEEP VISIBILITY IN AN AI DATA CENTER

AI NETWORK TELEMETRY  
BUFFER/LATENCY/PACKETS



#### Traditional SNMP is no longer relevant

The polling cycle with a 5-minute interval misses all of the millisecond micro-bursts of AI.

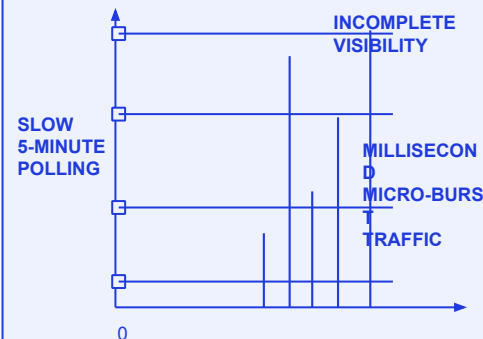
#### Streaming Telemetry

Instantly push buffer occupancy, drop packet, and latency metrics by sub-second.

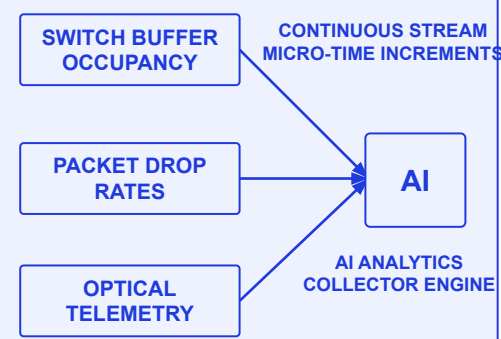
#### Inband Network

Telemetry (INT) measure the delay of each hop on the path of the RoCEv2 packets.

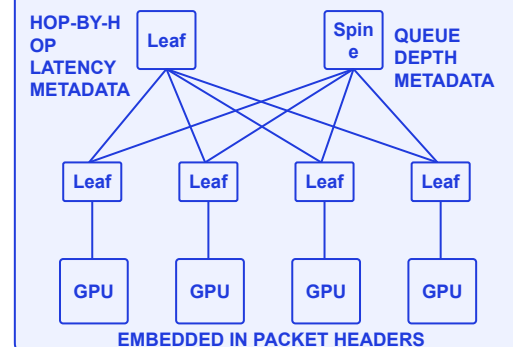
#### 1. TRADITIONAL SNMP INADEQUACY



#### 2. REAL-TIME STREAMING TELEMETRY



#### 3. INBAND NETWORK TELEMETRY (INT)



#### DEEP VISIBILITY ECOSYSTEM

SYNTHESIS AI DATA

NETWORK ANALYTICS

TELEMETRY INSIGHTS

OPERATIONAL INSIGHTS



# RECOMMENDATIONS FOR AI DATA CENTER IN VIETNAM

## 1. Design by Block (SU)

### Standardization & Expansion

- Start building AI Factory according to the Scalable Unit (SU) standard.
- Makes it easy to calculate CAPEX/OPEX costs, power supply, heat dissipation and floor area .
- Ensures the ability to upgrade to next generation of GPUs without breaking the design.

## 2. Choose the right technology

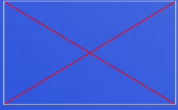
### InfiniBand vs RoCEv2

- Choose InfiniBand if building a monolithic AI Supercomputer for your organization, leveraging Nvidia's ecosystem.
- Choose RoCEv2 (Ethernet) if you want to build a flexible AI Cloud infrastructure that is not locked by vendors.
- Always make sure the DPU/NIC network card is in sync with the switch.

## 3. Invest in automation & monitoring software

### Telemetry & Automation

- AI networks cannot be manually administered using traditional CLIs.
- Synchronous investment in automatic telemetry solutions.
- Proactively detect network congestion and isolate errors before they affect AI Jobs.



# Thank you very much to the delegates and the Organizing Committee of VNIX-NOG 2026!

## 04 Key takeaways:

1. AI network infrastructure is the determinant of performance: Lossless, Rail-optimized and ultra-low latency networks to promote GPU power.
2. Separation of 04 independent network modules: Separation of Compute, Storage, In-Band and OOB helps eliminate the risk of I/O disputes and ensure system safety and stability.
3. Leverage a Standardized Reference Architecture: Applying the Physical Twin philosophy and Scalable Unit (SU) for fast deployment, easy scalability, and reliable operation.
4. Automation is Required: Manual operation is not feasible at the scale of an AI Factory

