Networking For Gen AI

Marco Redaelli – Datacenter Sales

Dell Technologies

Agenda

- Gen AI Networking Intro
- AI/ML Fabrics
- IB or Ethernet for GPU Back-End?
- Design best practice
- Nvidia Competitive

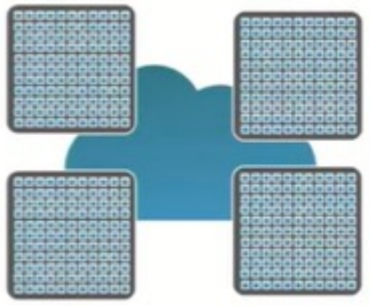
Evolution of Compute

CPU
Optimized for Serial Tasks

Multiple Cores

GPU
Optimized for Parallel Tasks

Thousands of Cores



GPU Clusters
Scale-up GPU Clusters for AI workloads

Tens of Thousands of Cores

Generative AI Models

BARD AI

OpenAI

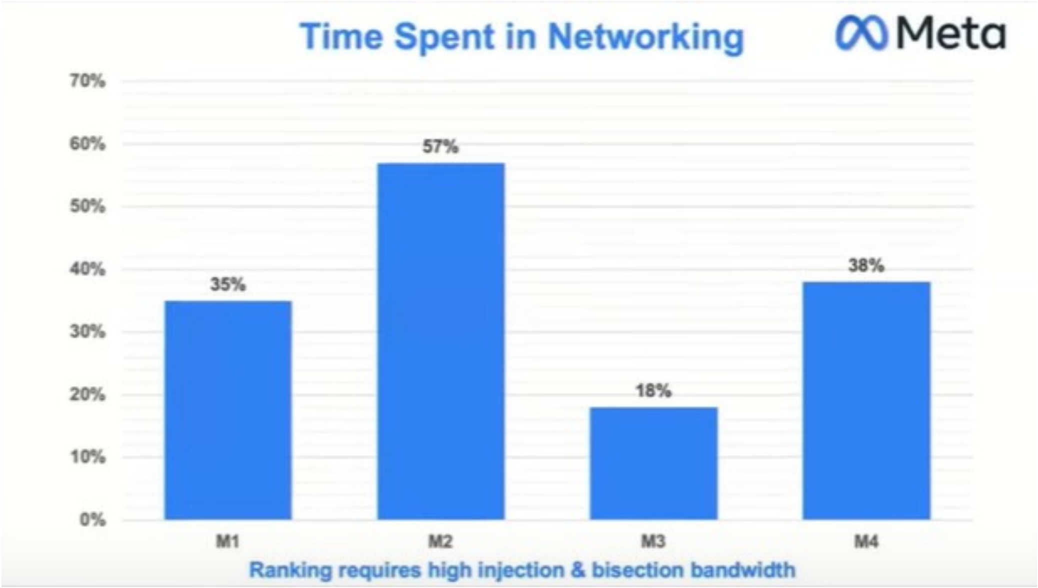
LLaMA by Meta

BLOOM

DELL Technologies

What makes AI Networking unique?

- GPU to GPU Communication Drives higher bandwidth flows
- Bursty traffic
- Links are saturated in Micro-seconds
- Training jobs run for long periods
- Tail latency impacts job completion time



OCP Keynote by Alexis Bjorlin at 2022 Global Summit

Gen AI Workloads

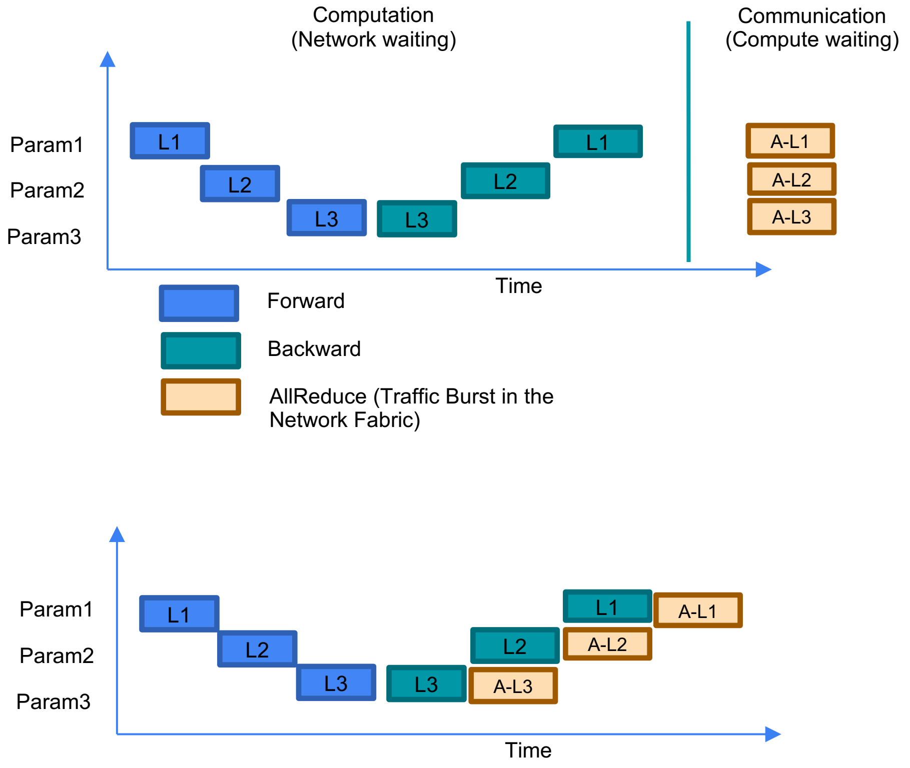
Training

Fine Tuning

Inferencing

AI Training – Optimization - Data Parallelism

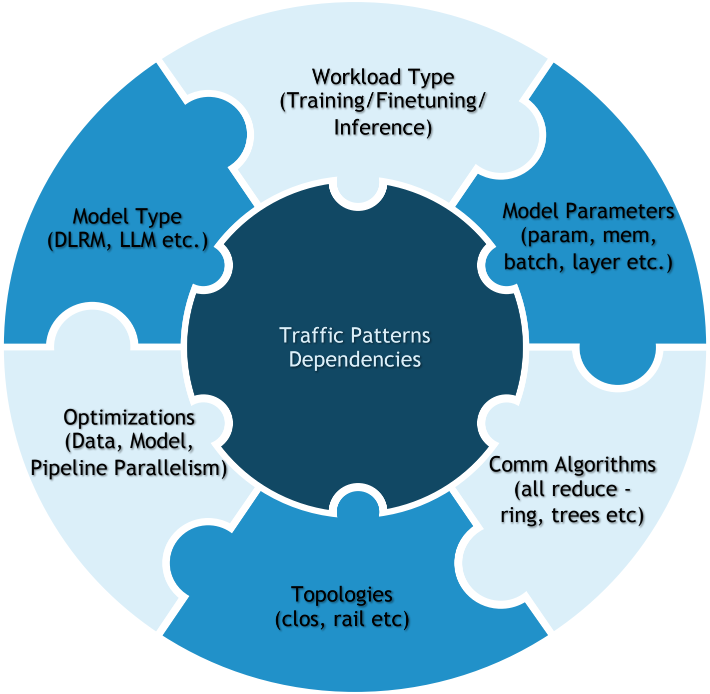
- Data Parallelism
  - Large data batches are divided into multiple mini-batches
  - Training performed in parallel across GPUs using mini-batches
- Without optimization, the first stage network waits for computation to complete, and during the second stage, computation waits while communication is ongoing.
- Interleaving optimization enables efficient utilization of Compute and Network Resources. Communication can commence immediately after the completion of the L3 backward pass.



Data Parallelism Optimization by Interleaving

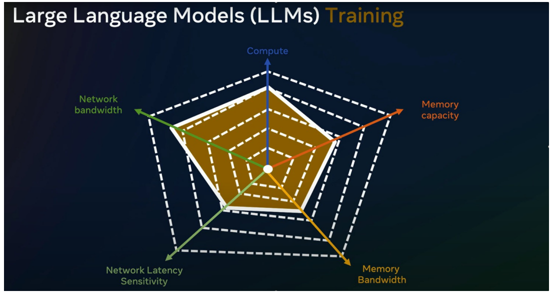
Gen AI Traffic Patterns

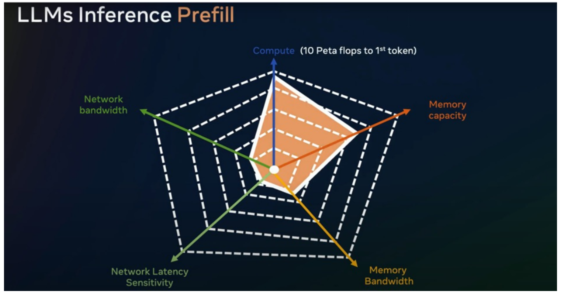
- Large volume of data exchanged.
- Traffic exhibits a diverse set of patterns.
- quasi-periodic – with peaks and valleys
- Highly ordered and predictable (training)
- Heterogenous –
    - Large flows (gradient, weight exchange)
    - Small flows (Ctrl)

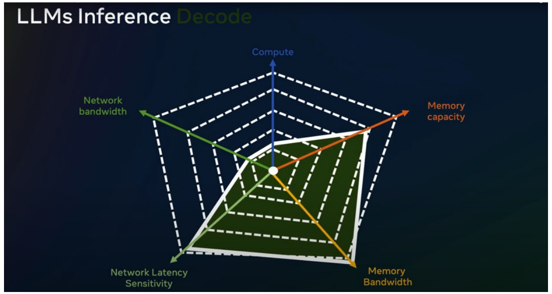


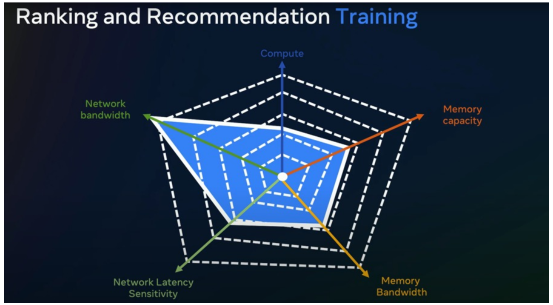
DELL Technologies

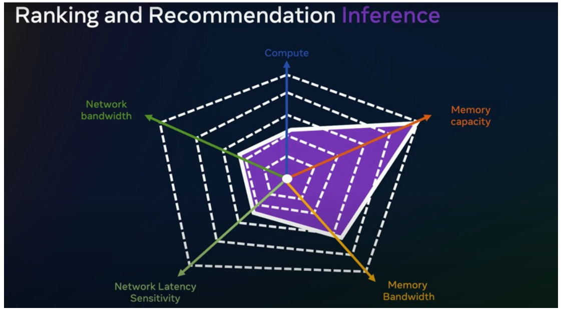
Gen AI Traffic Patterns (presented by Meta at OCP 2023 keynote)

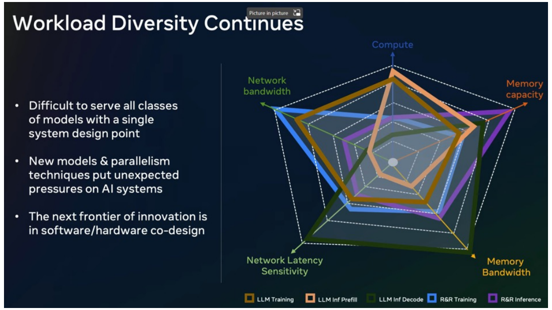












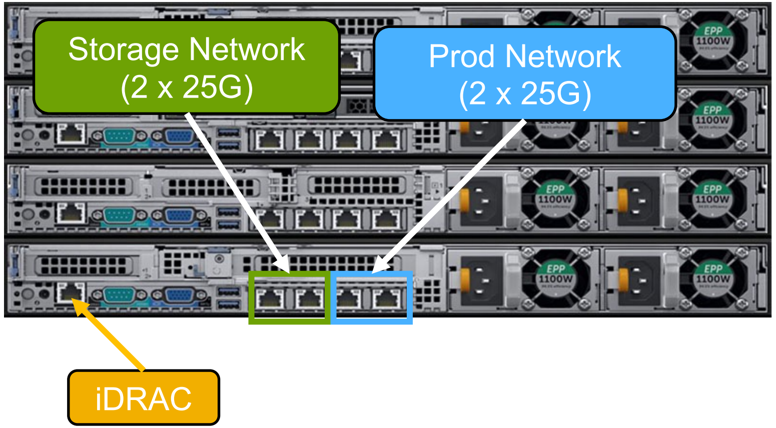
Ref: https://www.youtube.com/watch?v=dTeEwG2Bx-k

AI/ML Fabrics

DELL Technologies

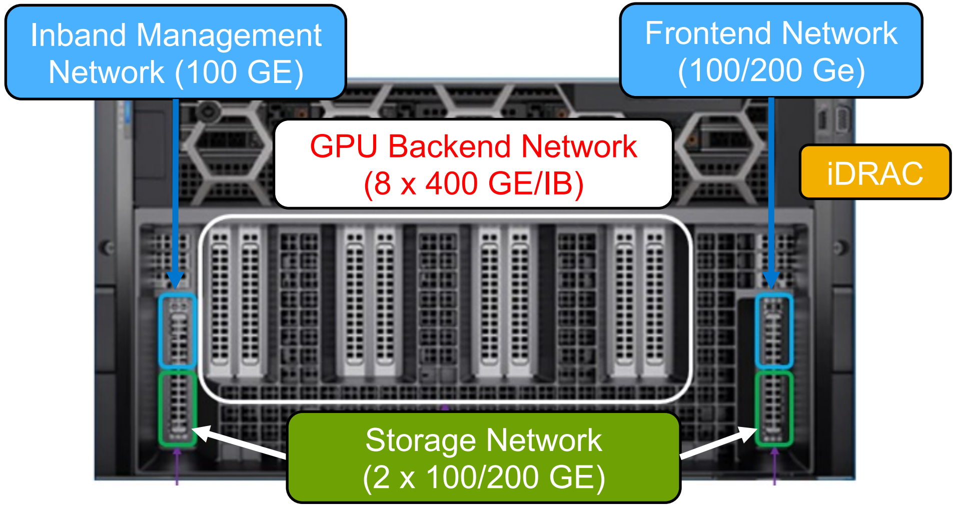
How Networking For Gen AI is different ?

DC “Classical” Network



- ❑ 2 Network Fabric @25G
- ❑ Design with oversubscription

Gen AI Network



- ❑ 3 Network Fabric @400G & @100G
- ❑ “GPU Backend” not oversubscribed
- ❑ “GPU Backend” not redundant

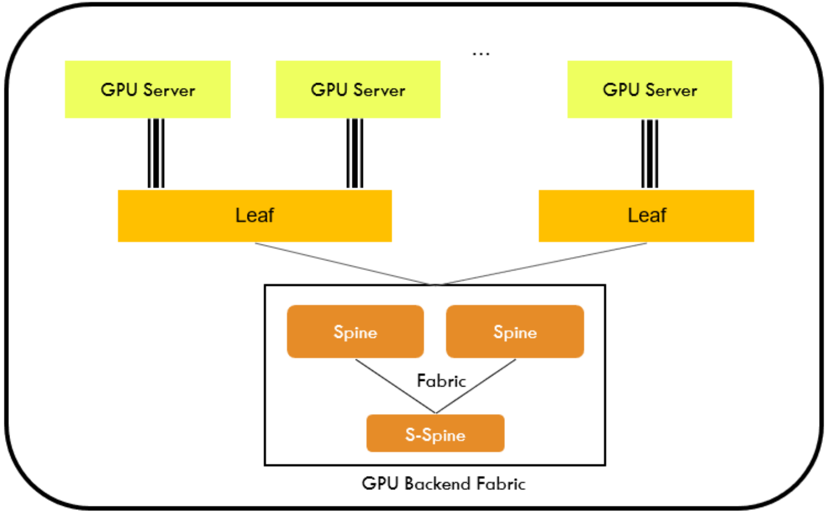
GenAI Infrastructure Building blocks – **Compute Backend (GPU) Fabric**

- **Objective:** GPU to GPU connectivity to execute an AI/ML training or inference job. This fabric is where GPUs are going to perform hyper-parameter optimizations
- **Fabric Highlights**
  - Dedicated fabric for GPU <-> GPU communication.
  - Model training and inferencing traffic
  - Ethernet solutions evolving as a preferred choice
  - Performance approaching InfiniBand specs
  - Each GPU-Server will have 8x400G or 8x(2x200G) connectivity to leaf switches.
  - NIC is connected to GPU & CPU
  - Software Requirements :

    Lower latency

    High Radix switches

    Lower tail latency

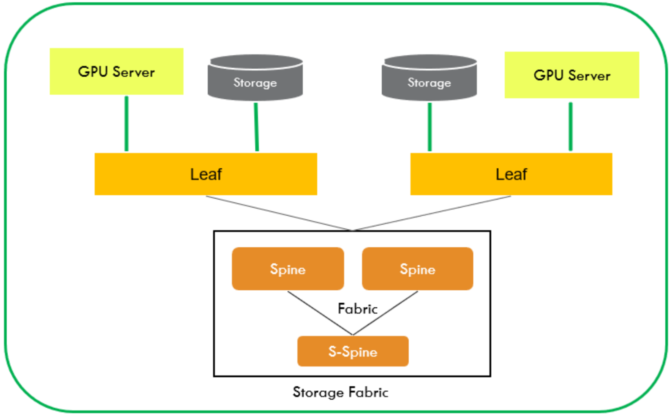


DELL Technologies

GenAI Infrastructure Building blocks – **Frontend Fabric**

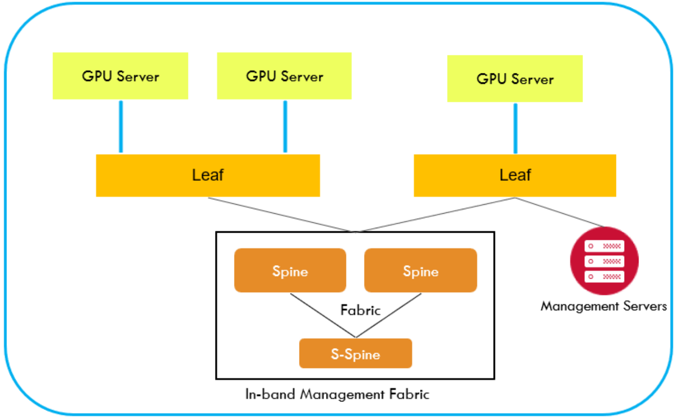
Storage Fabrics

- **Objective:** Storage fabric provides access to large-scale shared storage infrastructure. This storage is used as a shared resources for GPUs to communicate hyper-parameters during AI/ML training or inference jobs
- **Fabric Highlights**
  - Fabric for GPU to storage server communication.
  - Typically, 25G/100G connectivity with **ethernet solutions**



In-band/Access Management

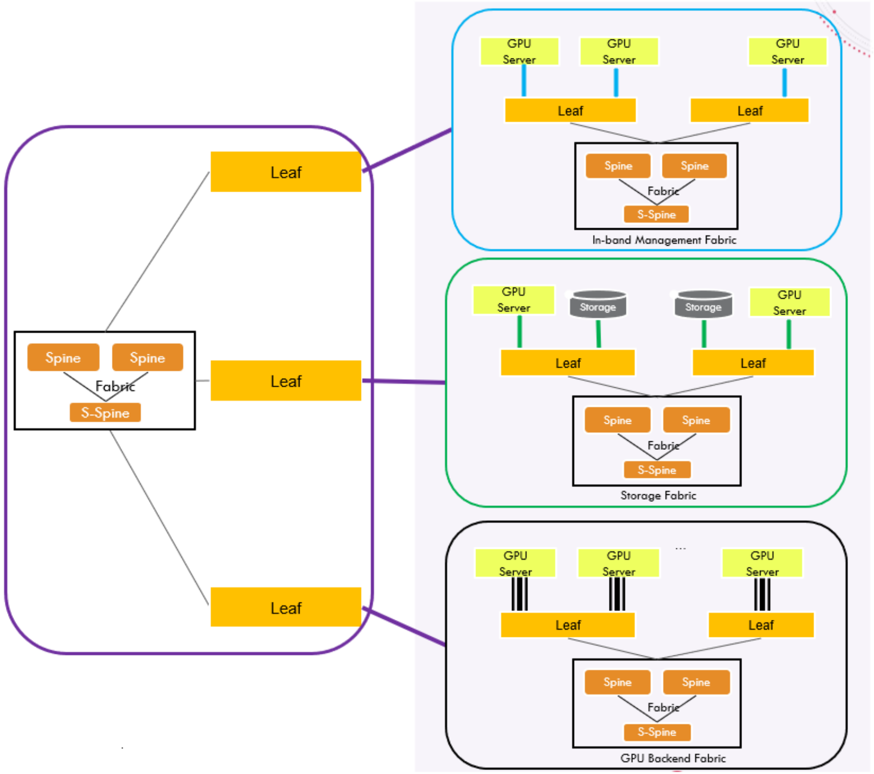
- **Objective:** This fabric is used to distribute the AI/ML jobs on to the Data Center back-end network on GPU. In-band Management prioritizes, batches, and provides/allocates the necessary resources (GPUs, Storage, Network) for AI/ML applications.
- **Fabric Highlights**
  - Fabric for managing the AI/ML jobs assignment on GPUs
  - Typically, 25G/100G connectivity with **ethernet solutions**
  - Multitenancy use-cases



D&LLTechnologies

GenAI Infrastructure Building blocks – **OOB Use case**

- **Objective:** OOB Fabric provides management for GPU/Storage servers, Ethernet / InfiniBand switches, appliances (firewall, load balancer) etc.
- OOB network on server use iDRAC interface to read temperature/thermals, CPU/GPU utilization, miscellaneous sensor information.
- 1G **ethernet** connectivity solution with basic L2/L3 features



DELL Technologies

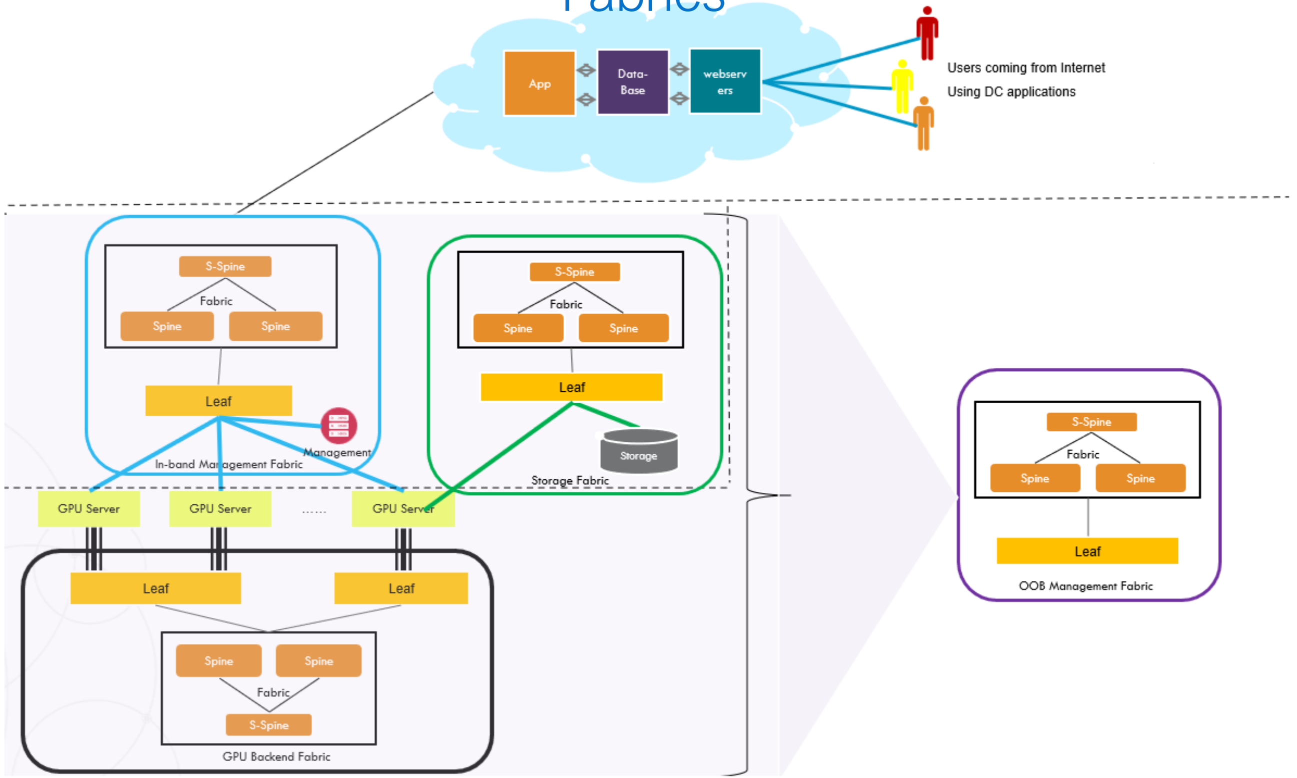
Bringing it ALL together – AI fabric

Back-End (GPU Fabric) has most demanding requirements for raw performance, lossless attributes and lowest latency

Front-End fabrics support application traffic, storage access and connection to the general network

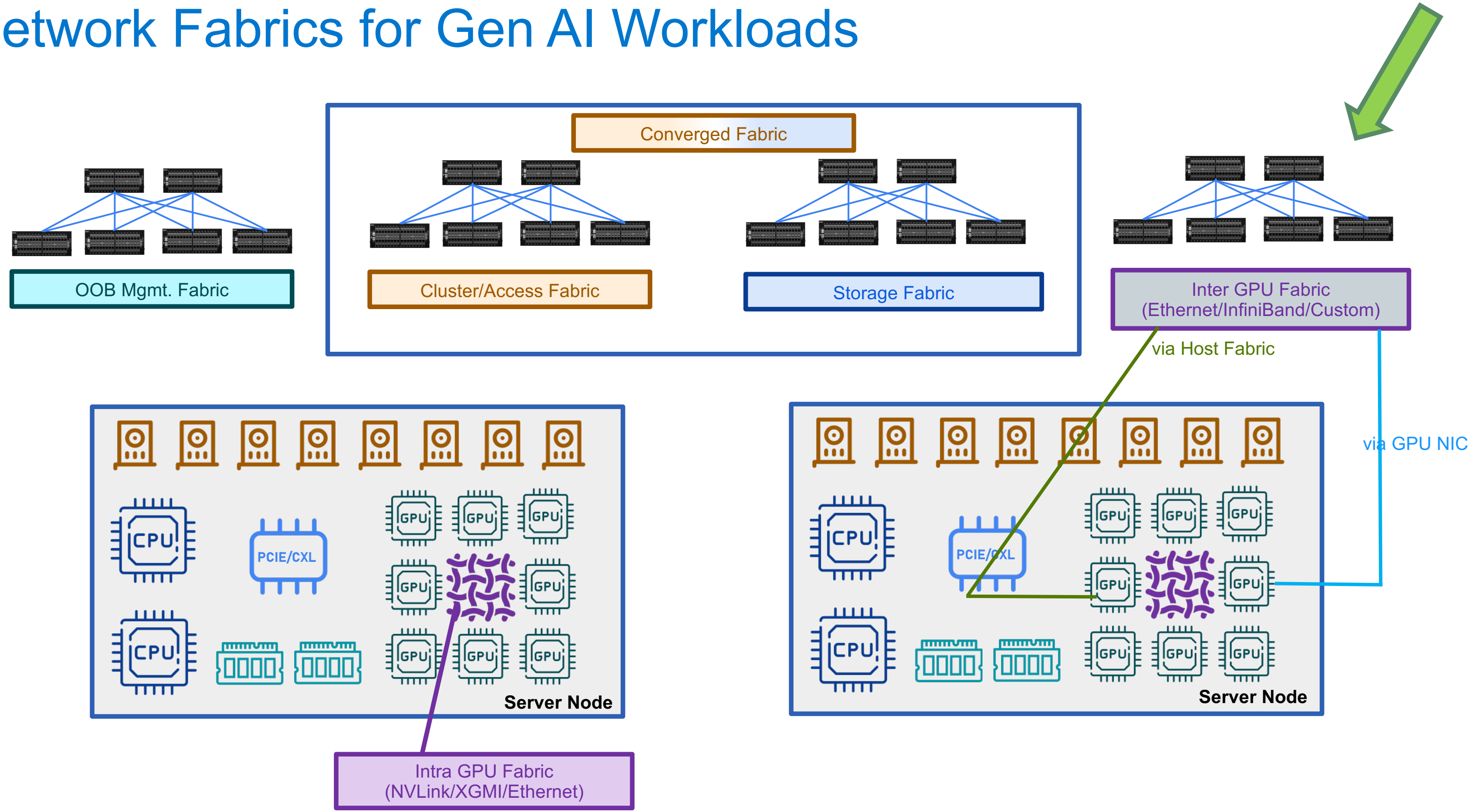
OOB Mgmt Network for administration and fabric management

Delivering Ethernet Solutions across all use cases within AI Fabrics



DELL Technologies

Network Fabrics for Gen AI Workloads



DELL Technologies

IB or Ethernet for GPU Backend ?

DELL Technologies

Ethernet evolving to be the preferred choice for backend AI fabrics

- Market inflection points for Ethernets powered by AI fabrics
  - **Availability** of High Radix switching with next-Gen silicon technologies – 64x400G (25.6T), 64x800G(51.2T), 102.4T…
  - Improved **congestion monitoring, flow control, and Transport (RoCEv2)** protocol availability in NOS
  - Community effort to drive Ethernet Standards – **Ultra Ethernet Consortium**
  - Desire for **no-vendor lock-in** infrastructures
  - **Silicon and supplier diversity**
  - Lower **Total Cost of Ownership (~3x lower)**
  - **Latency improvements** with next Gen Silicon **from 800ns to 200ns**

DELL Technologies

GenAI Fabric requirements

- Data Intensive – High Injection and Bisection Bandwidth
- High sustained traffic – Links are saturated in Micro-second
- Low entropy flow identification for carrying RDMA Messages
- Lossless Network
- Tail Latency Sensitive – Job Completion Time
- Drop and Order Sensitive
- Optimized Topologies
- Latency Important for Inferencing

DELL Technologies

Ethernet for GPU Backend : Server Side

- ✓ RoCE (RDMA over Converged Ethernet) enables RDMA (Remote Direct Memory Access) encapsulation over Ethernet. RoCE transport is fully supported by the Open Fabrics Software since OFED 1.5.1.
- ✓ InfiniBand natively supports RDMA encapsulation

→ RDMA encapsulation happen in the Network Card of the servers in hardware so **no performance gap between Ethernet and IB**

→ **Application layer is not aware** of the underlying encapsulation method

Ethernet for GPU Backend : Switch Side

GPU Fabric is not redundant

Scale

Large and non-blocking Network fabric

→ Spine-Leaf design
→ High port count switches

Speed

The fastest the better

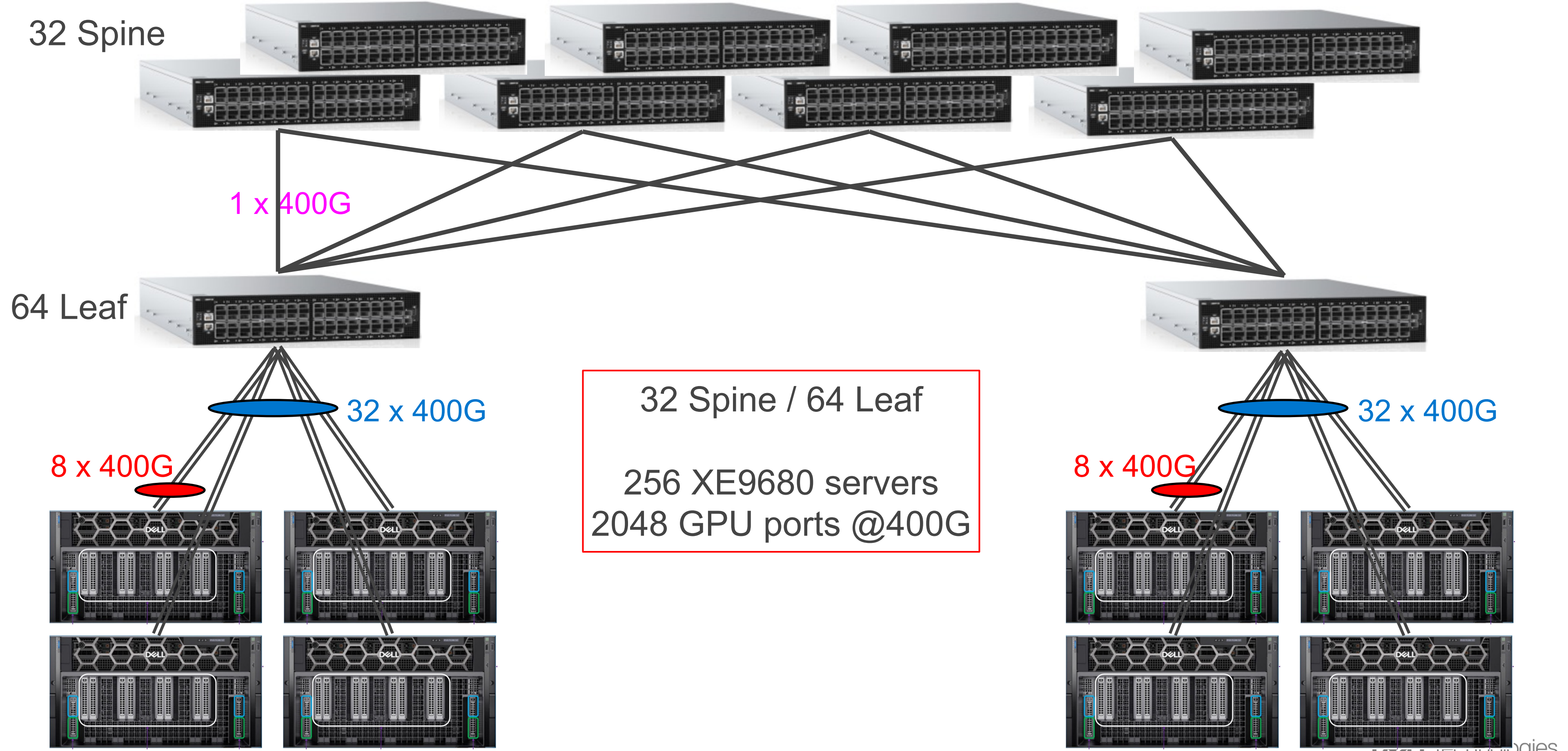
→ 400G Eth @ server
→ Manage "elephant" flow

Latency

Is switch latency impacting the global performance ?

→ Eth switch latency is 0,4 µs versus 0.2 µs for IB

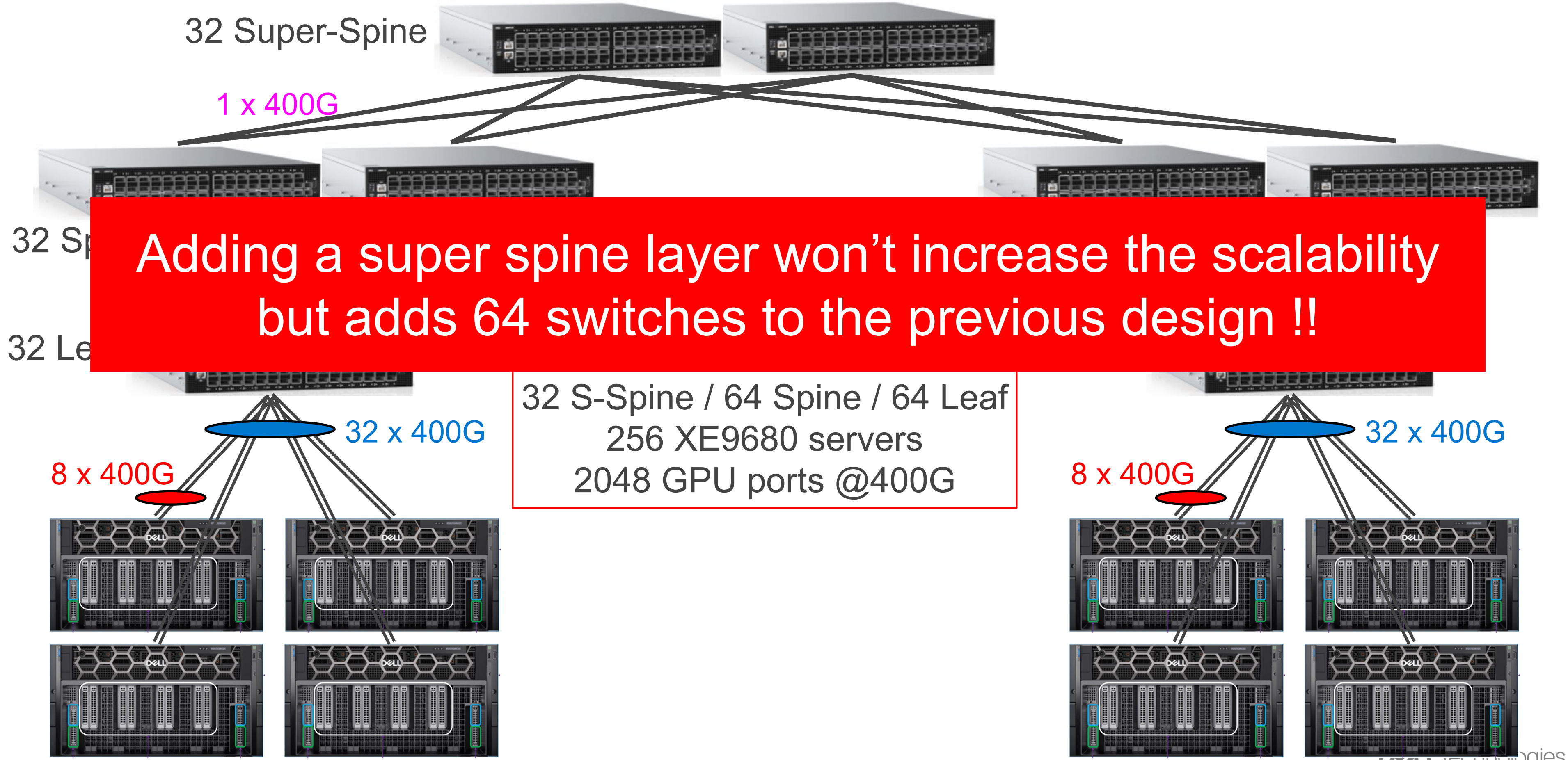
Max scale with Z9664F



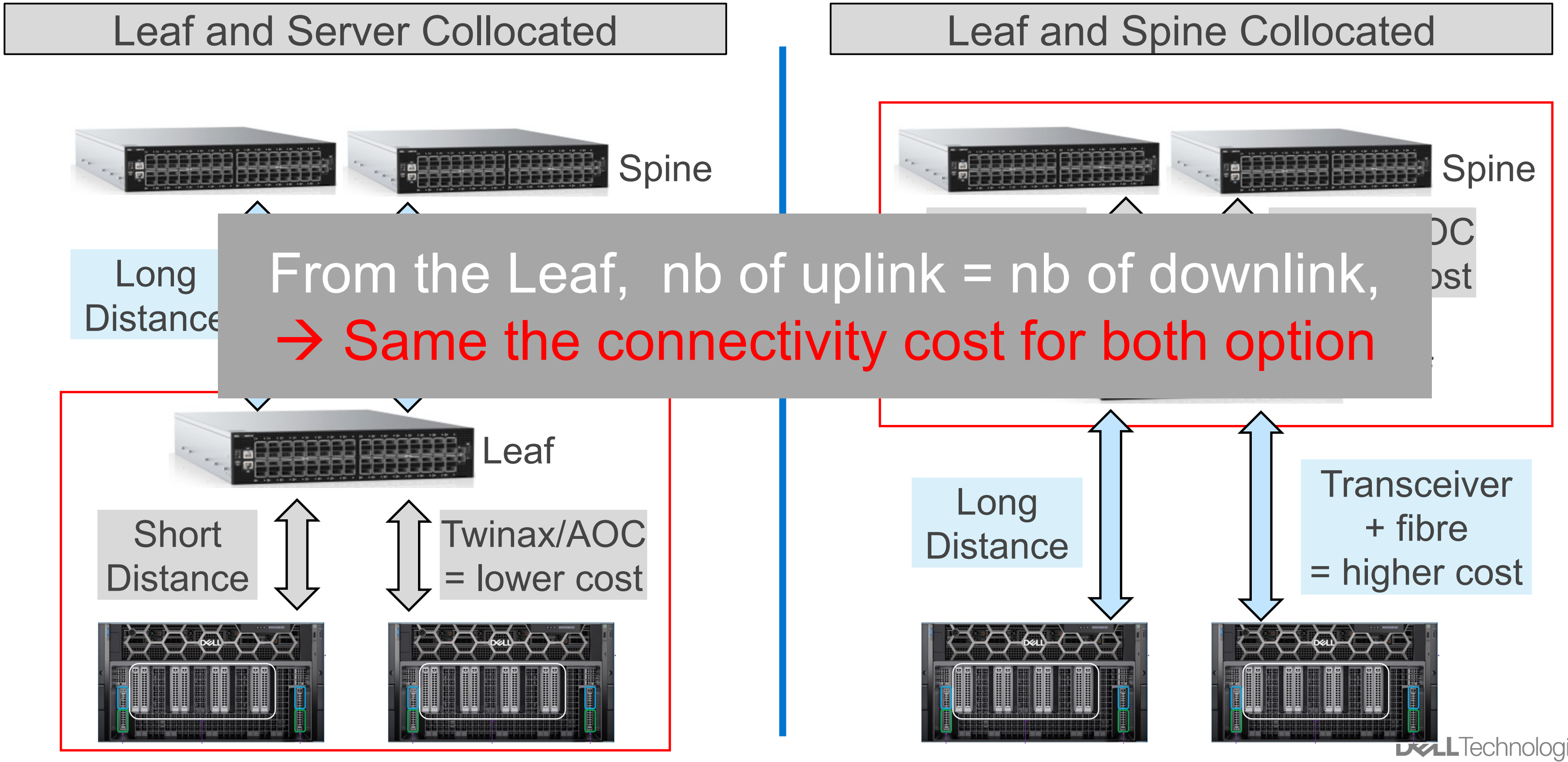
32 Spine / 64 Leaf

256 XE9680 servers
2048 GPU ports @400G

Does adding a "Super Spine" layer help ?



What about my Datacenter Urbanization ?



From the Leaf, nb of uplink = nb of downlink,

→ Same the connectivity cost for both option

Networking Criteria in Gen AI

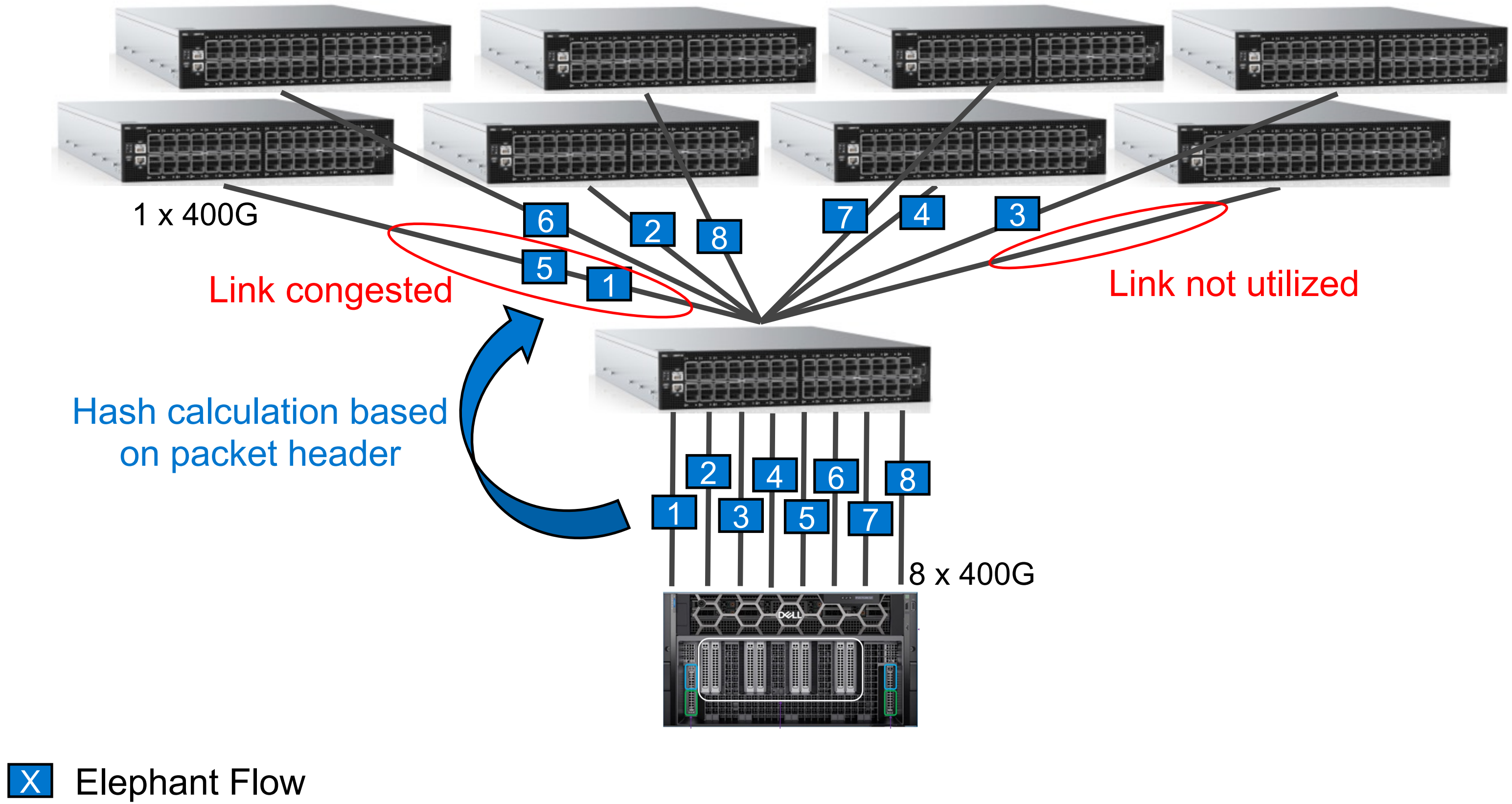
Speed

The fastest the better

→ 400G Eth @ server
→ Manage "elephant" flow

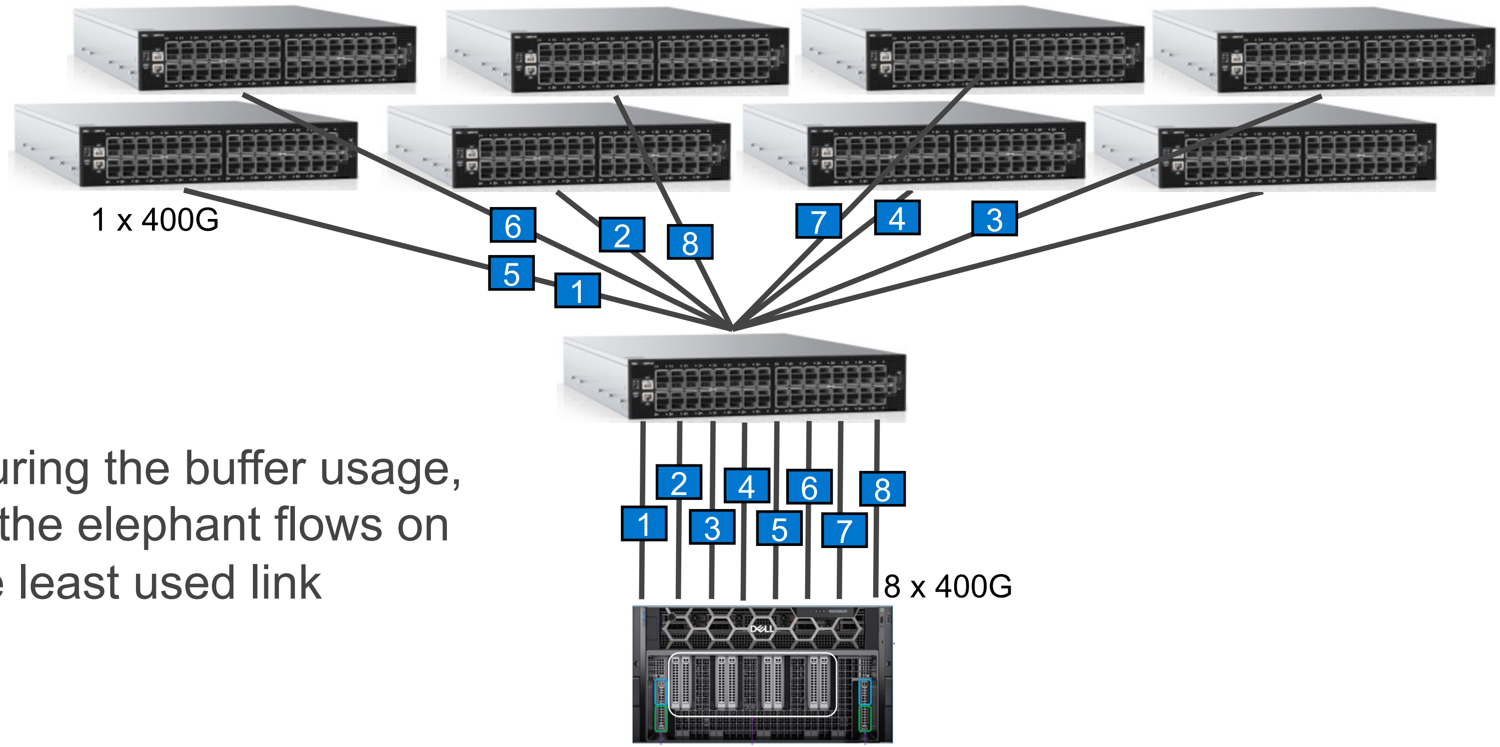
Dynamic Load Balancing (DLB)

Problem statement



Dynamic Load Balancing (DLB)

Solution



By measuring the buffer usage, DLB put the elephant flows on the least used link

Networking Criteria in Gen AI

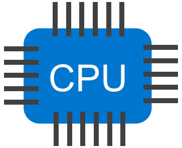
Latency

Is switch latency affecting the global performance ?

→ Let's deep dive into switch latency impact …

Latency in HPC and GPU context

HPC



3 → 20 µs

1 µs → 5 µs

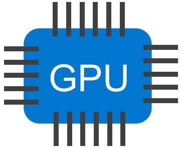
0,05 → 0.5 µs

Ethernet : 0,4 → 0.8 µs
InfiniBand : 0.2 µs

The switch counts for :

- 5% → 20% of the latency (Ethernet)
- 1% → 3% of the latency (IB)

GPU Interconnect



20 → 200 µs

1 µs → 5 µs

0,05 → 0.5 µs

Ethernet : 0,4 → 0.8 µs
InfiniBand : 0.2 µs

The switch counts for :

- 0.6% → 5% of the latency (Ethernet)
- 0.2% → 1.4% of the latency (IB)

How to improve latency ?

Cut-Through Switching improves latency by 20%
Cut-Through switching is already supported

For Z9664F :

- ❑ Store and Forward latency : 946 ns → 1054 ns
- ❑ Cut-Through latency : 709 ns → 867 ns

Z9864F latency should be around 400ns

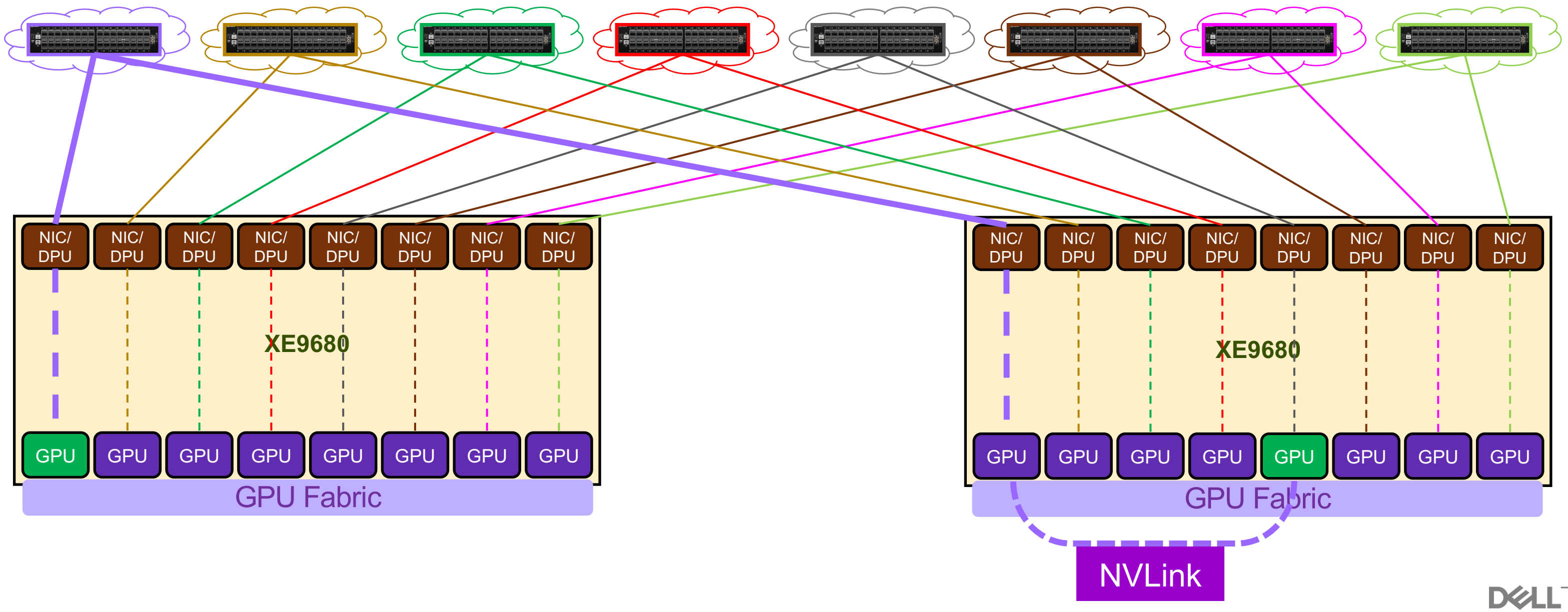
Network Design Best Practice

DELL Technologies
DELL Technologies

“Rail” Design (Nvidia GPU only)

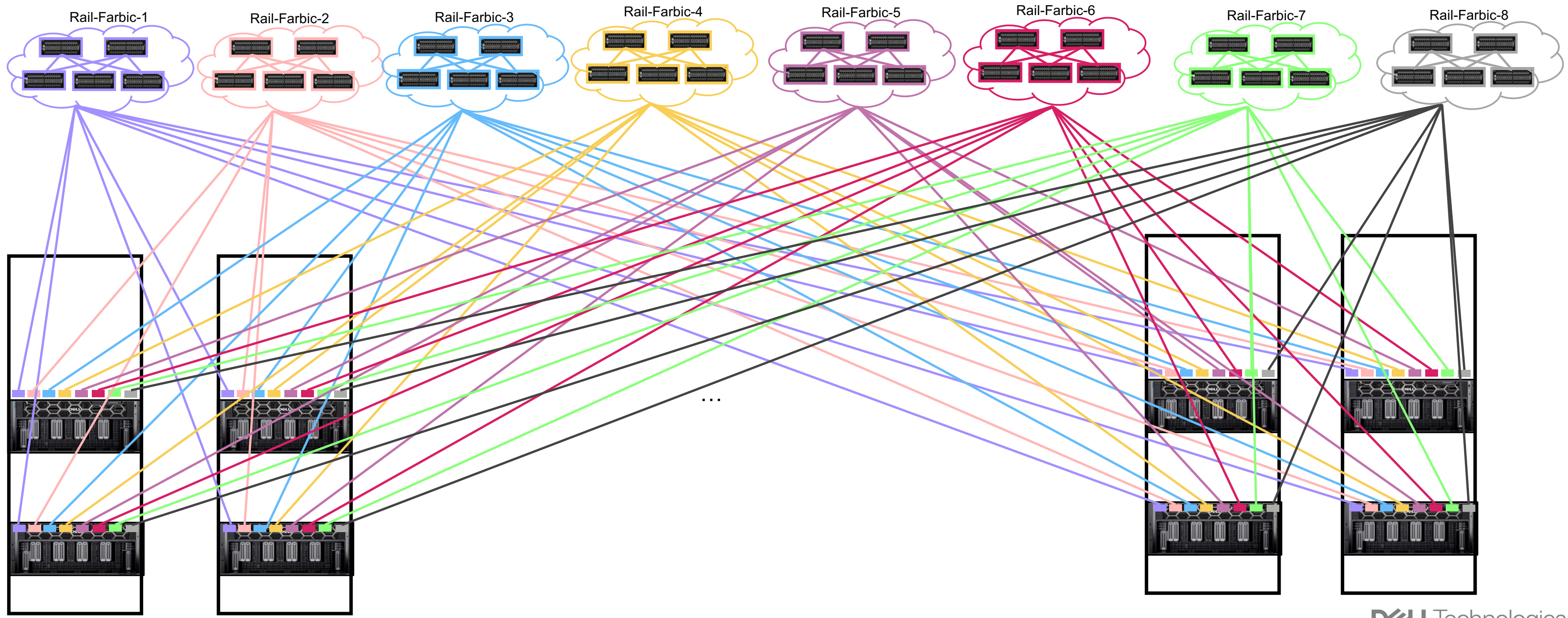
“Rail” design optimizes the GPU interconnect by leveraging “NVLink” feature available on Nvidia NIC that provides direct GPU-to-GPU communication path within the servers.

Building 8 separated network fabric for each “Rail” instead of a single fabric for all GPU ports



“Rail” Design with high scale

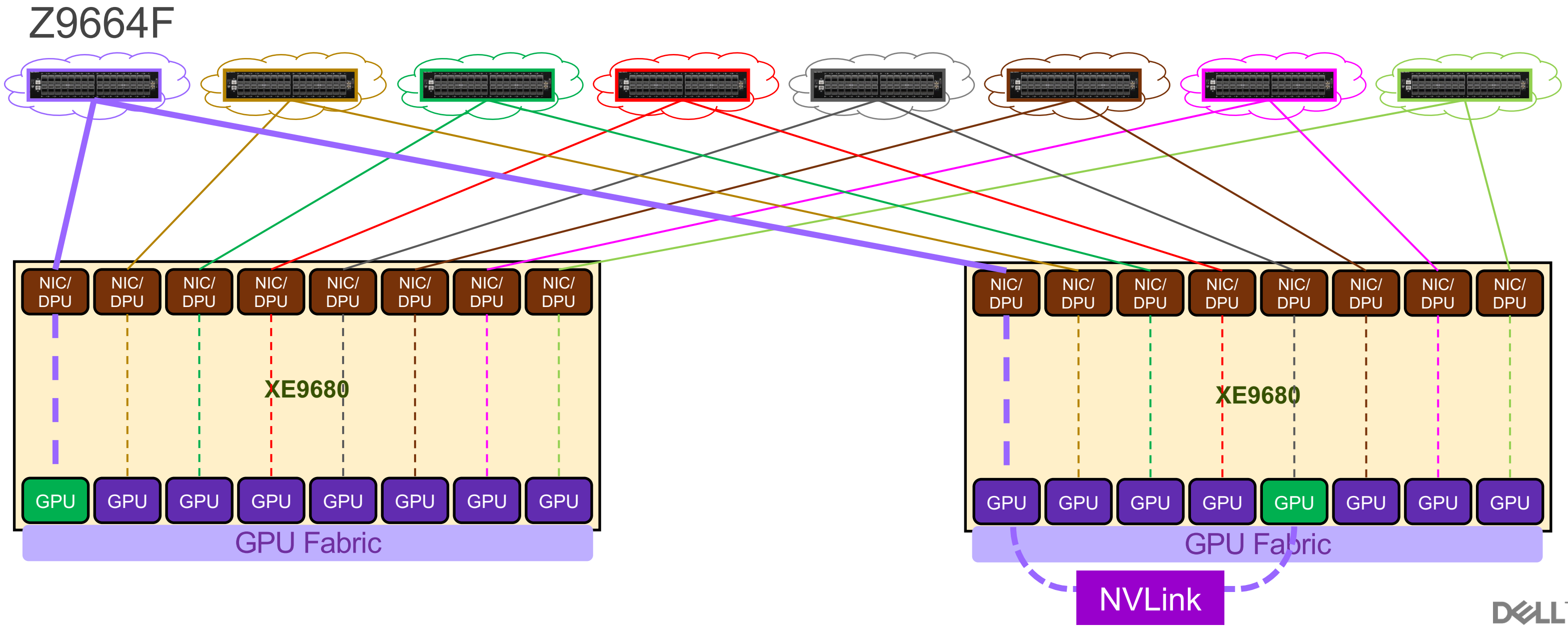
By leveraging “Rail” design, you can increase the max scale by 8 time !!



“Rail” Design for small / mid scale

A “Rail” design with a single switch per fabric can scale up to 64 XE9680 server / 512 GPU ports (400GE)

1 flow per link + single switch latency (instead of 3)



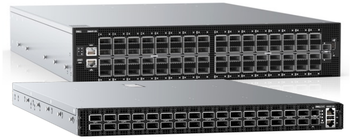
Dell PowerSwitch Portfolio (SONiC)

GPU Backend

800G

Z9864F — 64 x 800GE

400G



Z9432F — 32 x 400GE

Z9664F — 64 x 400GE

Storage / Inband

100G / 400G

S5232F — 32 x 100GE

S5448F — 48 x 100GE + 8 x 400GE

25G / 100G

S5296F
S5248F
S5224F
S5212F

96/48/24/12 x 25GE + 8/4/3 x 100GE

OOB Mgt

1GE / 10G

N3248TE-ON
E3248P/PXE

48 x 1G/10G BT + 4 x 10GE

DELL Technologies