DataCenter Networking Desing Evolution

Oktober 2025

Who am I ?

- Andreas Roeder / andreas.roeder@nokia.com
- Nokia System Engineer based in Germany
- Almost 20 Years in Networking Industry including Nuage, Cisco, VMware, F5
- Endurance Sports Nerd

Agenda

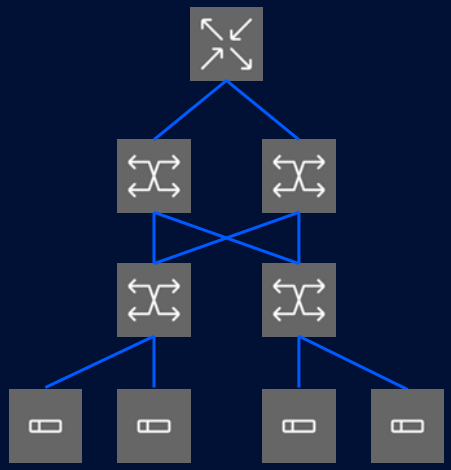
- How we ended where we are today?
- Problem to solve
- Introduction to UltraEthernet

How we ended where we are today?

Evolution

2005 – 2009 : Classic 3-tier (Core/Agg/Access) with STP

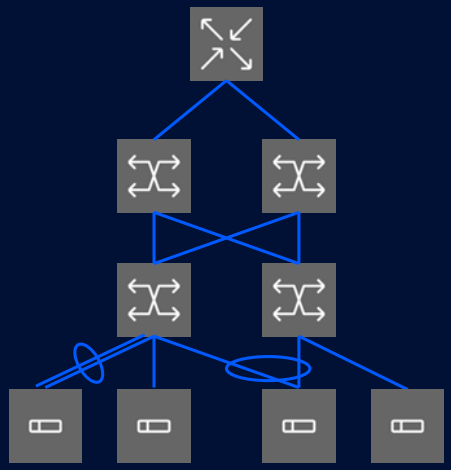
Characteristics	Pattern
Why it emerged	Virtualization just starting; simplicity; vendor reference designs
Defining characteristics	VLANs stretched across Access; north-south traffic; oversubscription acceptable
Typical tech/protocols	STP/RSTP/MSTP, 802.1Q trunks, HSRP/VRRP, LACP
Common pain points	Blocking links, slow convergence on failures, L2 loops risk, limited east-west bandwidth



Evolution

2009 – 2012 : L2 Fabrics & MLAG

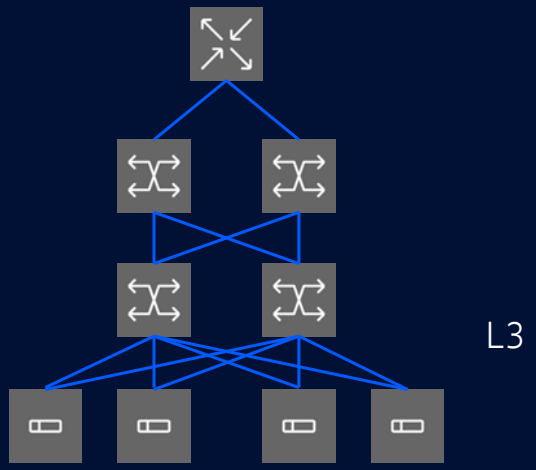
Characteristics	Pattern
Why it emerged	Server virtualization drove east-west; need active/active at L2
Defining characteristics	Flattened L2 domains; multi-chassis link aggregation; first “no-STP” fabrics
Typical tech/protocols	MLAG/vPC, TRILL / SPB, Cisco FabricPath, Juniper QFabric, Brocade VCS
Common pain points	Proprietary control planes, scale ceilings for single L2 domain, troubleshooting complexity



Evolution

2012-2014: Leaf-Spine Clos based underlay + Early Overlays

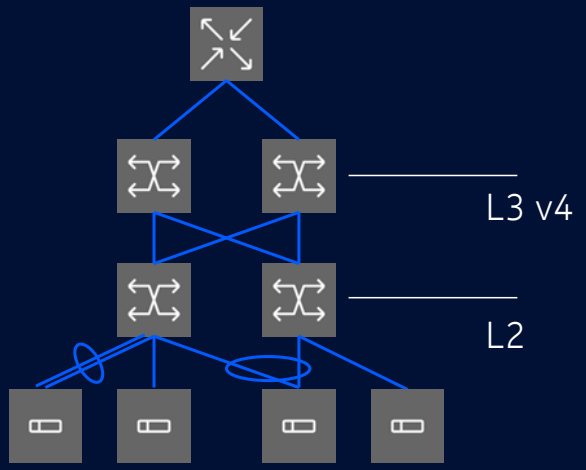
Characteristics	Pattern
Why it emerged	Scale-out apps; need uniform latency & ECMP; L3 everywhere
Defining characteristics	Small, predictable hops; ECMP load-sharing; L2 at the server edge only
Typical tech/protocols	L3 Clos, OSPF/IS-IS/BGP underlay, early overlays (VXLAN/NVGRE/STT), Nicira/NSX
Common pain points	Overlay control limited (flood-and-learn), ops/tooling still immature



Evolution

2014-2017: EVPN-VXLAN becomes the Standard

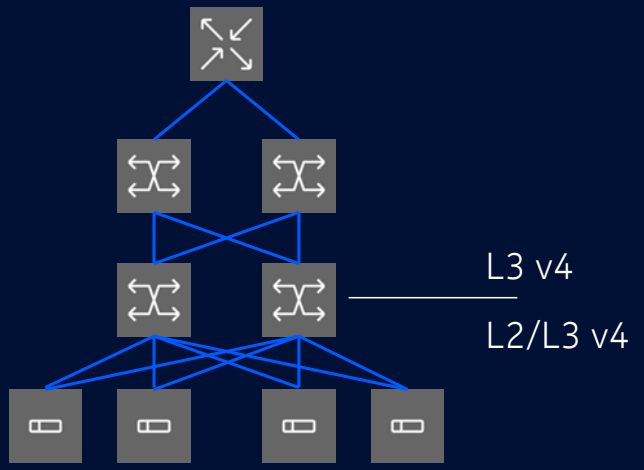
Characteristics	Pattern
Why it emerged	Need standards-based multi-tenant L2/L3 over IP with good control plane
Defining characteristics	Any-to-any L2 stretch with L3 gateway anywhere; ARP/ND suppression; MAC/IP learning in control plane
Typical tech/protocols	VXLAN (RFC 7348), EVPN (RFC 7432), eBGP/iBGP underlay, ECMP
Common pain points	Interop growing pains; new control-plane skill set; multicast-free but more BGP to manage



Evolution

2016-2019: Disaggregation + Intent Automation

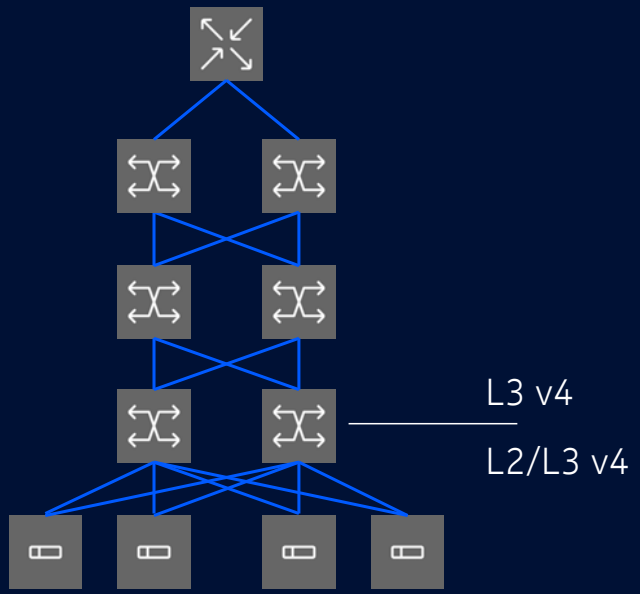
Characteristics	Pattern
Why it emerged	Cloud scale ops; vendor-agnostic choices; faster feature velocity
Defining characteristics	Whitebox/britebox, NOS choices (e.g., SONiC, Cumulus), infra as code
Typical tech/protocols	eBGP underlay, EVPN-VXLAN, Ansible/Terraform, gNMI/streaming telemetry
Common pain points	Toolchain sprawl; day-2 ops maturity; skill gap in automation/CI for networks



Evolution

2018-2021: Mature EVPN fabrics | 25/100/400G | RDMA/RoCEv2

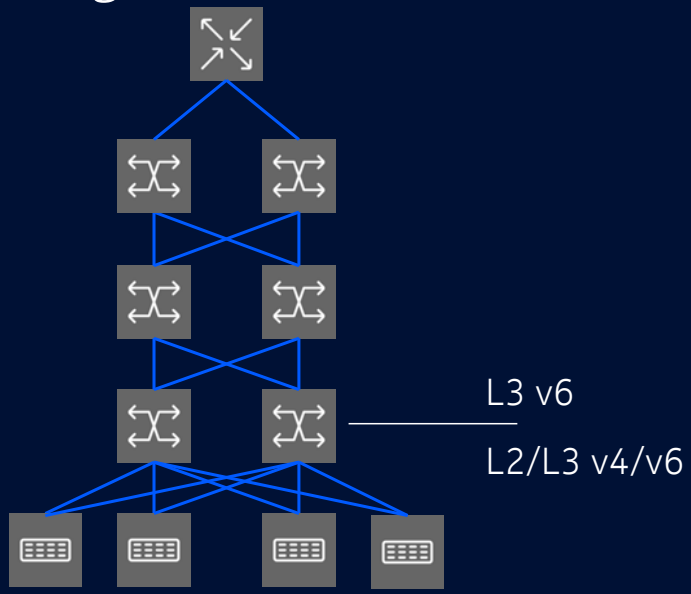
Characteristics	Pattern
Why it emerged	Microservices; HCI; storage/compute convergence; low-latency needs
Defining characteristics	Larger Clos stages; ToR L3 gateways; QoS/ECN; PFC to support RoCEv2
Typical tech/protocols	EVPN-VXLAN, ECN, DCB/PFC, DCQCN, ACI/Apstra “intent”
Common pain points	PFC issues (pause storms), congestion-hotspots, buffer/headroom tuning complexity



Evolution

2021-2023: IPv6-only, SRv6 trials, advanced load-balancing

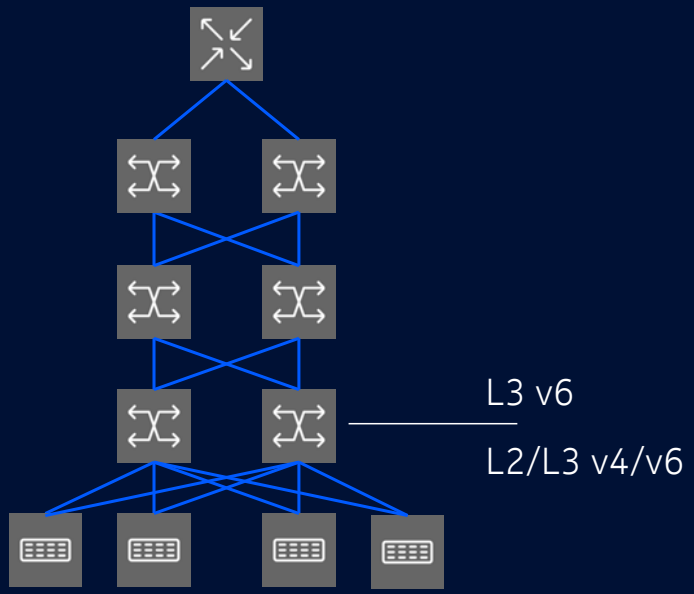
Characteristics	Pattern
Why it emerged	Address scale; simpler IP; segment routing experiments
Defining characteristics	v6 underlays, some SRv6 DC designs, better hashing/flowlets
Typical tech/protocols	IPv6-only Clos, SRv6 (limited DC adoption), flowlet-based LB (CONGA/HULA concepts), INT/telemetry
Common pain points	SRv6 hardware support variance; mixed vendor maturity; ops familiarity



Evolution

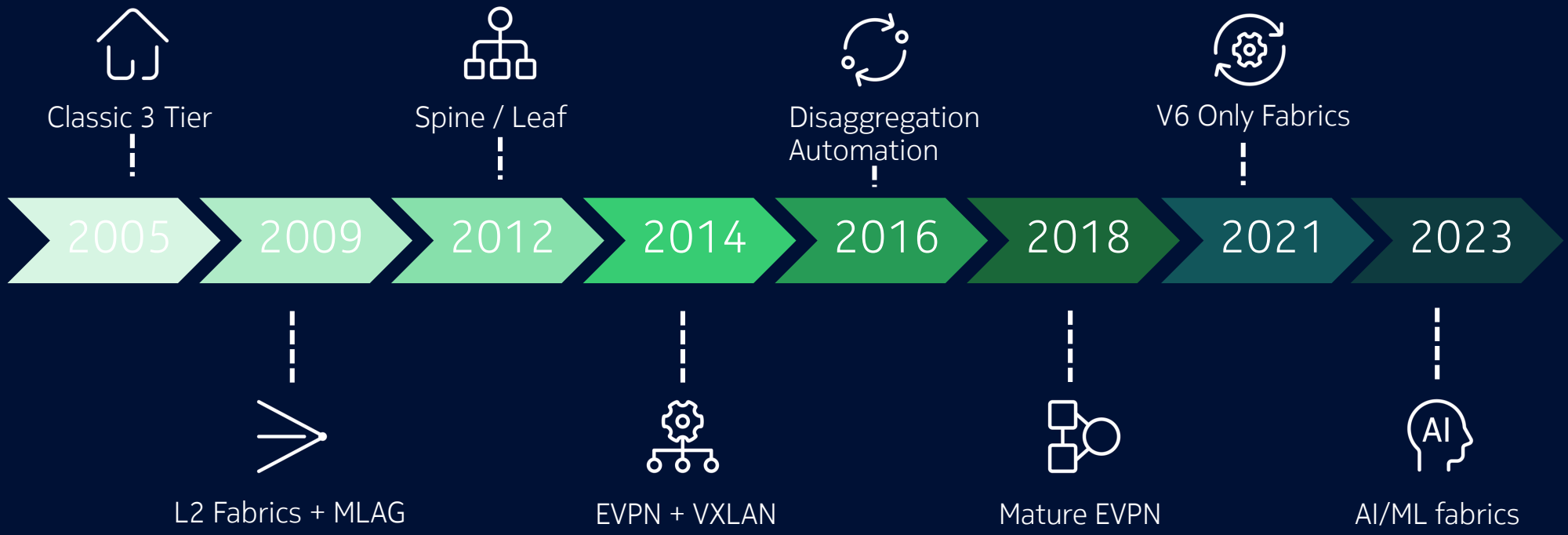
2023-today: AI/ML fabrics on Ethernet

Characteristics	Pattern
Why it emerged	GPU clusters dominate; need ultra-low tail latency & high bisection
Defining characteristics	Deeper Clos; adaptive routing; better congestion control; fine-grained QoS/telemetry; PTP for sync
Typical tech/protocols	EVPN-VXLAN underlay/overlay, ECN+DCQCN, selective PFC, advanced ECMP/flowlets, in-band telemetry, 400/800G optics; emerging **UE**; vendor AI fabric stacks
Common pain points	Still tricky to make Ethernet “Infiniband-like”; PFC hazards; topology-aware scheduling and failure handling at scale

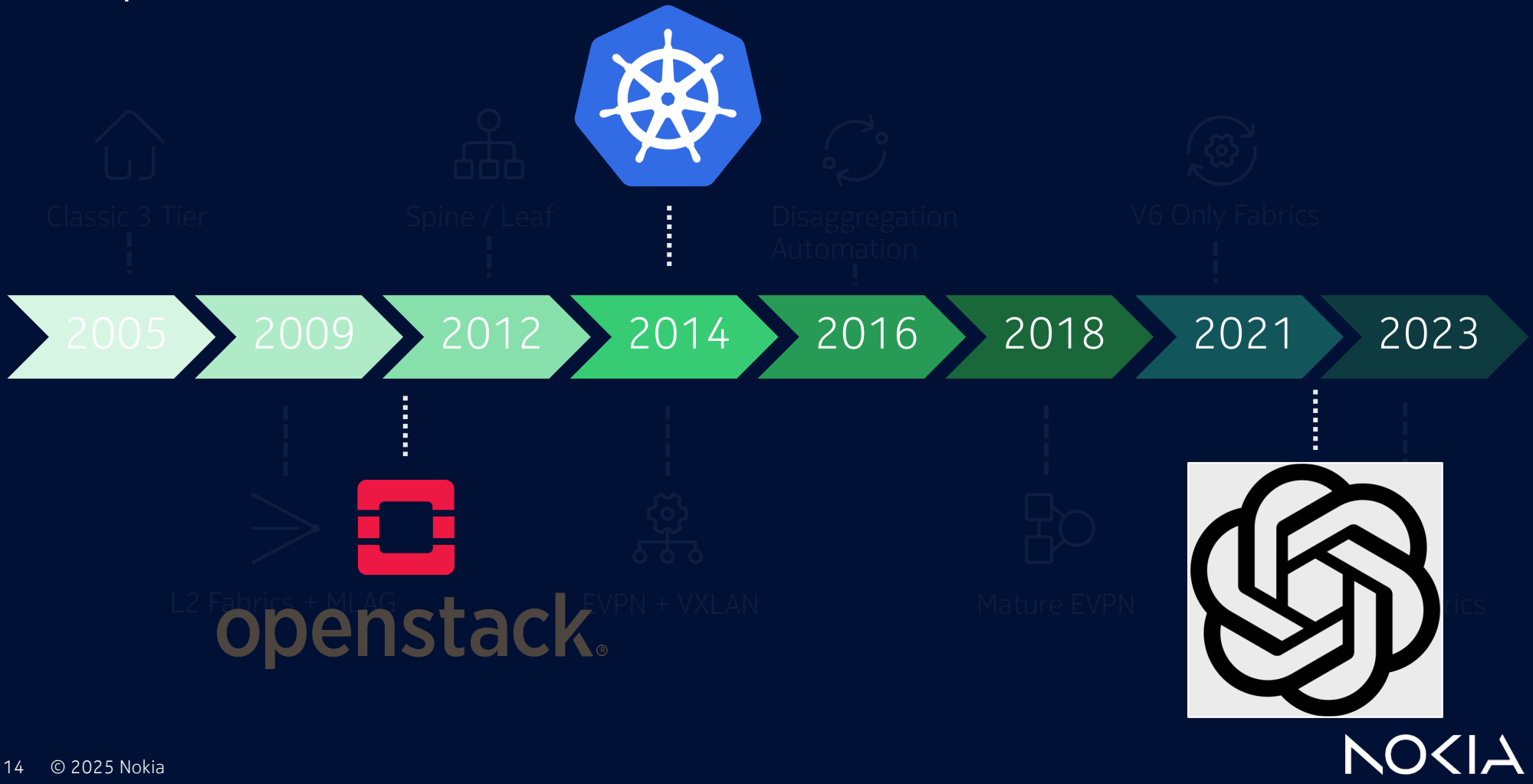


NOKIA

Timeline

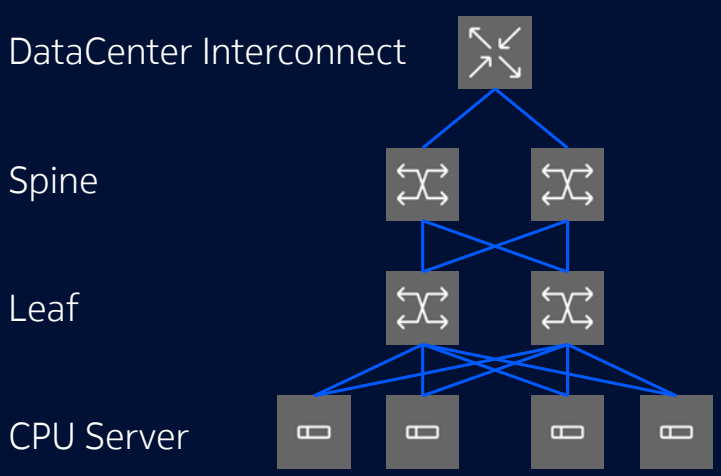


Disruption

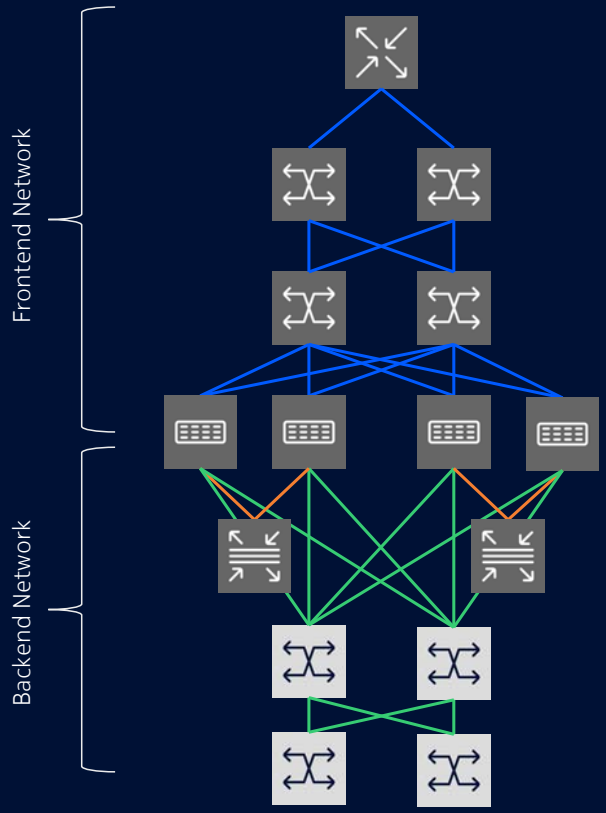


Rise of AI

Another Problem to solve...



Ethernet Links
InfiniBand Link
GPU Fabric Link



DataCenter Interconnect

Spine

Leaf

CPU Server + GPU Server

GPU Fabric

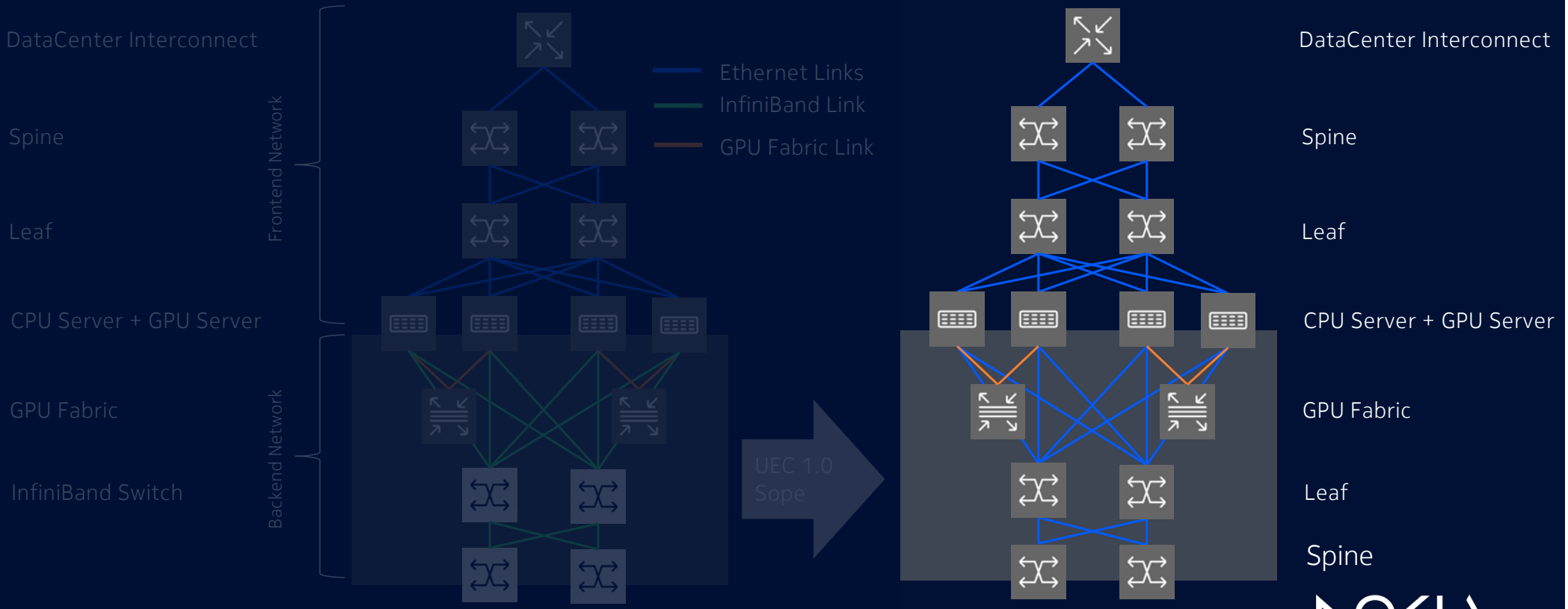
InfiniBand Switch

Problem to solve

Why Ethernet for Backend Networks?

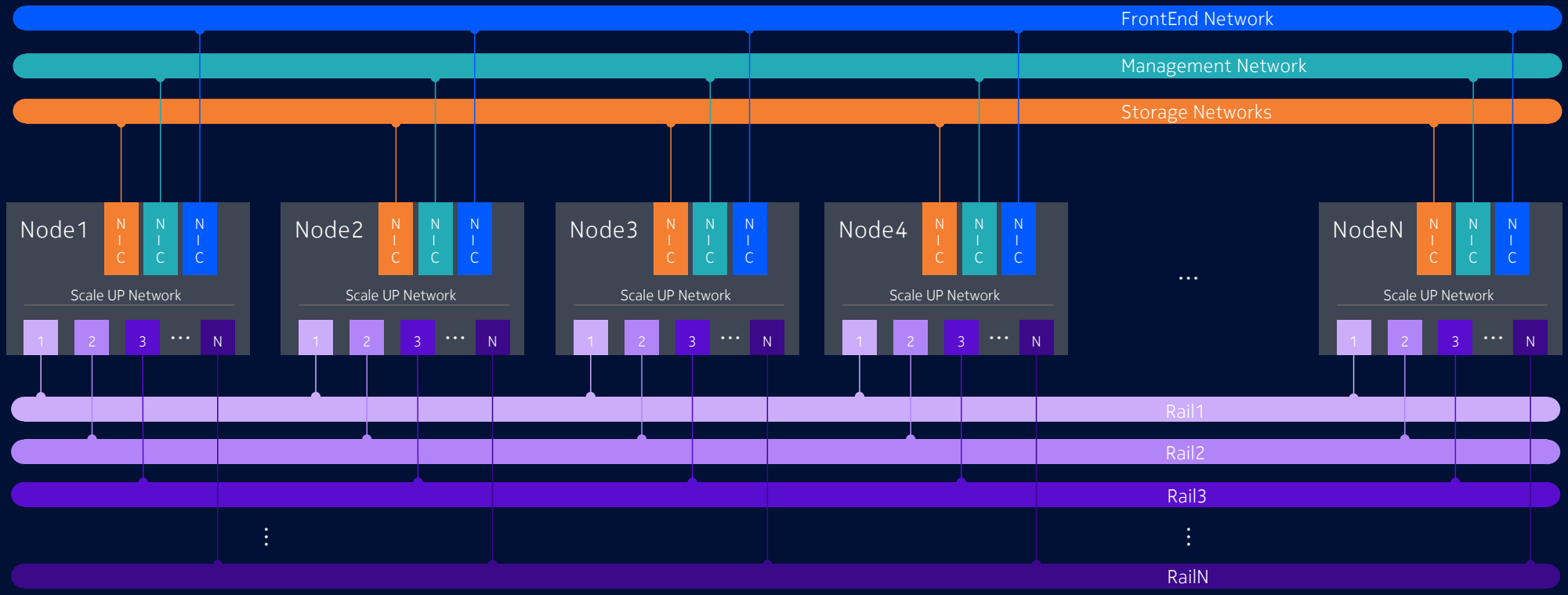
- InfiniBand is not considered the right technology given its limited scale (clusters with few thousands of GPUs) and supplier diversity (NVIDIA is the only one)
- The large Ethernet community has acknowledged deficiencies in Ethernet, mostly related to the use of RDMA and congestion management at scale.
- The Ultra Ethernet Consortium (UEC) - https://ultraethernet.org was formed in 2023 to address these architectural and technology challenges to enable replacement of InfiniBand with Ethernet in the backend

Evolution



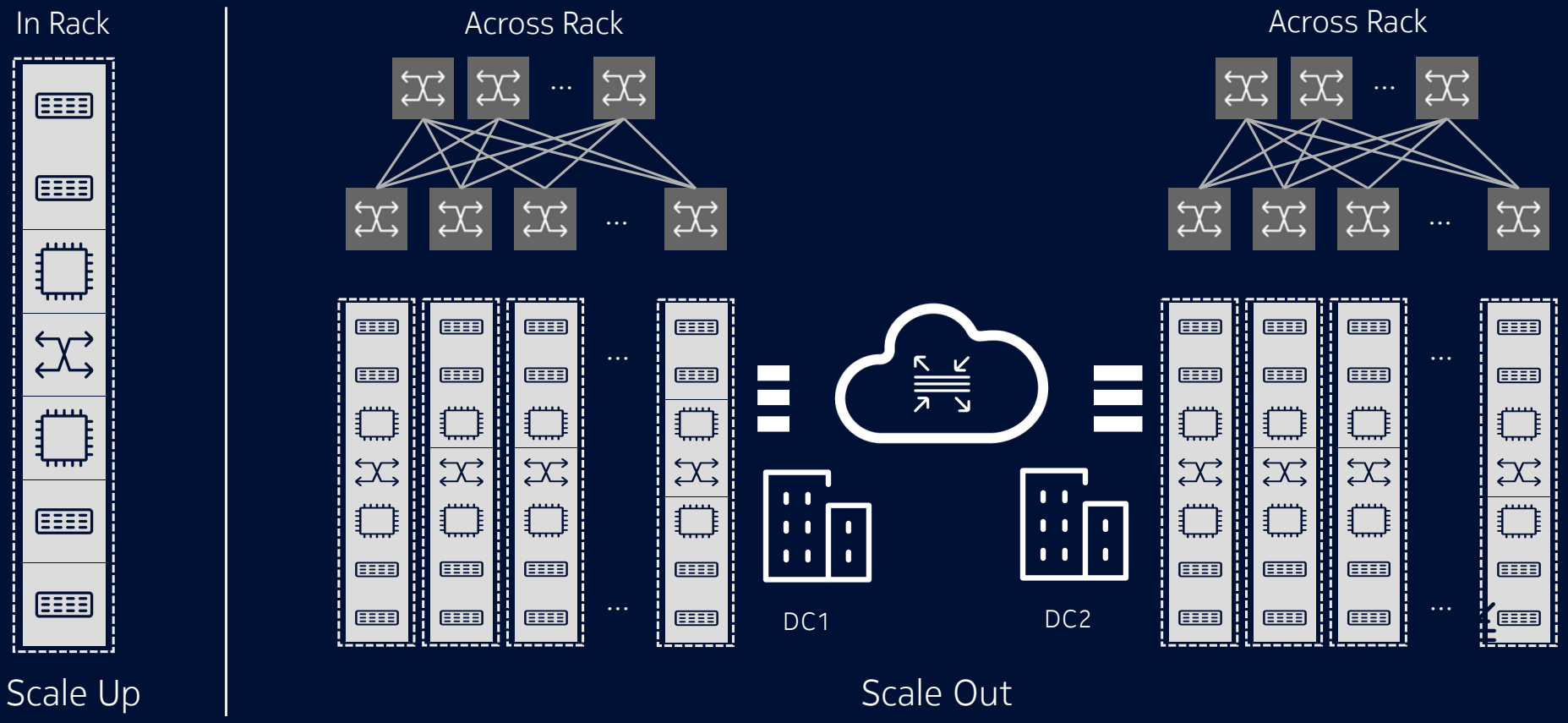
Overall Design for an AI DC

What kind of Networks are needed?



Introduction to UltraEthernet

AI Scale-UP and Scale-Out Networking



Ultra Ethernet Consortium

What are we trying to solve with RMA / RoCE Implementation

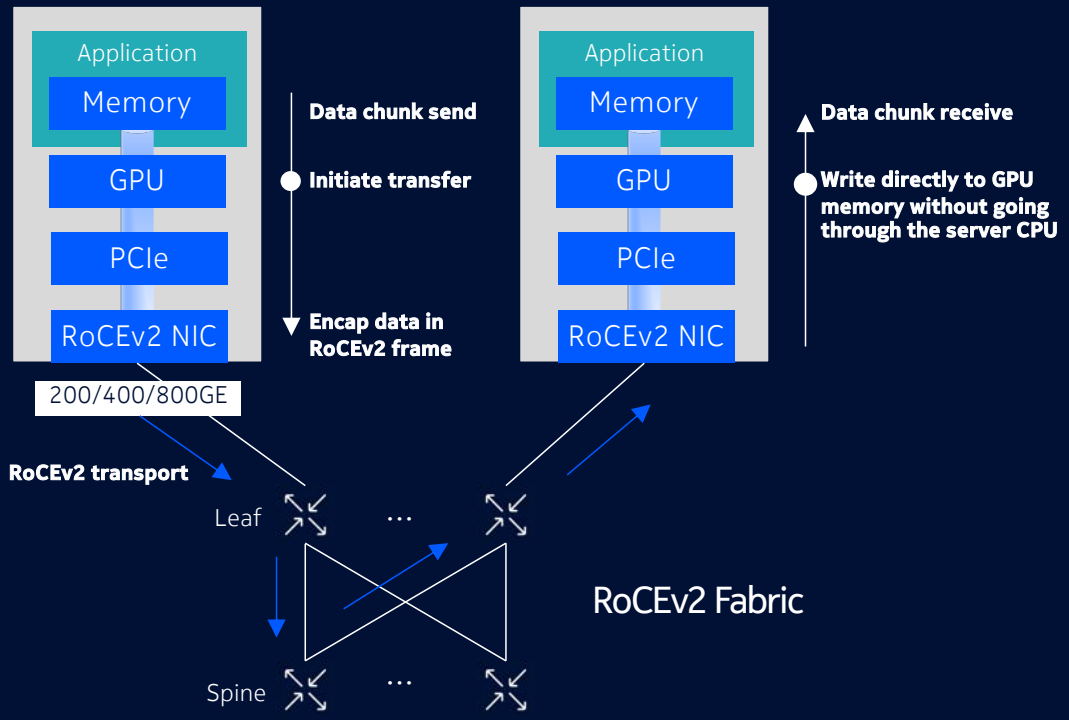
- Lack of Multipathing in current Solution
- Message and initiator/target based communications
- Fine grained congestion control with rapid response
- Unordered and ordered packet delivery and packet spraying
- Native support for RDMA and collective operations

Ultra Ethernet Transport Goals

- Multipathing RMA
- Relaxed Delivery Ordering
- Rapid Loss Recovery
- Modern congestion control for the DC – Rapid Startup and Slowdown, Multipath Aware
- Run on IPv4/IPv6
- Lossy and Lossless Operation
- Ordered and Unordered Delivery
- Day-1 Security

Optimized GPU-to-GPU Data Transfers

RDMA over RoCEv2 Lossless Ethernet Fabric

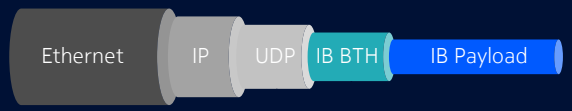


RDMA (Remote Direct Memory Access) delivers the following performance attributes for massive data movements with DC environments:

- Zero-copy data transfers
- Kernel bypass
- Low latency
- High throughput

RoCEv2 enables an InfiniBand transport layer to run over UDP/IP (IPv4 or IPv6) to support RDMA:

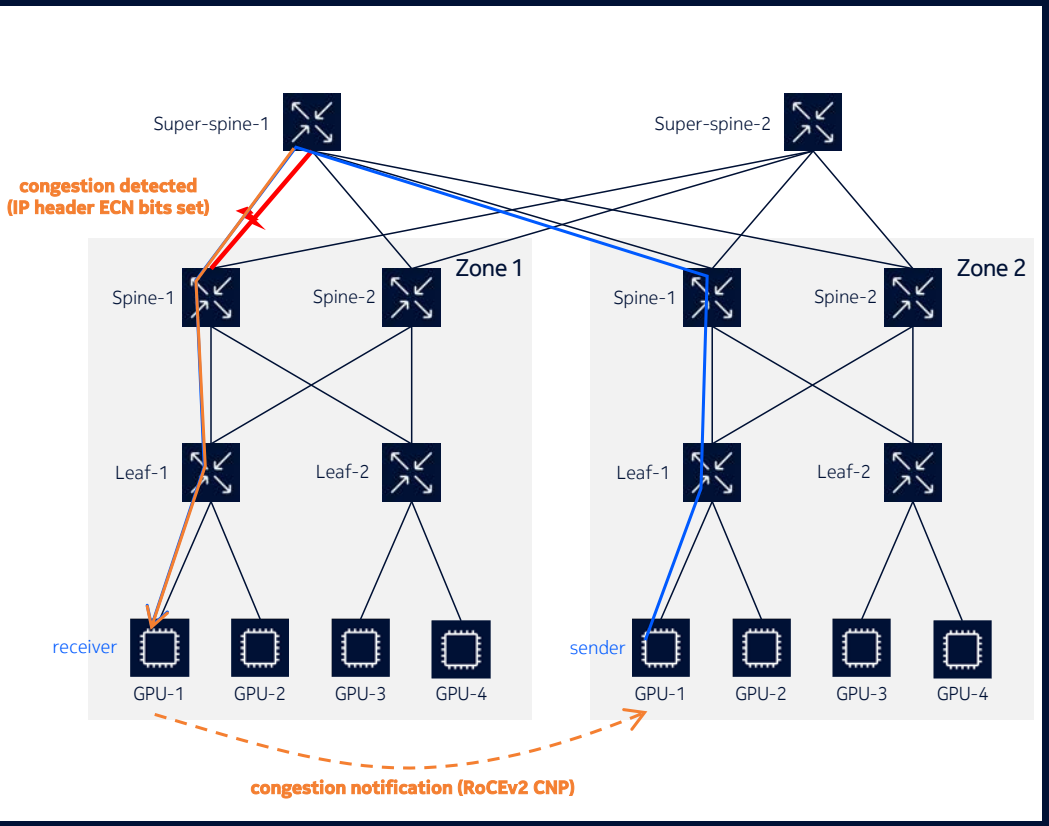
- Flow are uniquely identified by Queue-Pair ID's
- Relies upon a lossless Ethernet fabric

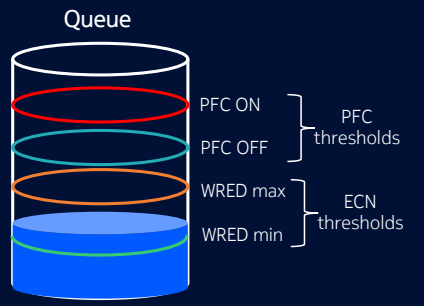


InfiniBand Base Transport Header
- Opcode: transport type
- Dest queue-pair
- Packet sequence number

RoCEv2 Lossless Ethernet Fabric

DCQCN (Data Cetner Quantized Congestion Notification) – ECN

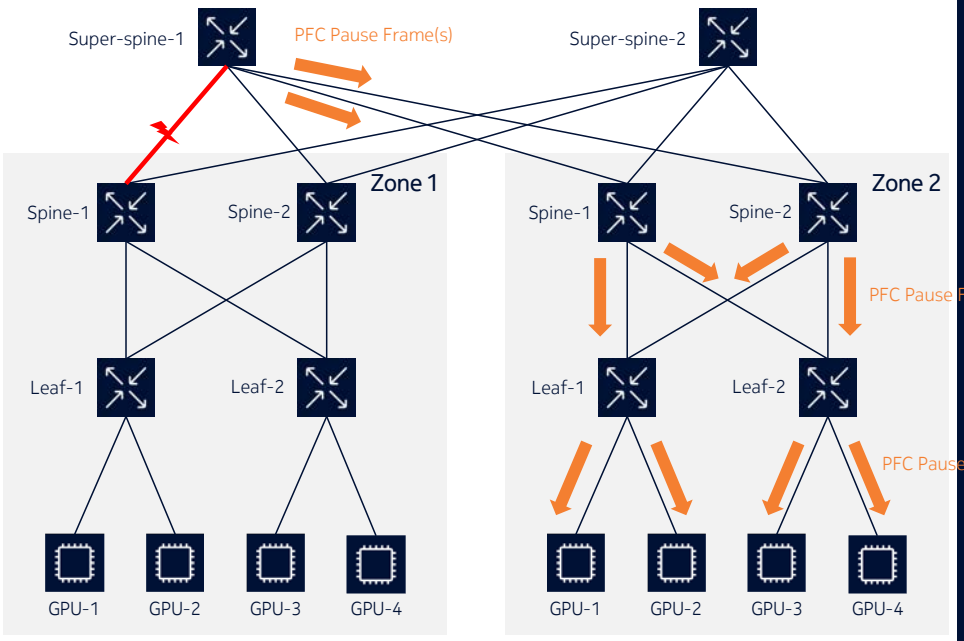


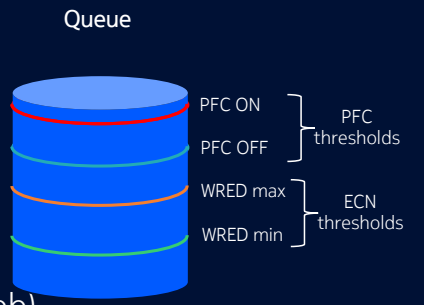


- ECN (Explicit Congestion Notification)
  - Provides end-to-end congestion control between two endpoints (RDMA NICs)
  - Upon interface congestion, DC fabric switches set the ECN bits (IPv4 ToS / IPv6 TC fields) of transiting packets to notify downstream receivers
  - ECN slope profile defines the proportion of packets marked w/ ECN bit
    - Q-depth < WRED min = no packets marked
    - WRED min < Q-depth < WRED max = linearly increasing proportion of packets marked
    - Q-depth > WRED max = all packets marked
  - During congestion, sender(s) reduce their transmission rate until congestion clears
  - Suitable to resolve minor congestion within DCF
- Advanced operational capabilities
  - ECN statistics

RoCEv2 Lossless Ethernet Fabric

DCQCN (Data Cetner Quantized Congestion Notification) – PFC

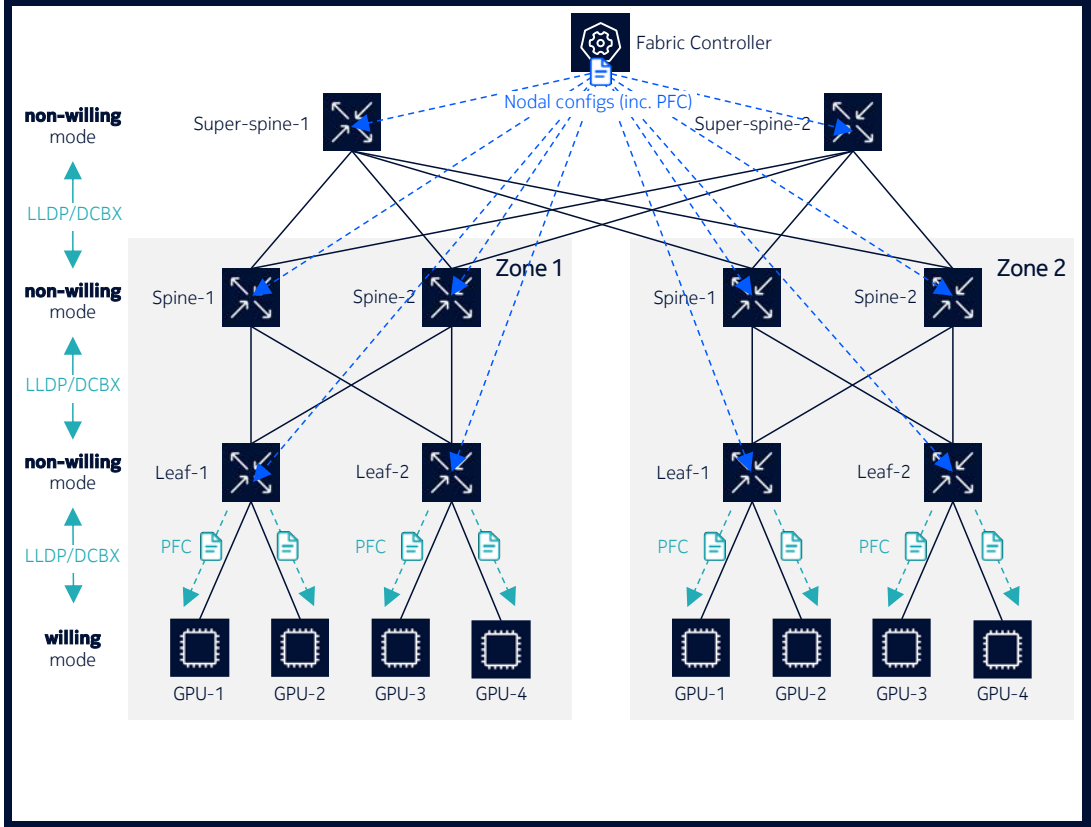




- PFC (Priority Flow Control) (IEEE 802.1Qbb)
  - Enables hop-by-hop per-priority flow control within the DC fabric and NIC's
  - When a buffer is becoming full, the switch/NIC requests the sender to temporarily stop sending traffic for a specific priority class (PFC pause frame)
  - The sender halts transmission until is receives a signal that the congestion has cleared
  - PFC ON/OFF thresholds define PFC pause frame behavior
    - Q-depth > PFC ON = send PFC pause frame to peer
    - Q-depth < PFC OFF = cease sending PFC pause frames
  - Suitable to resolve severe congestion (inc. microbursts) within DCF
- Advanced capabilities:
  - DCBX (PFC TLV) for PFC capability and configuration exchange
  - PFC watchdog

RoCEv2 Lossless Ethernet Fabric

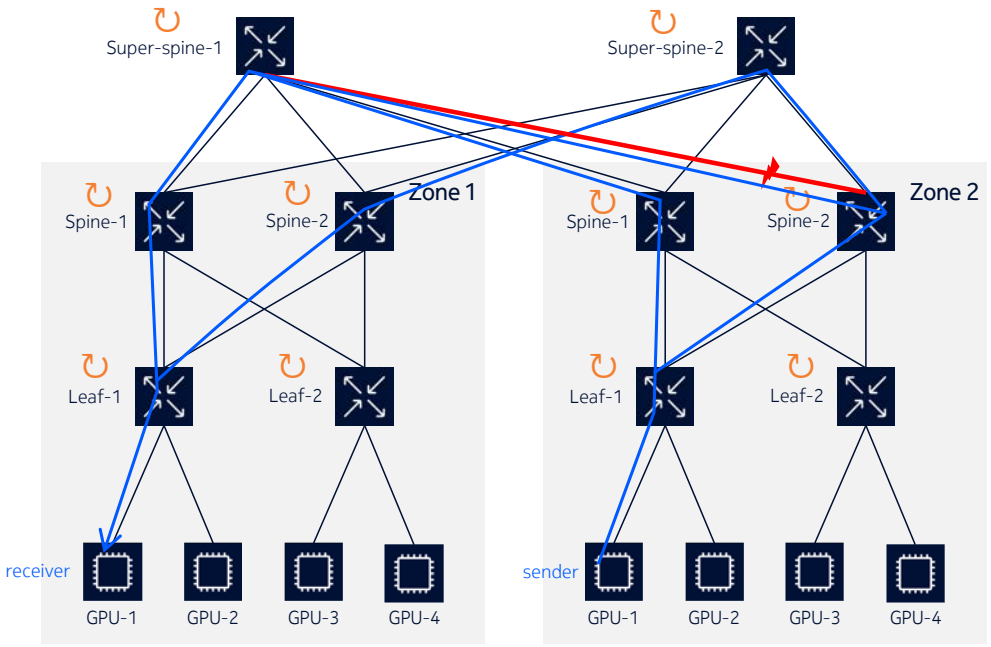
DCBX (Data Center Bridging Exchange Protocol)



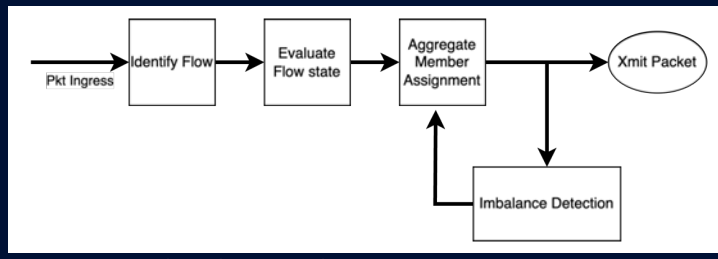
- DCBX (IEEE 802.1Qaz) runs over LLDP and exchanges DCB attributes between peers on a link, including PFC, ETS, application priority configuration, etc.
  - Useful to enforce feature configuration consistency or identify relevant feature misconfigurations across DC environment
- Attributes of different types are exchanged between peers:
  - Informational: exchanged via LLDP but do not influence the state or operation of DCBX.
  - Asymmetric: attributes are exchanged between two peers, whereby the *desired attributes on either peer may not match*. E.g. ETS configuration and recommendation TLVs
  - Symmetric: attributes are exchanged between two peers, with the objective of *both peers utilizing the same attributes*. E.g. PFC configuration TLVs
- When exchanging PFC attributes, two modes of operation are relevant:
  - Willing mode: node accepts PFC configuration from its peer
  - Non-willing mode: node enforces its own PFC configuration and pushes it to its peer
  - **Best practice**: centralized control of PFC parameters from switches (non-willing mode) towards NIC's (willing mode)

RoCEv2 Lossless Ethernet Fabric

Advanced AI Load Balancing: DLB



DLB Imbalance detection model



- DLB (Dynamic Load Balancing)
  - Imbalance model introduces a feedback mechanisms for egress hash flow assignment
  - Locally significant capability to improve ECMP load across all nodes in GPU fabric
    - Considers the state of the aggregate members when assigning flows
    - Allowing existing flows to consider changing load conditions
    - Identify instances where active flows can be moved by avoiding re-ordering
  - Imbalance detection model considers the following:
    - Egress Port Queue fill size, Egress Port Utilization, Ingress Traffic Manager (ITM) Port Queue Size
  - NIC support for packet re-ordering is mandatory
    - E.g. Nvidia ConnectX-7/8, Broadcom Thor 2, AMD Pensando Pollara 400, etc.

AI fabric choices

Our focus today is on Ethernet back-end scale-out fabrics..

DC integration strategy?

Integrated
- one fabric for cloud services and AI workloads

Separate
- front-end fabric connects to external users and data
- back-end fabric for AI workloads

Backend fabric technology?

Ethernet
- ROCEv2 + DCQCN
- standard 400G Ethernet NICs,
- switches and ISLs

DDC
- fully scheduled fabric
- standard NICs, proprietary ISLs

Infiniband
- carryover from HPC, expensive

Ultra Ethernet Consortium (UEC)
- emerging

Topology to support cluster scale?

Single switch
1x 7250 IXR-18e = 1K GPUs

1-tier rail-optimized
8x H5 leafs, one per rank
=1K GPUs

2-tier non-blocking
128x H5 leafs + 64x H5 spines = 8K GPUs

3-tier non-blocking
512x H5 leafs + 512x H5 spines +256x H5 super-spines = 32K GPUs

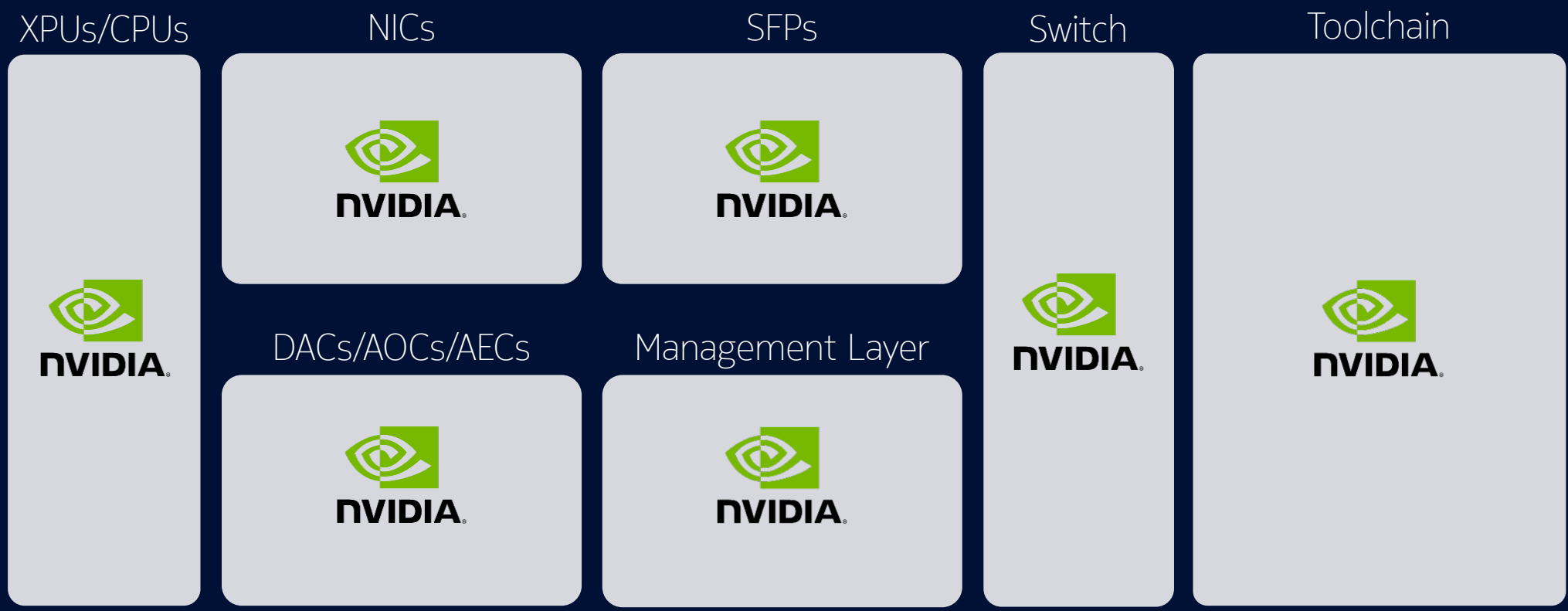
Multi-tenancy design?

Storage design?

Management design?

* Assume 8xGPUs + 8x400G NIC ports per server

The InfiniBand Lock-in Problem



Ethernet's Ecosystem Advantage

XPUs/CPUs

NVIDIA

AMD

intel

...

NICs

NVIDIA

AMD

BROADCOM

SFPs

Vendor SFPs +
3rd Party SFPs

Switch

NOKIA

ARISTA

JUNIPER NETWORKS

CISCO

NVIDIA

DELL

Hewlett Packard Enterprise

Toolchain

ANSIBLE

Grafana

netbox

Terraform

MuleSoft

DACs/AOCs/AECs

Vendor Cables +
3rd Party Cables

Management Layer

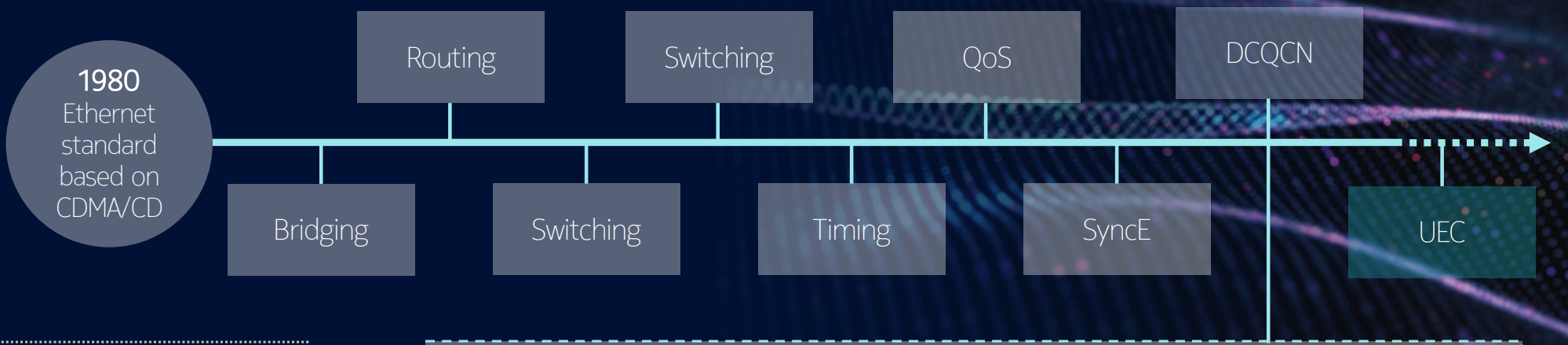
Vendors Platforms +
Open-Source and 3rd Party
Management Tools

Source: https://ultraethernet.org/wp-content/uploads/sites/20/2025/06/UE-Specification-6.11.25.pdf

NOKIA

Ethernet has a long history of winning…

… and is evolving for AI/HPC workloads



Meanwhile competing technologies fade away despite technology advantages:

- Serial P2P
- Token Ring
- SMDS
- ATM
- Frame Relay

RoCEv2 Priority Flow Control Dynamic Traffic Flow Balancing

Higher Data Rates Explicit Congestion Notification

Evolution for AI/HPC Workloads

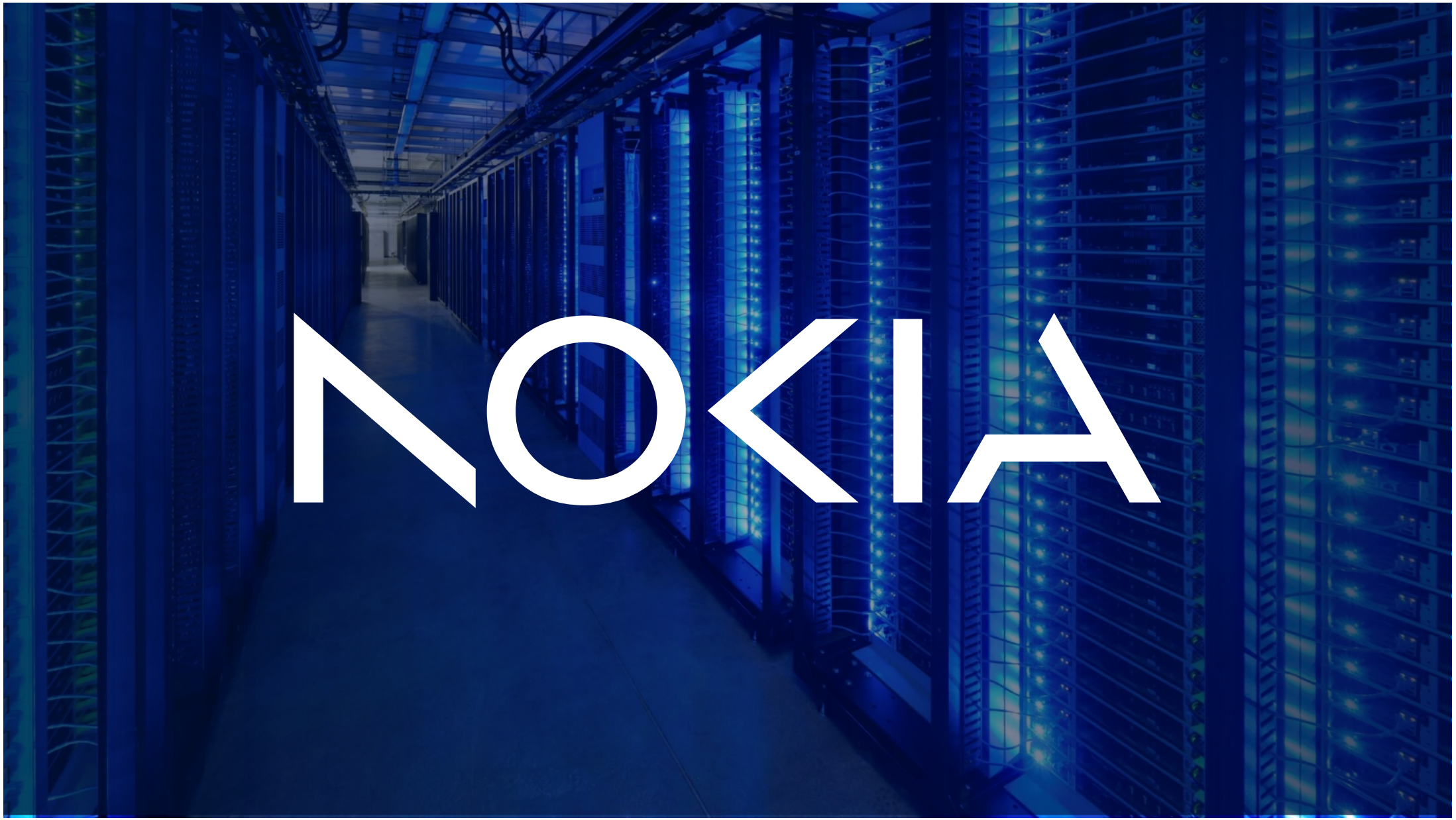
Conclusion

ETHERNET is here to Stay

ETHERNET will evolve (like done before)

ETHERNET will Scale

ULTRA ETHERNET is ready for AI and HPC of the Future


NOKIA