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

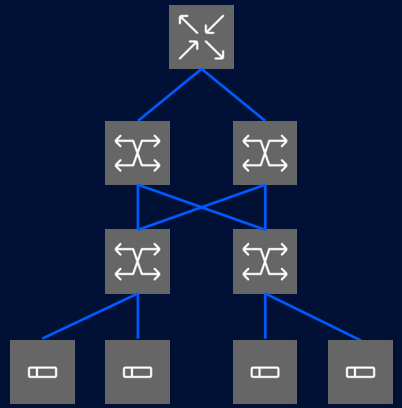
- How we ended where we are today?
- Problem to solve
- Introduction to UltraEthernet

How we ended where we are today?

Evolution

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

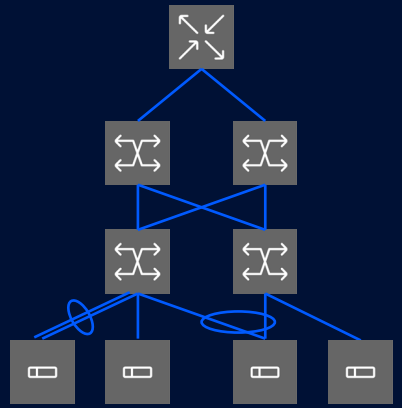
Characteristics	Pattern
Why it emerged	Virtualization just starting; simplicity; vendor reference designs
Defining characteristics	VLANs stretched across Access; north-south traffic; oversubscription acceptable
Typical tech/protocols	STP/RSTP/MSTP, 802.1Q trunks, HSRP/VRRP, LACP
Common pain points	Blocking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



Evolution

2009 – 2012 : L2 Fabrics & MLAG

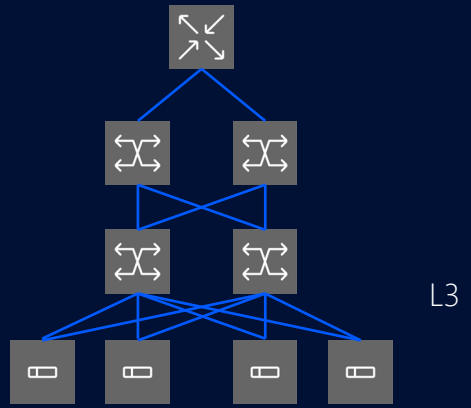
Characteristics	Pattern
Why it emerged	Server virtualization drove east-west; need active/active at L2
Defining characteristics	Flattened L2 domains; multi-chassis link aggregation; first “no-STP” fabrics
Typical tech/protocols	MLAG/vPC, TRILL / SPB, Cisco FabricPath, Juniper QFabric, Brocade VCS
Common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



Evolution

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

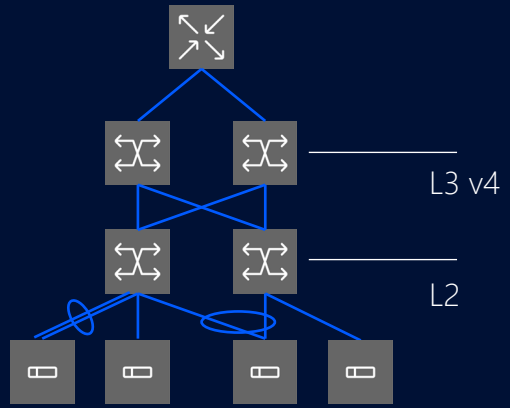
Characteristics	Pattern
Why it emerged	Scale-out apps; need uniform latency & ECMP; L3 everywhere
Defining characteristics	Small, predictable hops; ECMP load-sharing; L2 at the server edge only
Typical tech/protocols	L3 Clos, OSPF/IS-IS/BGP underlay, early overlays (VXLAN/NVGRE/STT), Nicira/NSX
Common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



Evolution

2014-2017: EVPN-VXLAN becomes the Standard

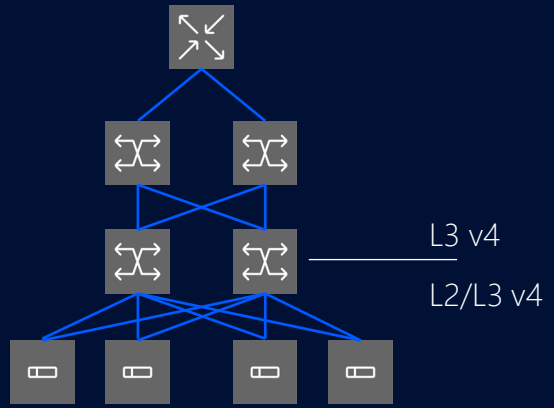
Characteristics	Pattern
Why it emerged	Need standards-based multi-tenant L2/L3 over IP with good control plane
Defining characteristics	Any-to-any L2 stretch with L3 gateway anywhere; ARP/ND suppression; MAC/IP learning in control plane
Typical tech/protocols	VXLAN (RFC 7348), EVPN (RFC 7432), eBGP/iBGP 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

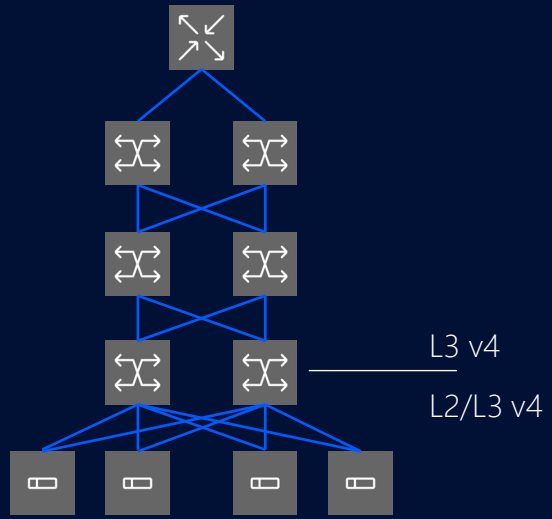
Characteristics	Pattern
Why it emerged	Cloud scale ops; vendor-agnostic choices; faster feature velocity
Defining characteristics	Whitebox/britebox, NOS choices (e.g., SONiC, Cumulus), infra as code
Typical tech/protocols	eBGP underlay, EVPN-VXLAN, Ansible/Terraform, gNMI/streaming telemetry
Common pain 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

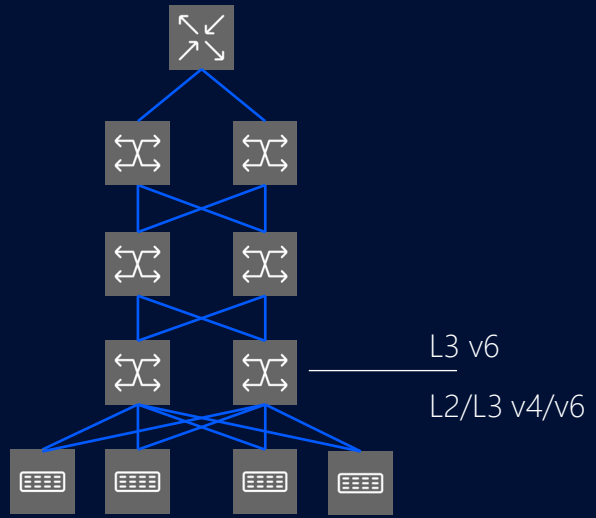
Characteristics	Pattern
Why it emerged	Microservices; HCI; storage/compute convergence; low-latency needs
Defining characteristics	Larger Clos stages; ToR L3 gateways; QoS/ECN; PFC to support RoCEv2
Typical tech/protocols	EVPN-VXLAN, ECN, DCB/PFC, DCQCN, ACI/Apstra “intent”
Common pain points	PFC issues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



Evolution

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

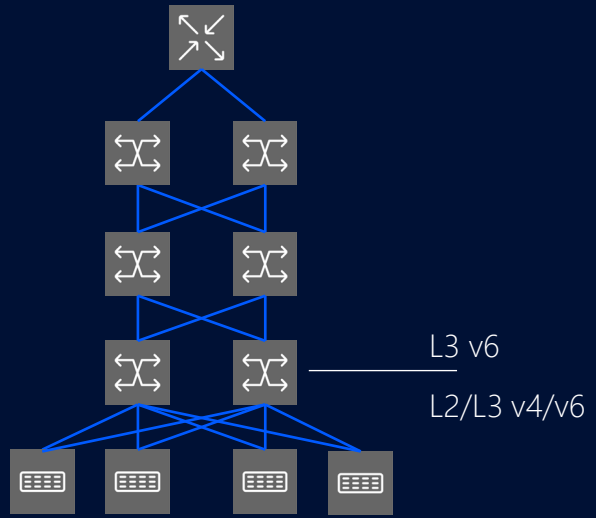
Characteristics	Pattern
Why it emerged	Address scale; simpler IP; segment routing experiments
Defining characteristics	v6 underlays, some SRv6 DC designs, better hashing/flowlets
Typical tech/protocols	IPv6-only Clos, SRv6 (limited DC adoption), flowlet-based LB (CONGA/HULA concepts), INT/telemetry
Common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



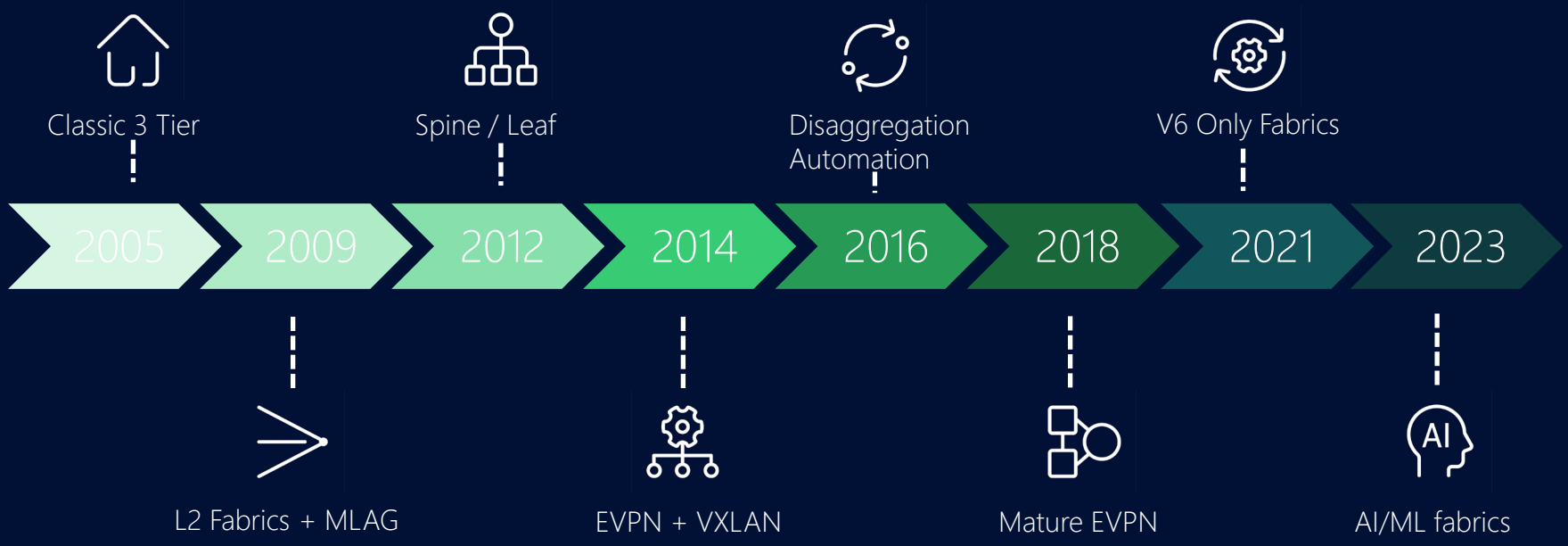
Evolution

2023-today: AI/ML fabrics on Ethernet

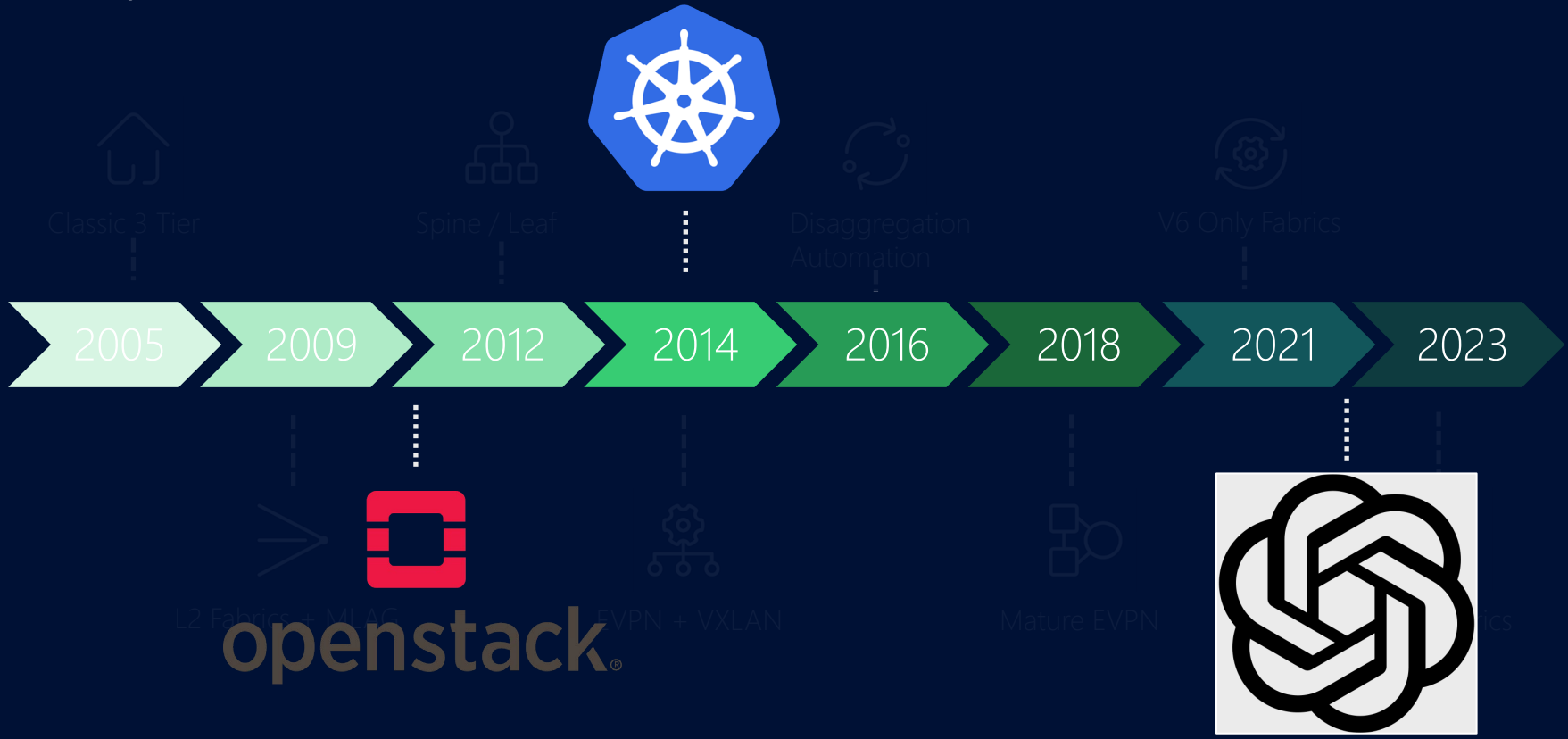
Characteristics	Pattern
Why it emerged	GPU clusters dominate; need ultra-low tail latency & high bisection
Defining characteristics	Deeper Clos; adaptive routing; better congestion control; fine-grained QoS/telemetry; PTP for sync
Typical tech/protocols	EVPN-VXLAN underlay/overlay, ECN+DCQCN, selective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



Timeline

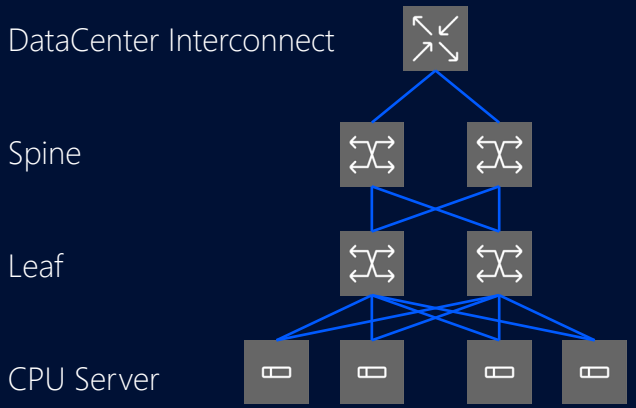


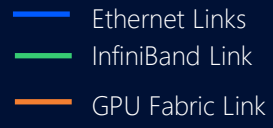
Disruption

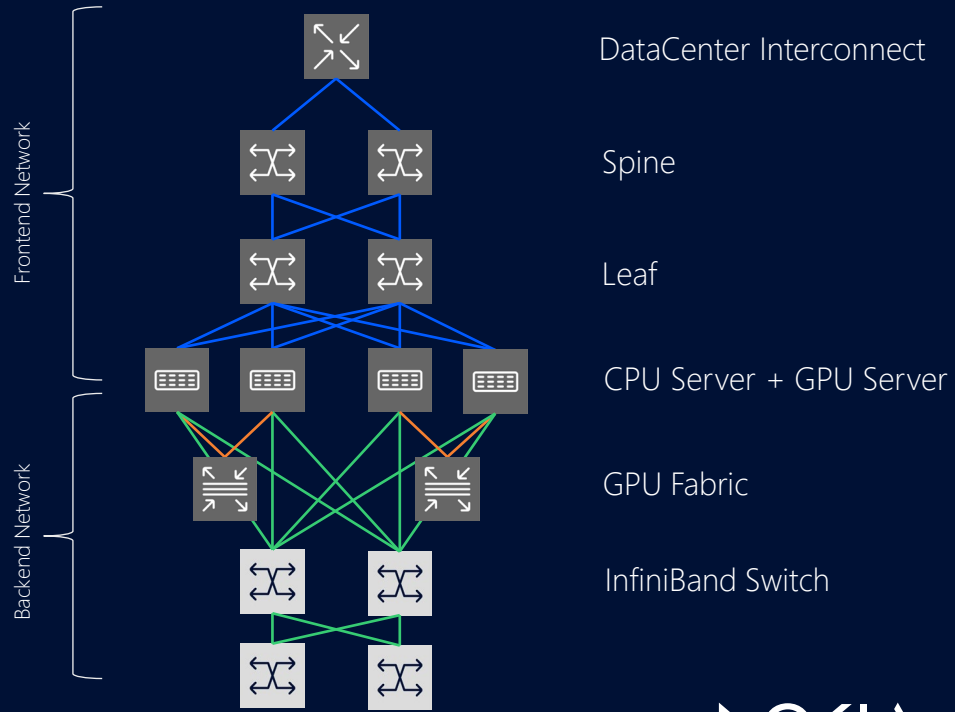


Rise of AI

Another Problem to solve…







Evolution

What happened?

- With the growth of AI applications and use of GPUs in servers, the traditional data center network comprising Ethernet switches and network adapters (or NICs) used in servers has become the “frontend network”
- Deals with movement of data among modern applications that run in CPUs and storage appliances (the east-west traffic), and data to other data centers or the Internet (the north-south traffic)
- A new network called the “backend network” has evolved with the sole purpose of handling data movement between GPUs. RDMA as Baseline technology to Access Memory of GPU’s
- GPUs used to process AI training and inference algorithms move data between them orders of magnitude more than CPUs have in the past
- One misbehaving element (an overburdened GPU or a congested link between a few of them) could throw the entire routine into disarray; as a result training algorithms could take way longer to complete, recommender systems could fail to do so in time annoying users, and expensive infrastructure (GPUs cost a lot, use a lot of power) could go underutilized

Rise of Backend Networks

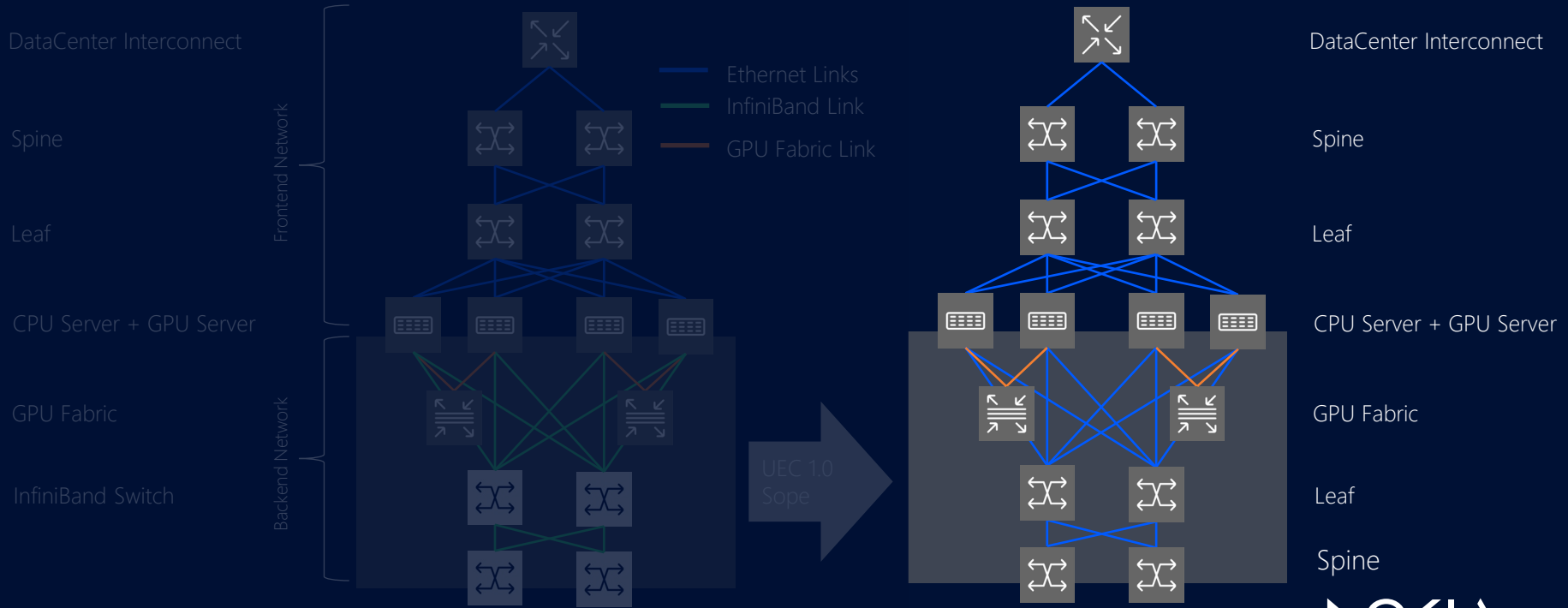
- 2 Parts of the Backend Networks
  - Proprietary GPU fabrics like NVLink and Infinity Fabric are used for GPU-to-GPU communication within the server or for smaller clusters of NVIDIA and AMD GPUs respectively. This part of the backend network is called the “scale-up” network
  - For communication between a larger set of GPUs, a complementary “scale-out” network is used, which today is serviced mostly using InfiniBand switches and host adapters used in servers for Backend Networks

Problem to solve

Why Ethernet for Backend Networks?

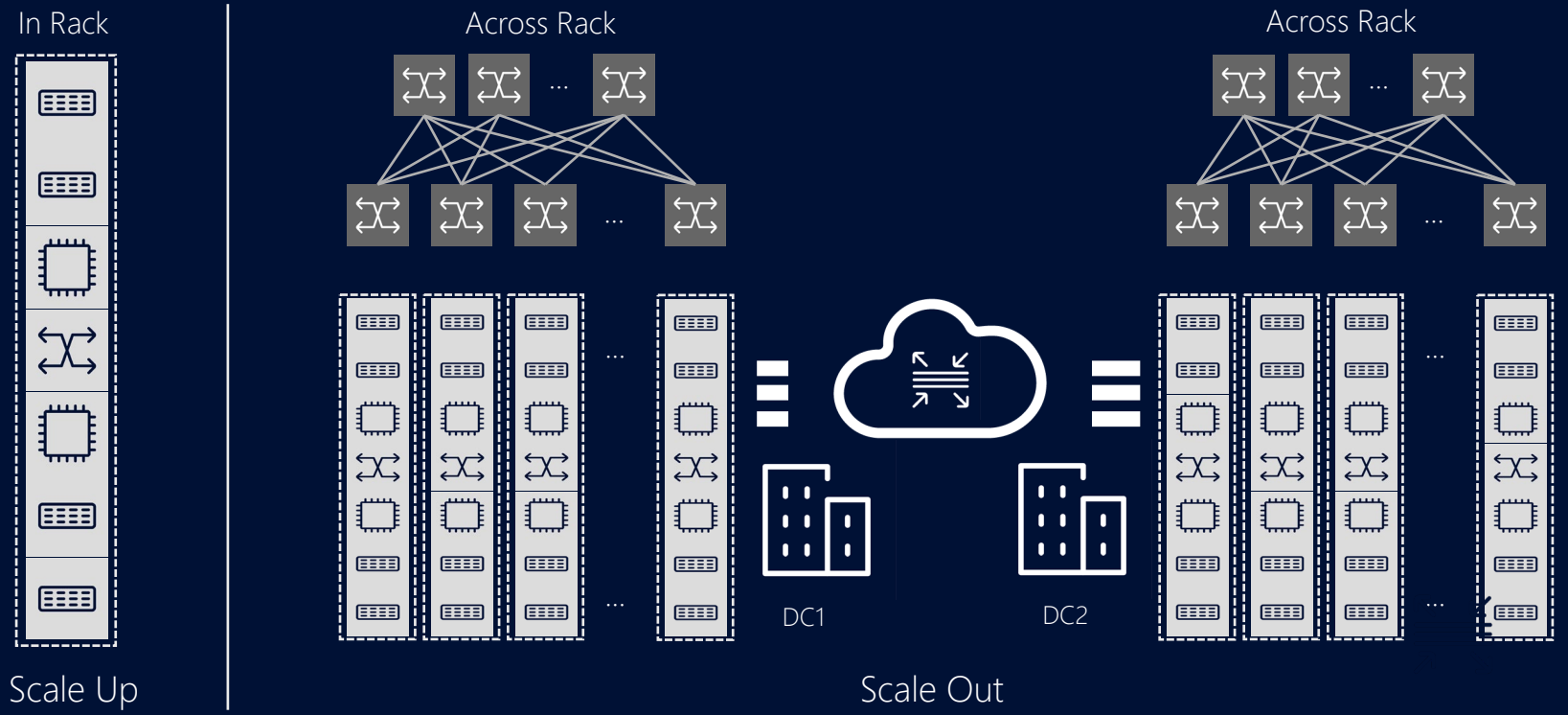
- InfiniBand is not considered the right technology given its limited scale (clusters with few thousands of GPUs) and supplier diversity (NVIDIA is the only one)
- The large Ethernet community has acknowledged deficiencies in Ethernet, mostly related to the use of RDMA and congestion management at scale.
- 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



Introduction to UltraEthernet

AI Scale-UP and Scale-Out Networking

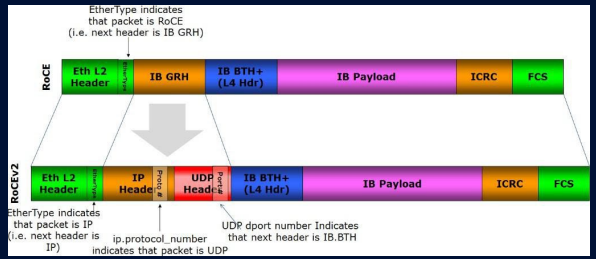


Ultra Ethernet Consortium

RMA and its relation to Ethernet

RMA Highlights

- Accelerators today communicate with RMA
- RMA is hardware delivery straight to/from memory
  - Kernel bypass, zero-copy
  - Hardware loss detection, retrans, loss recovery
- RDMA over IP (RoCEv2) is a widely deployed RMA implementation



Ethernet Highlights

- Broad Ecosystem
  - NIC's, Switches, Optics, Cables
  - Multi-vendor at all layers
- Rapid Innovation overall
- Tooling, Interops, Knowledge
- Overall Scaling

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

Overall Ultra Ethernet Specification Structure

Introduction

UE Software Layer

UE Transport Layer

UE Network Layer

UE Link Layer

UE Physical Layer

UE Compliance requirements

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

Congestion Management
Dual Congestion Control Modes

NSCC – Network Signal Congestion Control

- Reactive Server Side congestion Control method which act to things like
  - Trimmed Packets
  - ECN (Explicit Congestion Notification) marks
  - Increased network Latency

RCCC – Reciever Credit Congestion Control

- Reciever based with Focus on in-cast congestion – Multiple sources simultaneously send data to a single destination
- Credit based System based on available buffer Space

Both mechanisms are used in parallel

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

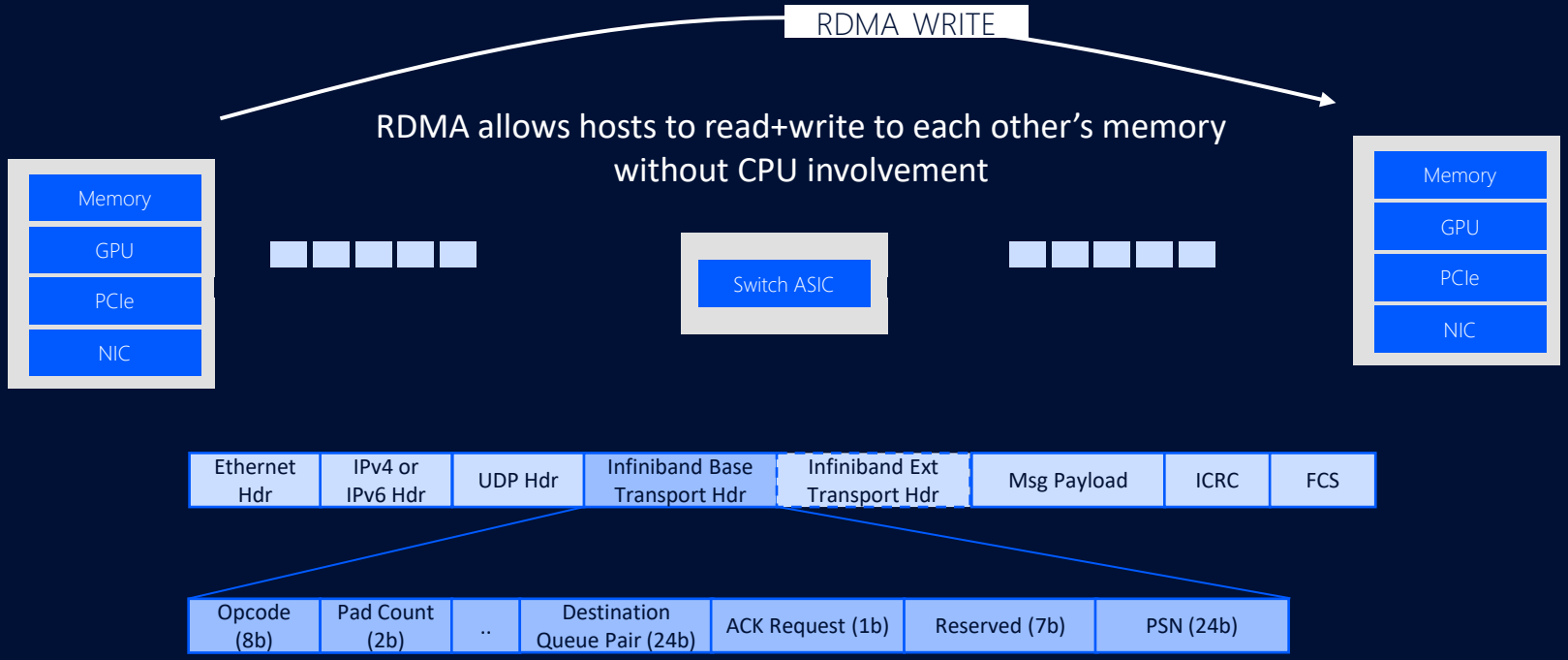
Multi-tenancy design?

Storage design?

Management design?

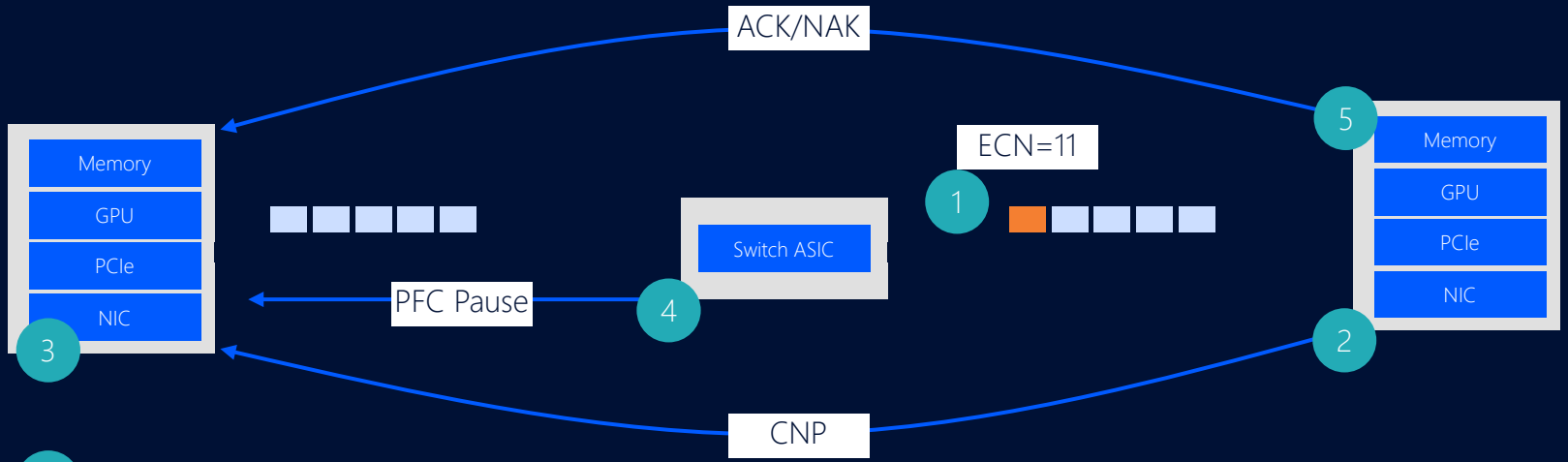
* Assume 8xGPUs + 8x400G NIC ports per server

Ethernet Backend Fabric Solution



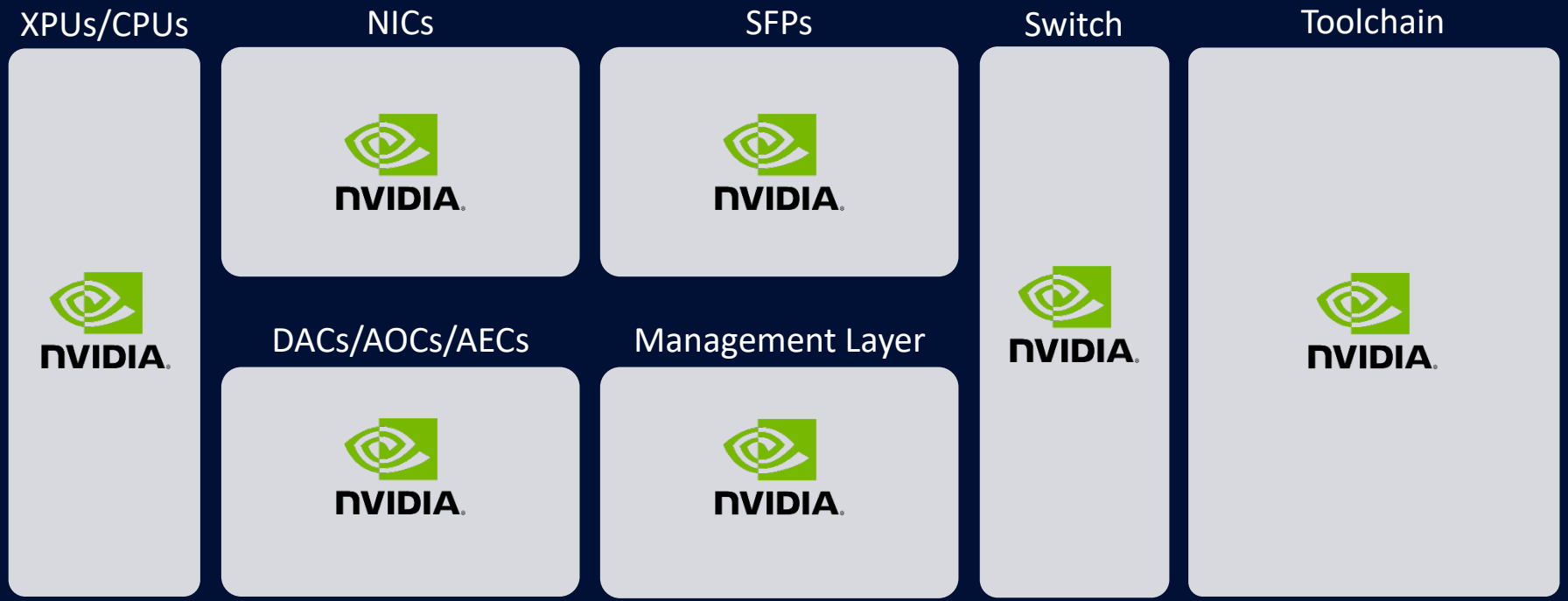
ROCEv2 is a method of sending RDMA messages using UDP/IP/Ethernet encapsulation

Ethernet Backend Fabric Solution – Congestion Control

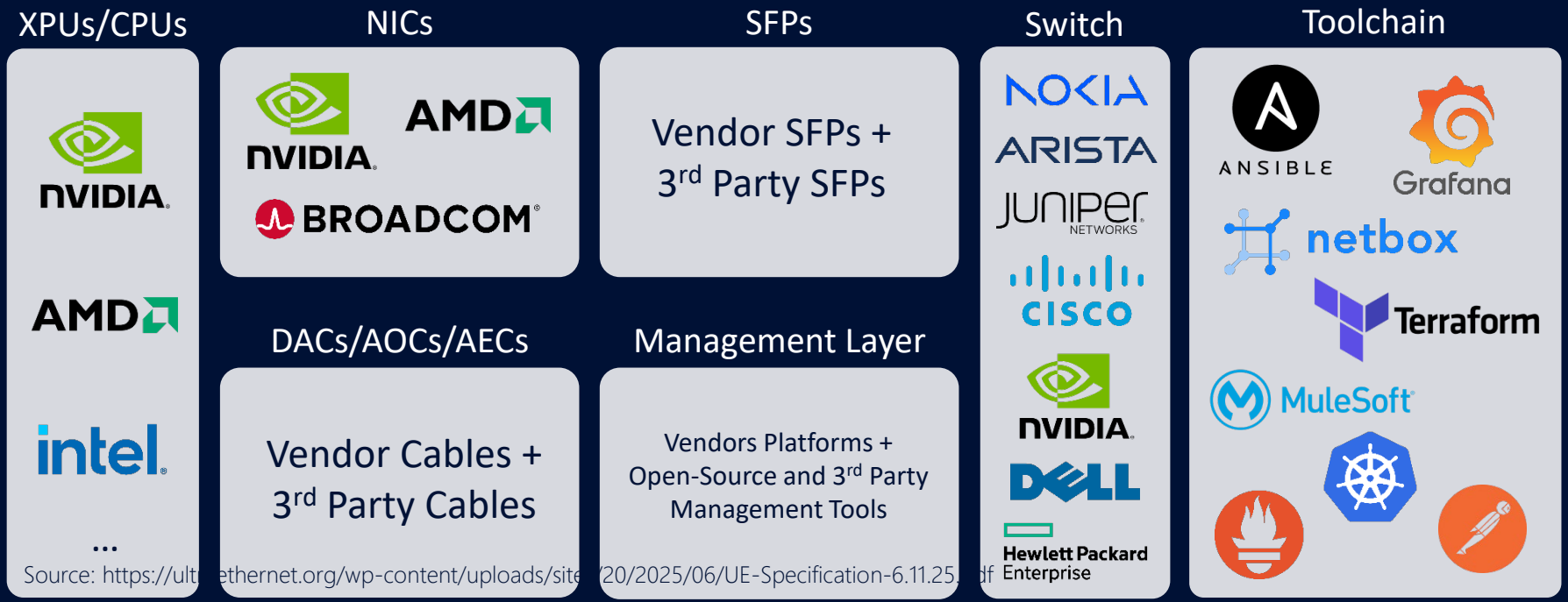


1. Leafs/spines selectively mark IP packets with ECN=11 to indicate they experienced congestion
2. Receiving RDMA NIC sends ROCEv2 CNP (Congestion Notification Packet) to sending NIC
3. Sending RDMA NIC (running DCQCN) slows its TX rate in response to received CNPs.
4. If 1-3 was not sufficient and queues continue to build, leafs/spines send PFC pause frames to link neighbors
5. If 1-4 was not sufficient and loss occurs, receiving RDMA NIC sends NAK and retransmission occurs

The InfiniBand Lock-in Problem



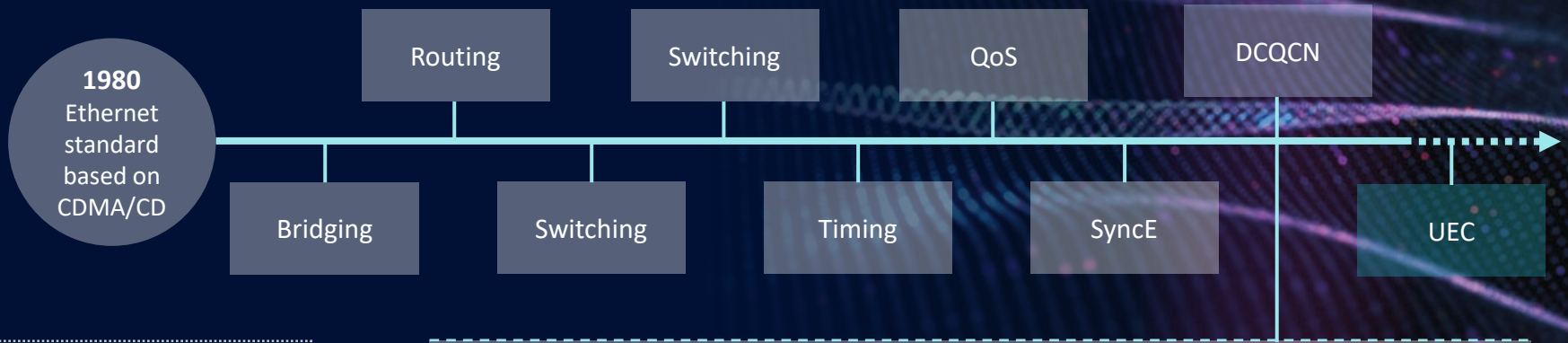
Ethernet’s Ecosystem Advantage



Source: https://ult ethernet.org/wp-content/uploads/site /20/2025/06/UE-Specification-6.11.25. df

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