HPE JUNIPER networking

AI Data Center Networking

Design and Advanced Load Balancing

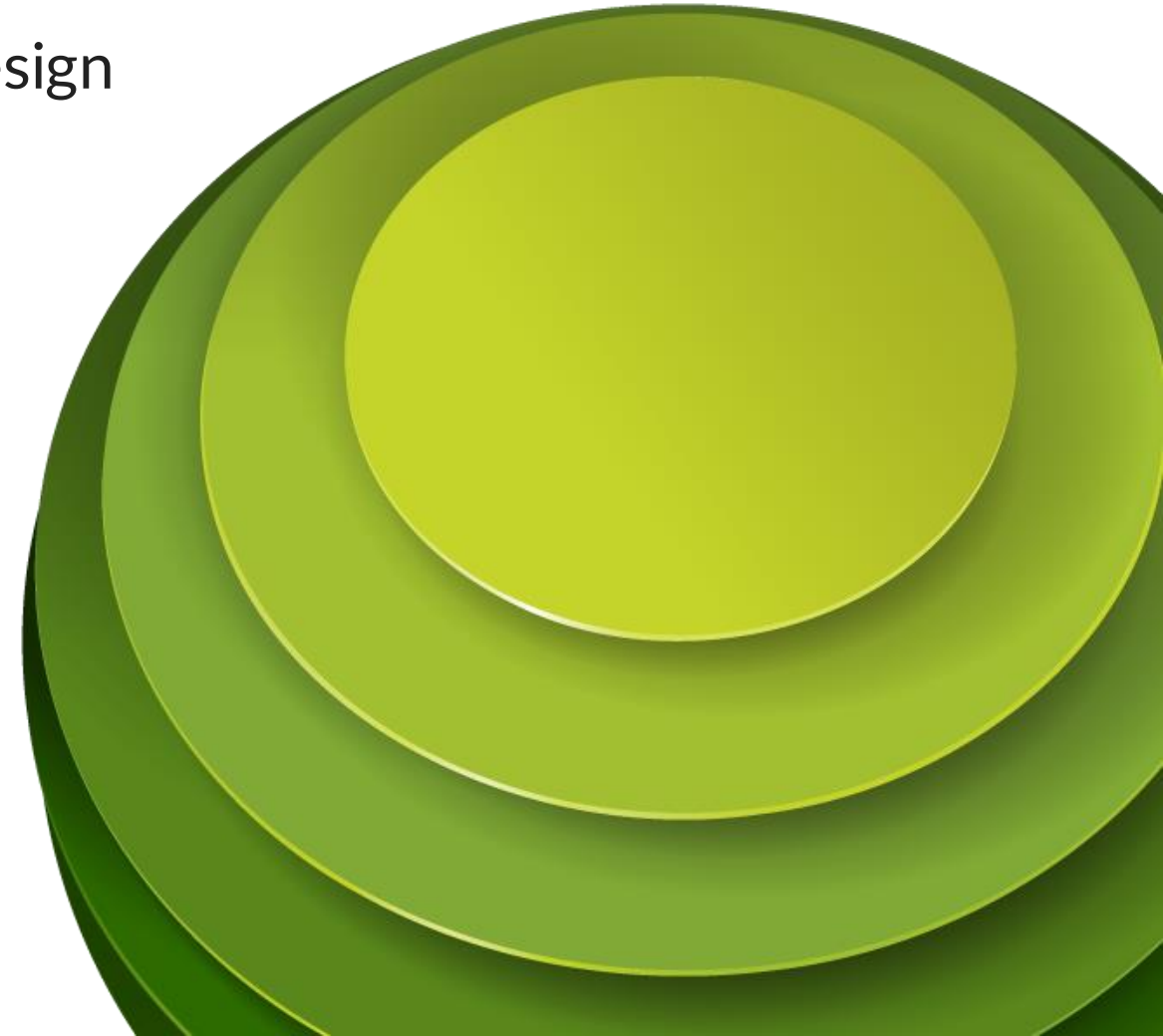
16th SIG-NGN Meeting, in collaboration with SIG-AI

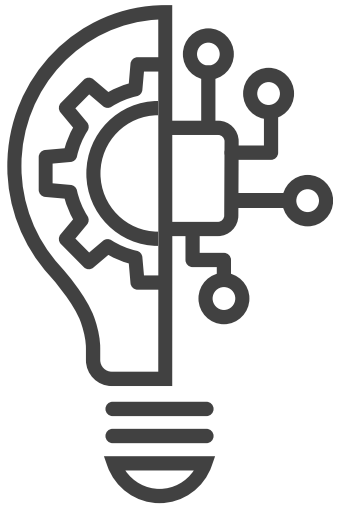
OCT 10th, 2025, KIT - Karlsruhe, Germany

Michal Styszynski
Director, DC Product Management

Agenda:

- AI/ML Scale-out DC key building blocks & design
- Advanced Load Balancing for ROCEv2
- UEC and Load Balancing
- LB & Multi-tenancy options in the AI/DC
- Key takeaways





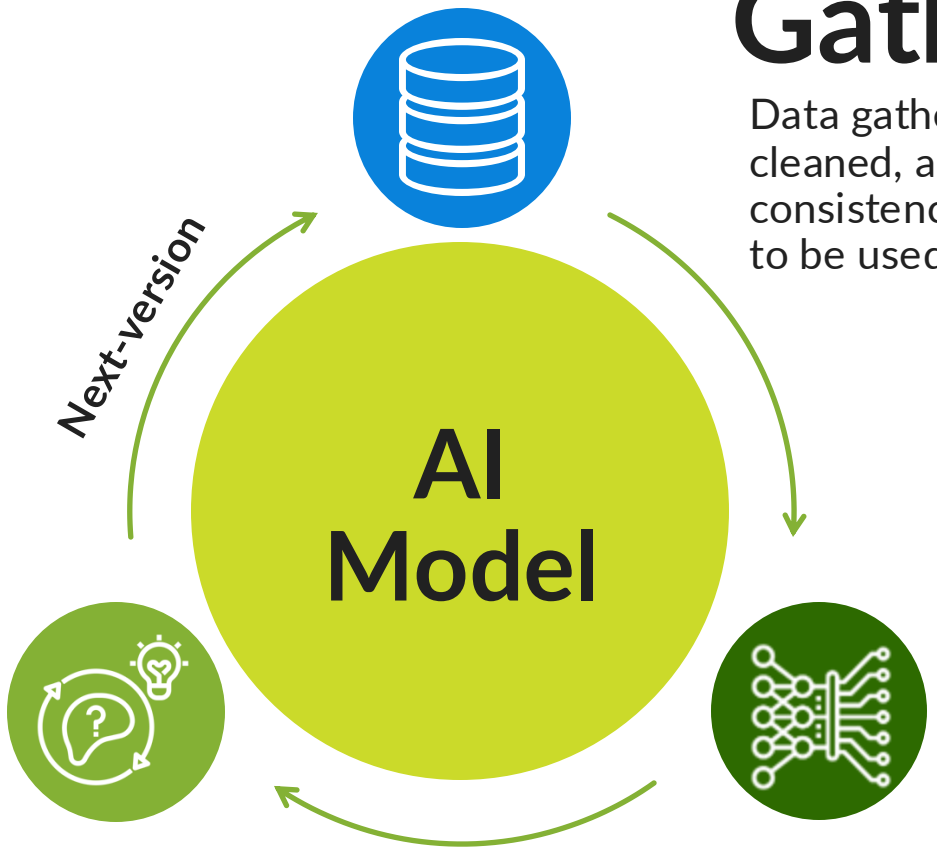
AI Data Centers

Introduction and design options

Lifecycle of an AI Model

Gather data

Data gathered from various sources is cleaned, and verified for reliability and consistency. It is then prepared and curated to be used by the training model.



Training

AI model is trained with the curated dataset and deep learning framework on GPU clusters.

Inference

The trained model is deployed on inference clusters to provide actionable outcomes from user inputs.

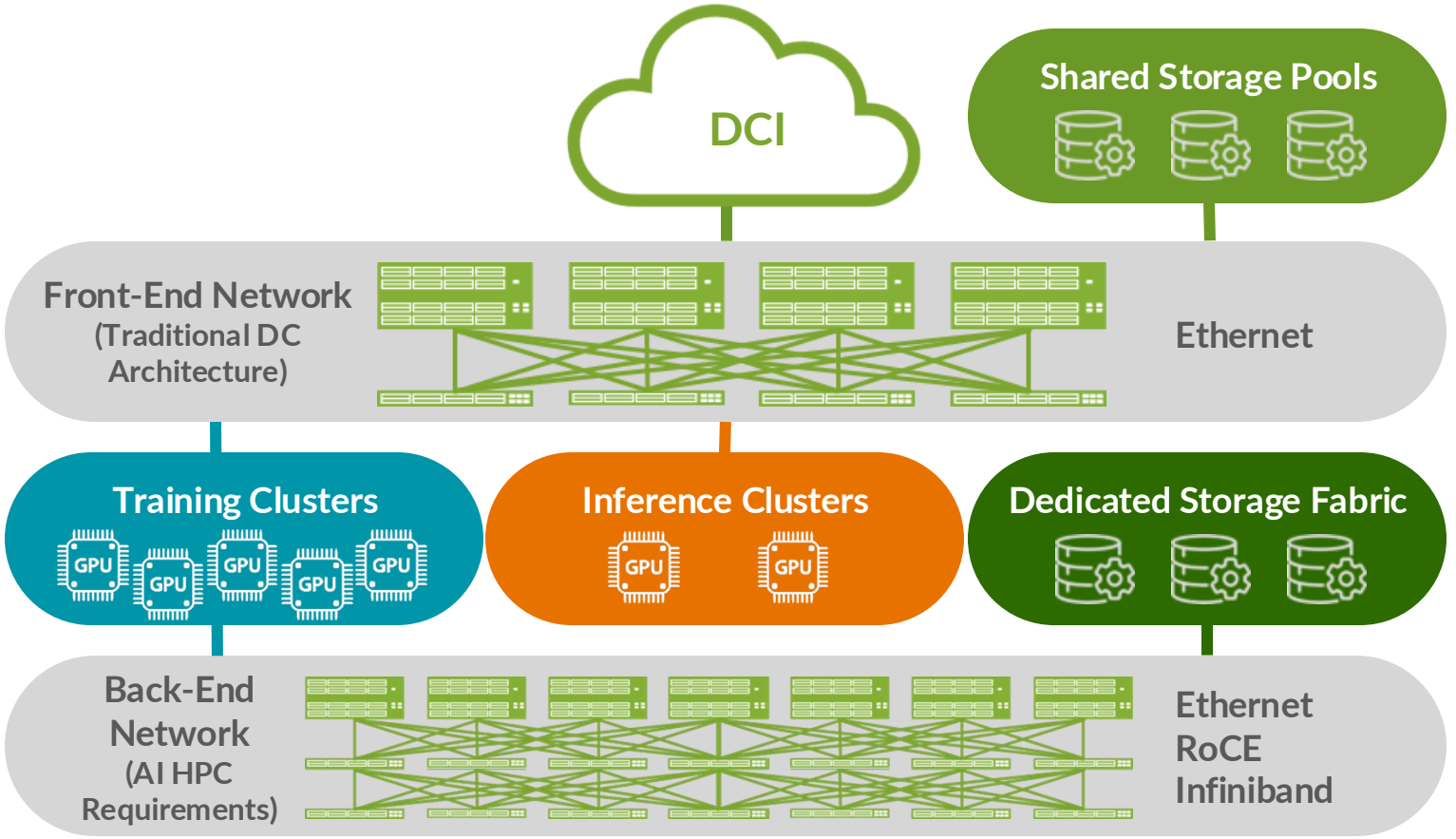
The AI/ML app lifecycle can be a continuous, iterative process of designing and developing models, training and validating them with curated data, and deploying them into production while monitoring their performance for constant refinement and improvement.

AI/ML DC Network Infrastructure - Building Blocks

Front-End Networks
- Inference clusters
- Shared storage pools
- Used for access, data movement, etc.

Back-End Networks
- GPU Compute Fabric
  - Connects GPUs/Compute in training clusters and storage clusters
- Dedicated Storage Fabric
  - May be sometimes converged with Compute

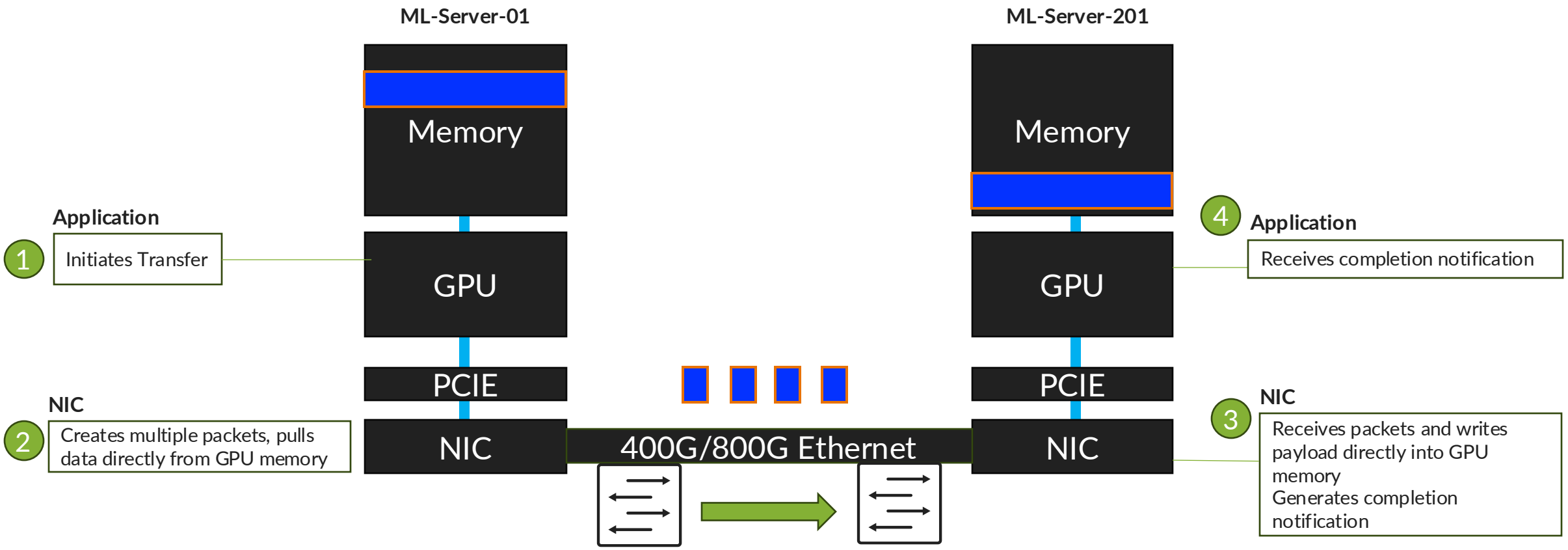
OOB Management Network



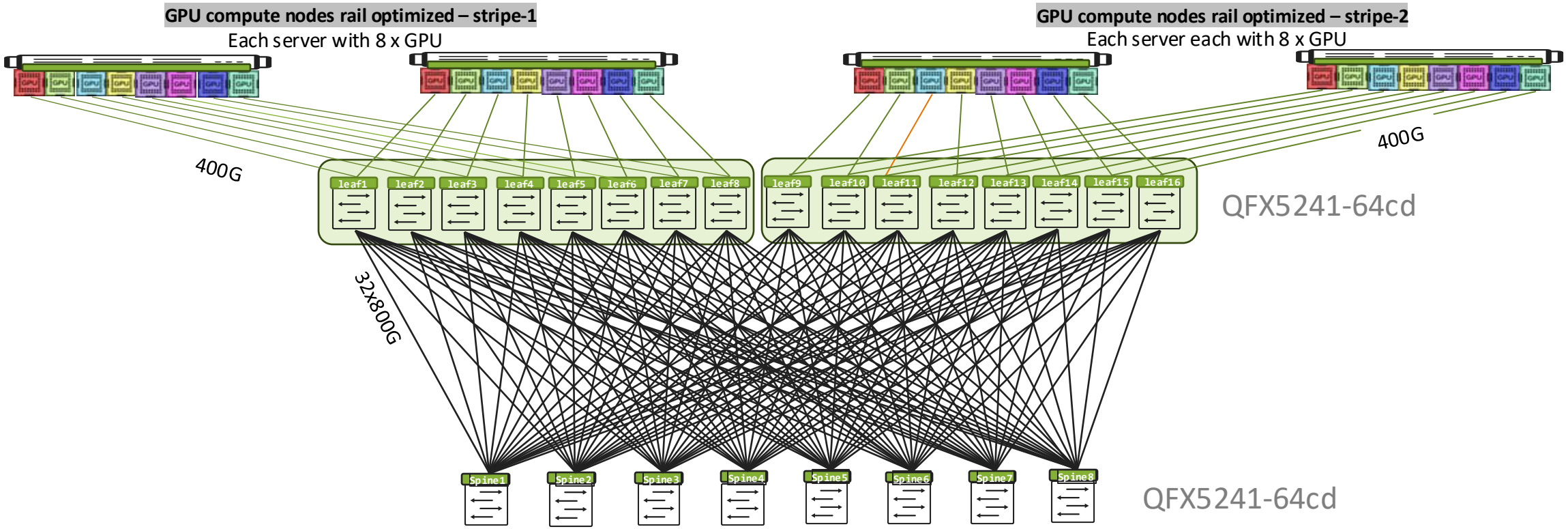
AI front-end and back-end networks have different, unique requirements

HPE Juniper networking

RDMA Transfer using ROCEv2



AI DC – Rail Optimized Design - scale out architecture



512 GPU Cluster, QFX5241-64OD Leaf/Spine - 1:1

Leaf:
- QFX5241-64OD/QD
- Interfaces:
  - 64 x 400G GPU facing
  - 64 x 400G Fabric facing

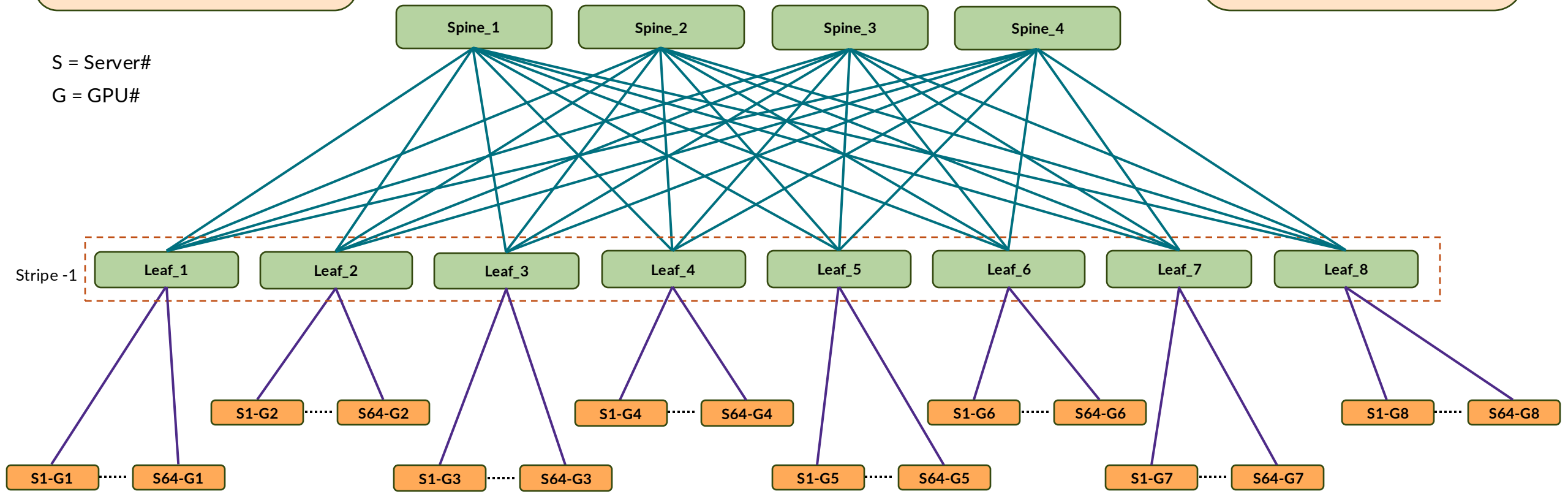
16 x 400G

1 x 400G

Spine:
- Four wide spine build
- 128 x 400G per switch using QFX5241-64OD ToR and Spine

S = Server#

G = GPU#



HPE juniper networking

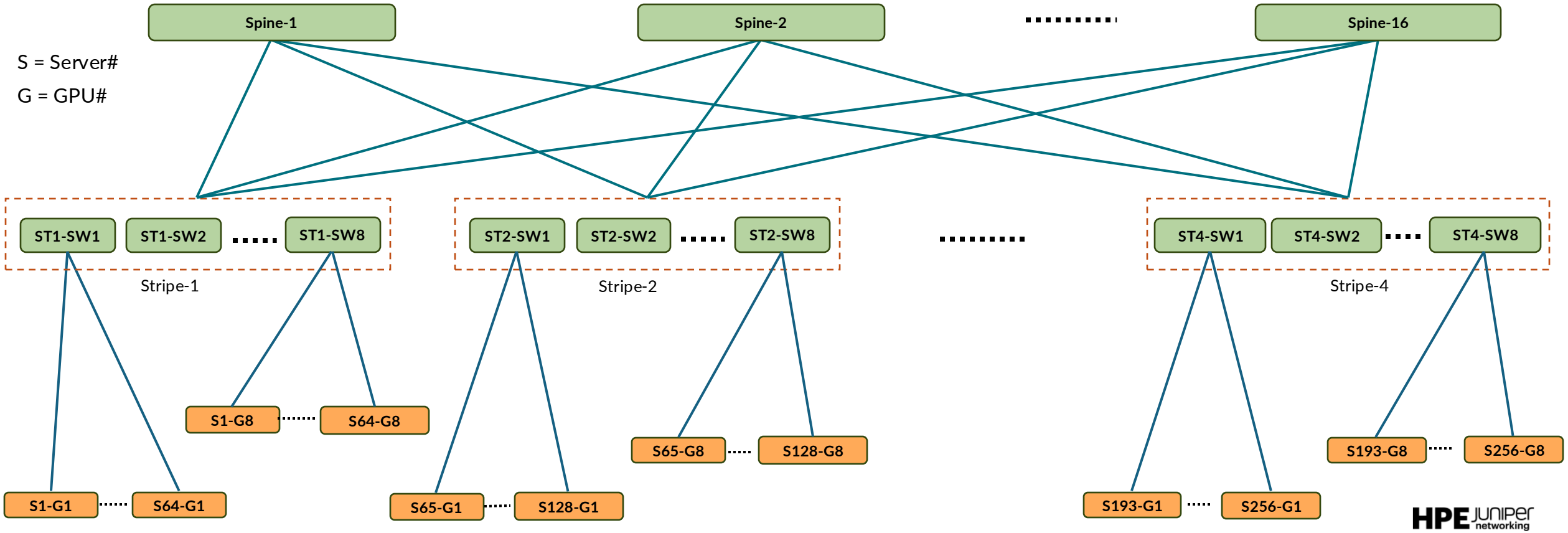
2048 GPU Cluster, QFX5241-64OD Leaf/Spine 1:1 (non oversubscribed)

Leaf:
- QFX5241-64OD/QD x 8 per Stripe
- 4 x Stripes (each stripe with 8. * QFX5241-64OD)
- Total of 32 x QFX5241-64OD leaf/ToR
- Interfaces:
  - 64 x 400G GPU facing
  - 32 x 800G Fabric facing

4 x 400G

1 x 400G

Spine:
- 16 x QFX5241-64OD/QD spines
- Sixteen wide spine build
- 64 x 800G per switch



HPE JUNIPER networking

AI Data Center Networking

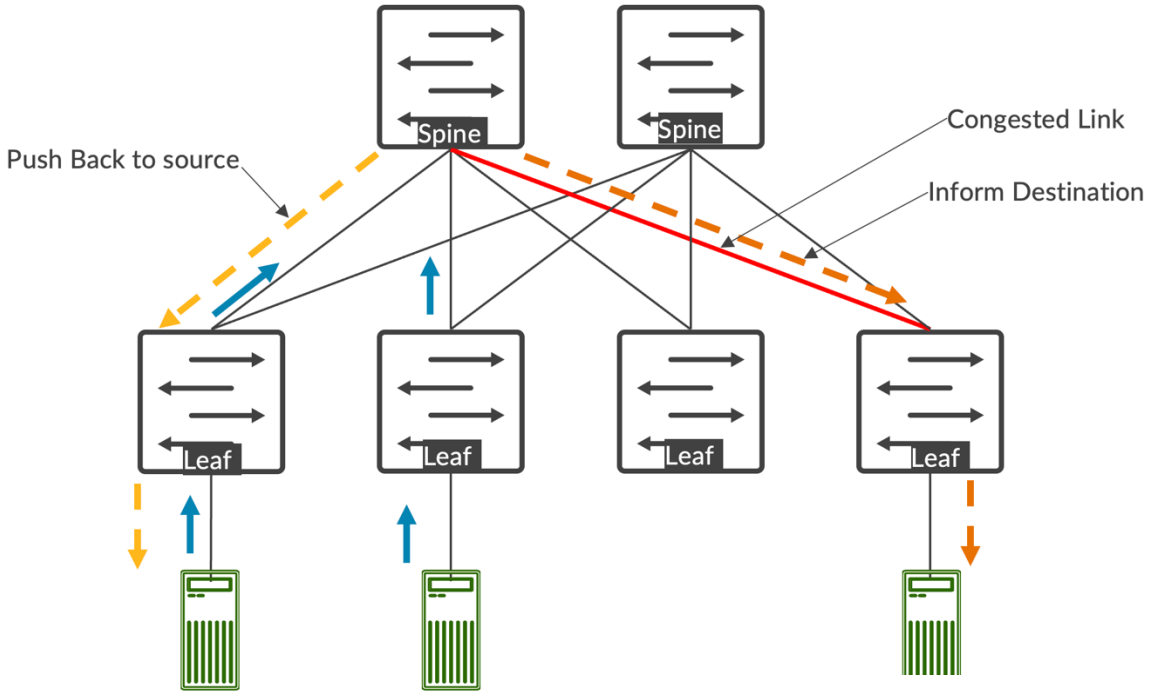
Advanced Load Balancing

HPE JUNIPER networking

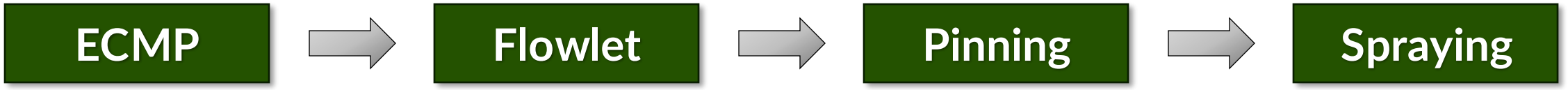
AI DC workloads characteristics and challenges

- Small number of flows per ToR switch port
- Low entropy of the ROCEv2 flows (same destination UDP)
- Long live elephant flows

- It may lead to:
  - Congestions at the spines
  - Longer Job Completion Time (JCT)
  - Tail Latency
  - Restart the given job execution
  - PFC regular pushbacks



Load Balancing Evolution



Static ECMP

Standard ECMP supported by everyone

DLB
(Dynamic Load Balancing - FLOWLET)

GLB
(Globa; Load Balancing)

- Dynamic Load Balancing
- Global Load Balancing

(GLB requires >TH5 hardware)

RLB
(RDMA-aware Load Balancing)

- RDMA-Aware Load Balancing

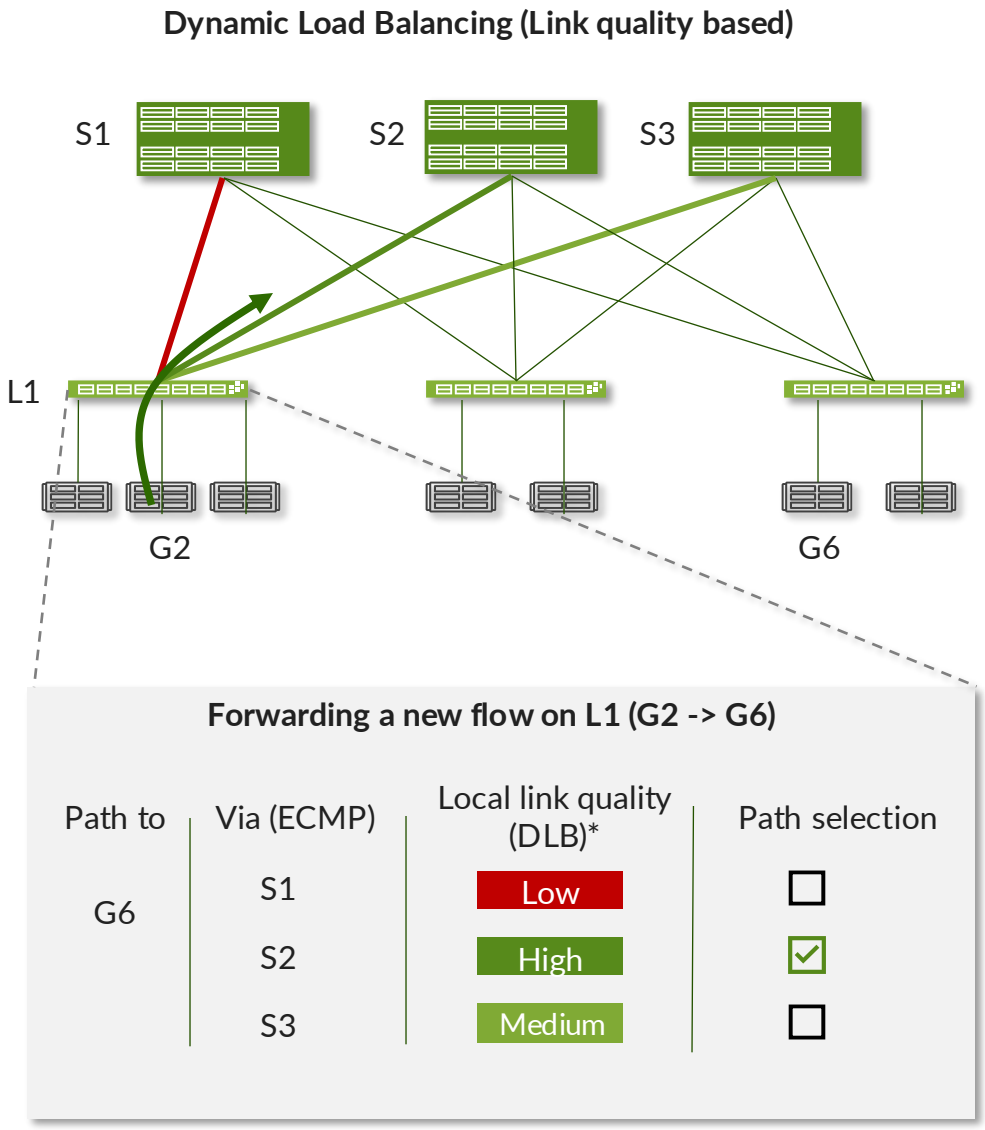
DLB
(Dynamic Load Balancing – PACKET SPRAY)

WPS
(Weighted Packet Spray)

- Dynamic Load Balancing using Packet Spraying
- Weighted Packet Spray

Dynamic Load Balancing (DLB)

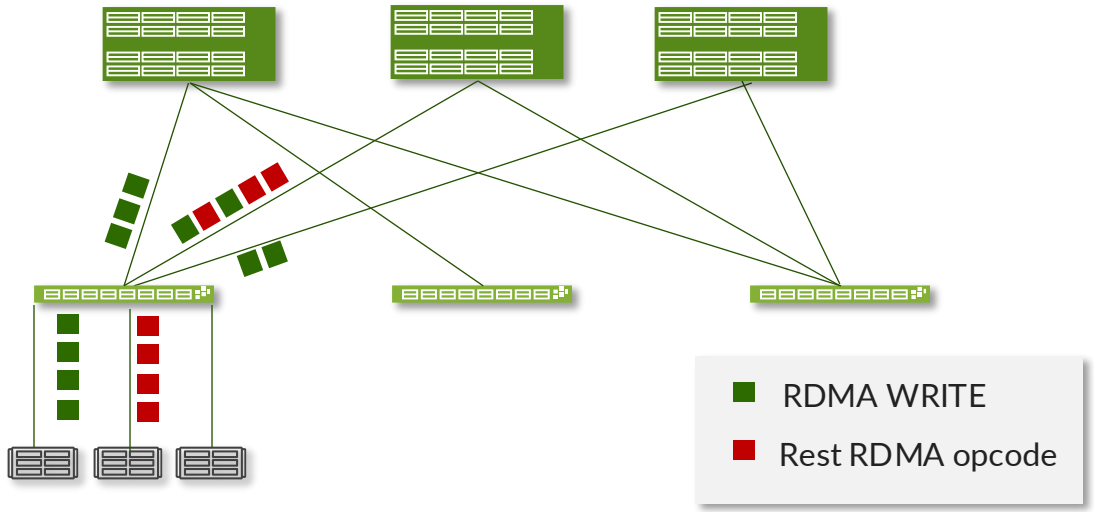
- DLB monitors link quality on local ECMP paths based on real-time link utilization and queue depths of each link.
- DLB factors local link quality and forwards on the least loaded path.
- Minimizes flow collisions and the consequent packet drop.
- DLB supports two modes
  - **Flowlet mode -** assigns links based on *flowlets* instead of flows. Flowlets are multiple bursts of the same flow separated by a period of inactivity between these bursts.
  - **Packet Spray mode -** DLB is initiated for each packet in the flow. Best for load-balancing traffic across elephant and mice flows and achieves fair BW utilization. Requires support from NIC to re-order



* Colors & High, med, and low are used for simplifica
Actual link quality is a numerical metric.

DLB – Dynamic Load Balancing – customization

If Customer uses NICs which support reordering of RDMA write verbs. Hence the need for an efficient Load Balancing technique that takes advantage of this feature.



Packet Spray for RDMA write flows

- Packet Spray makes efficient use of links if the destination NICs can handle out of order packets.
- In this scenario, since RDMA write flows can be received out of order, we would want the RDMA write flows to follow packet spray. But we would want the rest of the RDMA flows to follow SLB so that these are not re-ordered.
- *With Packet Spray, we can look into BTH header and match on Opcode/QPAIR using UDF and enable spray only for them.*

ACL can match the ROCEv2 OP-code and particular QPAIR & enable or disable the packet spraying

Global Load Balancing (GLB)

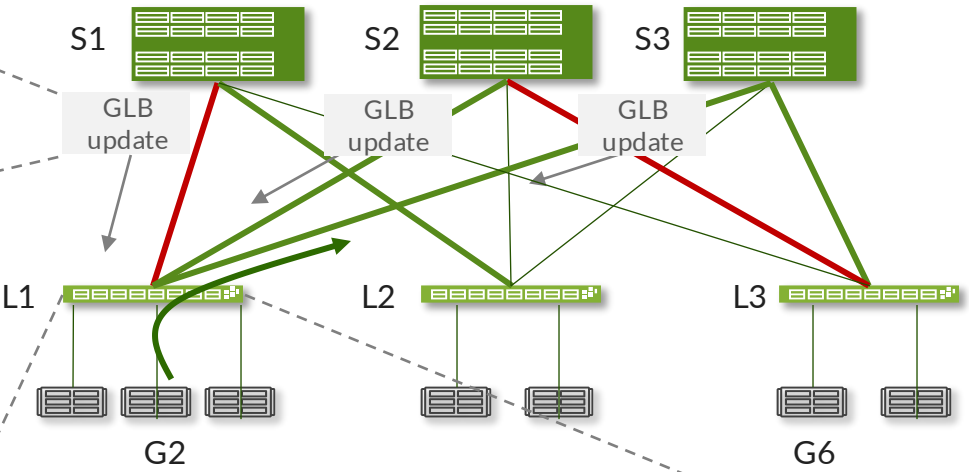
- Spines monitor local link quality to all leaves (DLB) and advertise these to all upstream leaves
- Leaves use received link quality from spines along with local link quality (DLB) for load balancing
- Upstream switches avoid downstream congested paths and select a better end-to-end path
- Like Google Maps traffic condition-based path selection

Dynamic Load Balancing (Link quality based)

GLB from S1	
To L2	High
To L3	High

GLB from S2	
To L2	High
To L3	Low

GLB from S3	
To L2	High
To L3	High



Forwarding a new flow on L1 (G2 -> G6)

Path to	Via (ECMP)	Local LQ (DLB)*	Global LQ (GLB)*	Path selection
G6	S1	Low	High	☐
	S2	High	Low	☐
	S3	High	High	☑

Like google maps traffic

* Colors & high, med, and low are used for simplification.
Actual link quality is a numerical metric.

HPE JUNIPER networking

RLB : RDMA-aware LB

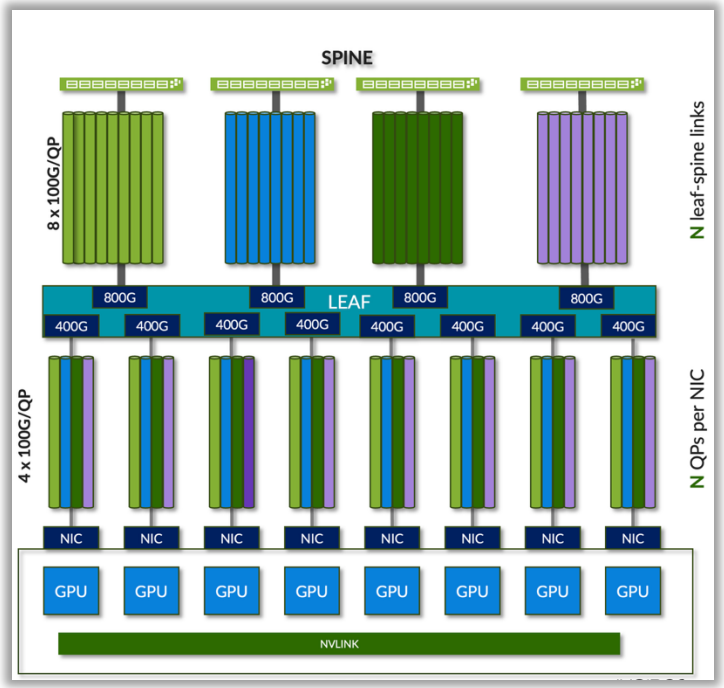
Highest Consistent Performance

Eliminates Congestion in most scenarios

In-order Delivery

Delivers packets in order improving JCT

RDMA-aware Load Balancing



Expensive HW not required

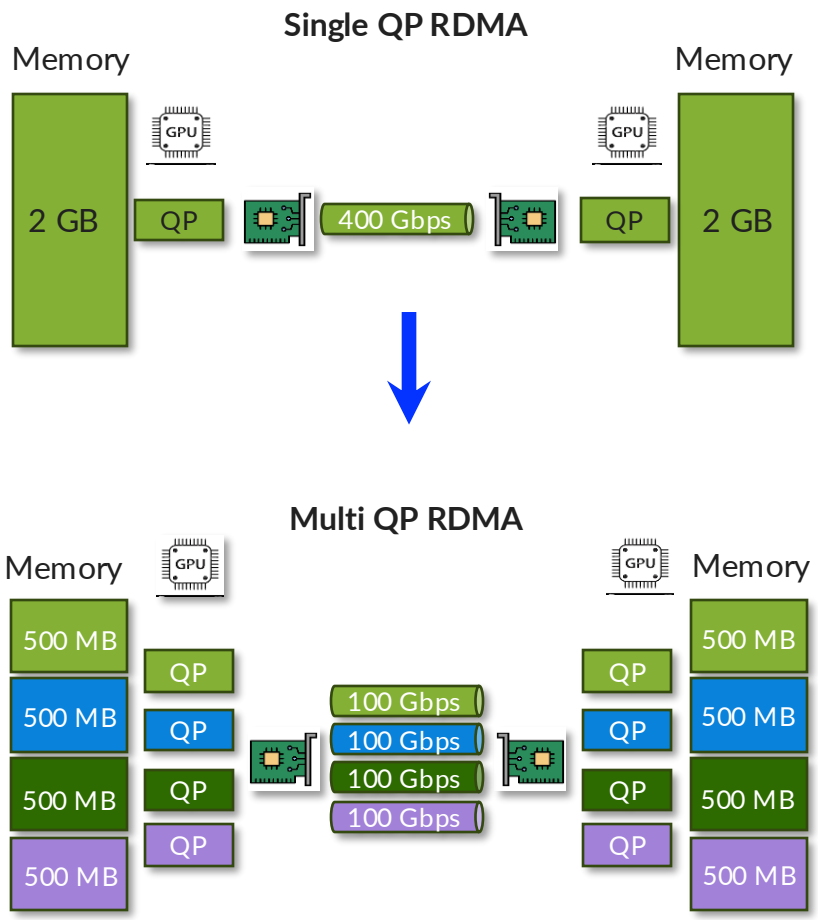
Doesn't require expensive HW (SuperNICs or deep-buffer switches)

Delivers non-rail performance

Delivers highest performance even for non-rail, lowering TCO.

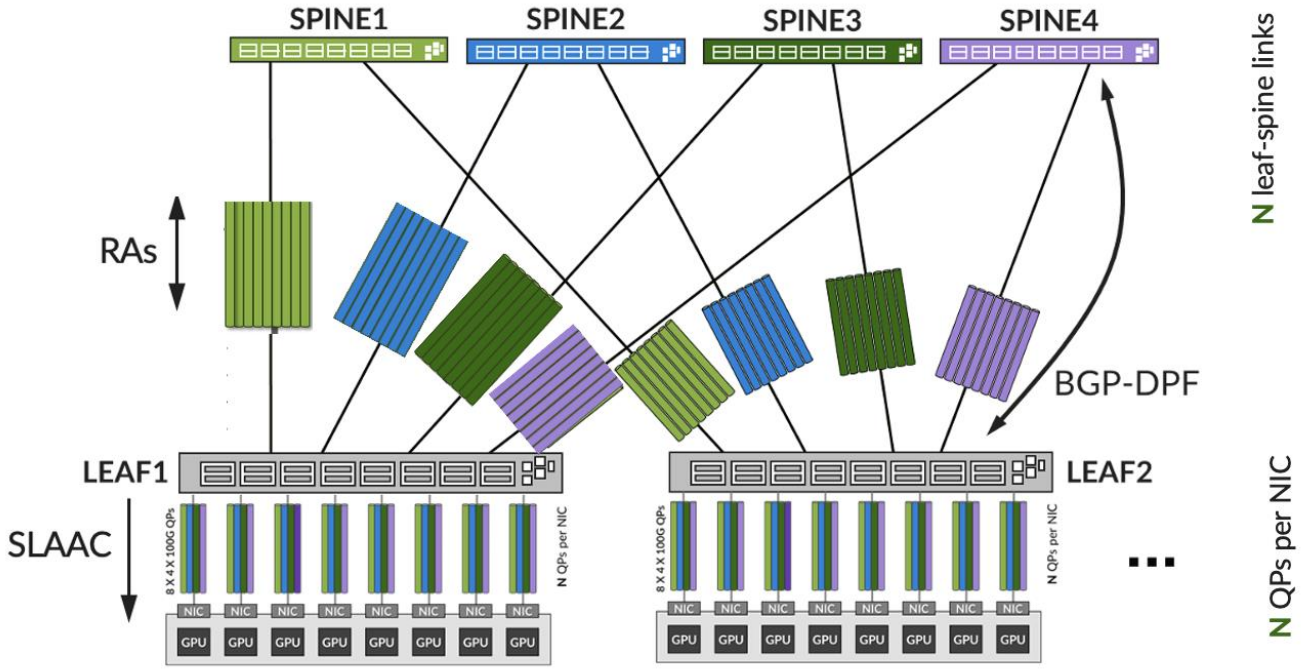
HPE juniper networking

RLB : RDMA-aware LB



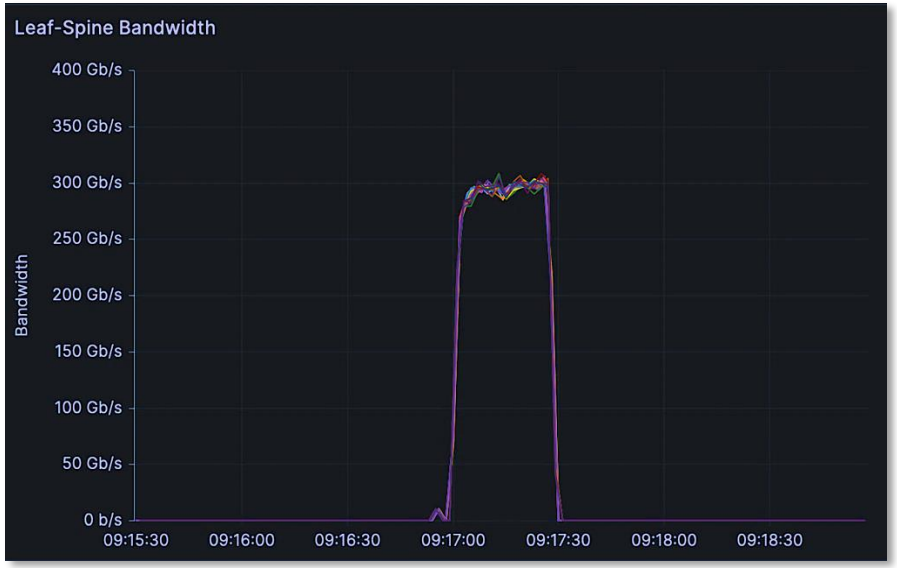
Num of QPs = Num of leaf-spine links = N = 4

HPE Juniper Networking - NCCL plugin sets up N QPs per GPU

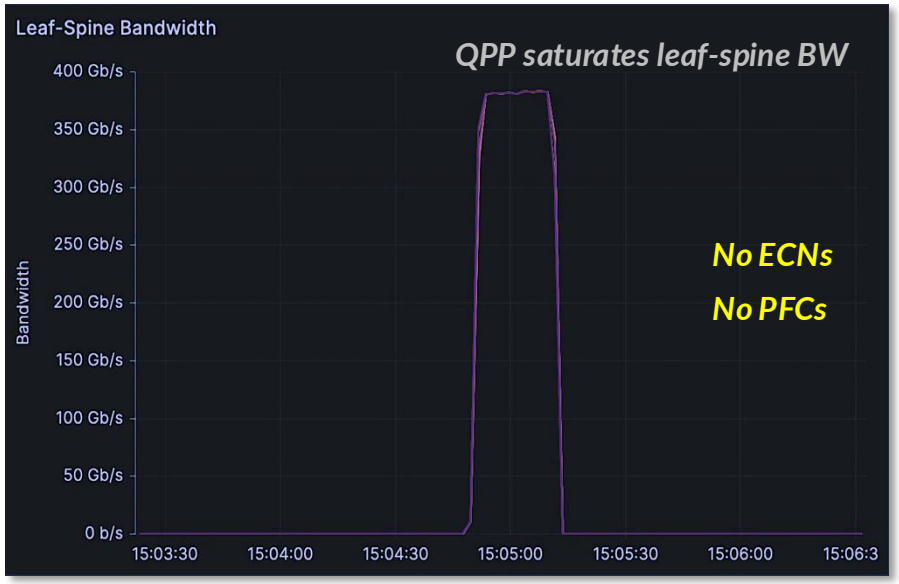


DLB vs. RLB performance

DLB



RLB



HPE JUNIPER networking

RDMA Load Balancing – using BGP-DPF

```
root@TOR101# show protocols bgp
fabric-advertise {
    route 2001:1:1::/64 {
        color red;
        backup-color lightred;
    }
    route 2001:2:1::/64 {
        color blue;
        backup-color lightblue;
    }
}
group underlay-red {
    type external;
    peer-as 65001;
    fabric-color red;
}
group underlay-blue {
    type external;
    peer-as 65002;
    fabric-color blue;
}
multipath;
root@TOR101#
```

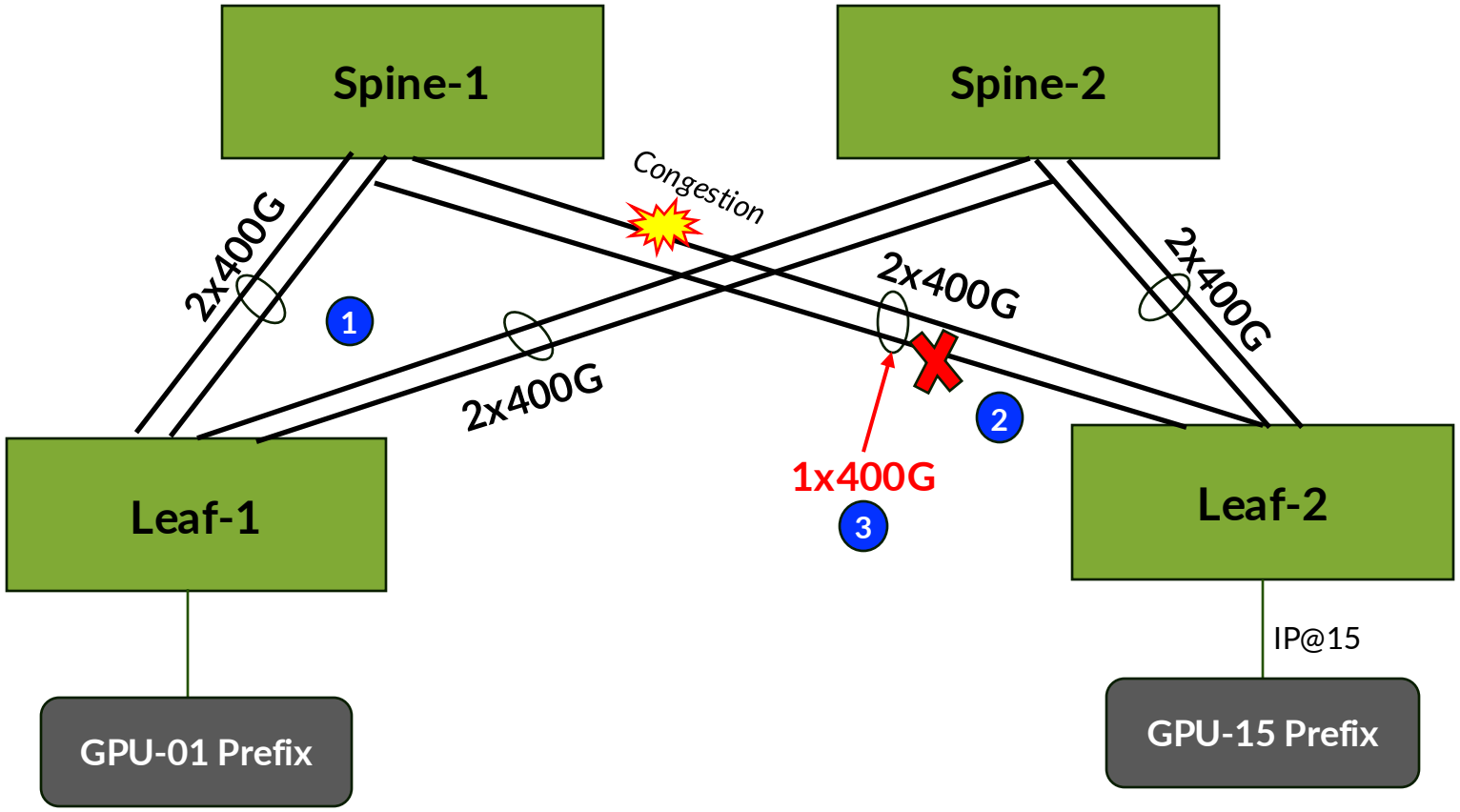
explicit backup colored path can be defined
OR all-paths backup with DLB can be used

x 16 QPAIRs per GPU server port == 16 x DPF colors

- If 16 x SPINES/links in the topology

HPE JUNIPER networking

Weighted Packet Spray



1 Normal Condition

- All leaf-to-spine connections use 2×400GE links.
- Spine-1 and Spine-2 each advertise the GPU-15 prefix to Leaf-1 with 800G aggregate bandwidth.
- Leaf-1 forwards traffic equally to both spines to reach GPU-15 **(1:1 ratio)**.

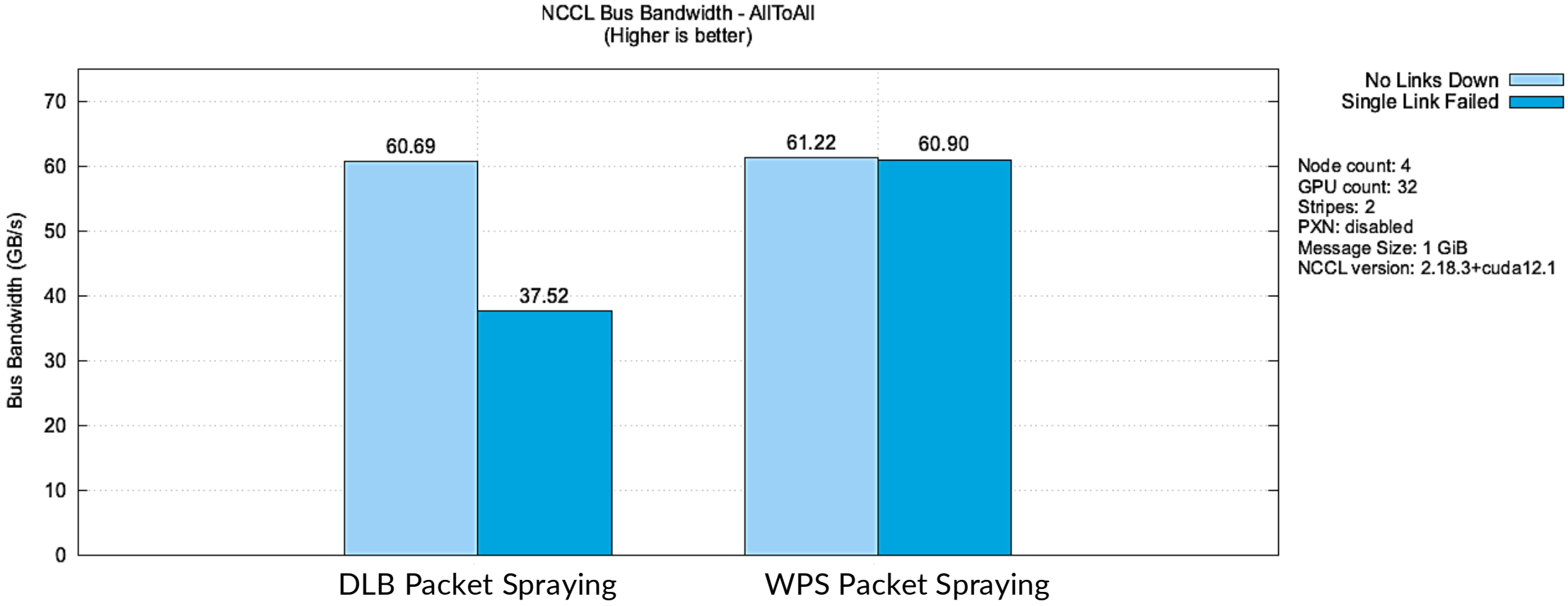
2 Failure Without WPS

- If the link between Spine-1 and Leaf-2 fails:
- Leaf-1 continues equal traffic distribution to both spines. Leaf-1 is unaware of reduced bandwidth on Spine-1.
- This will lead to congestion and packet loss on the remaining link.

3 Failure With WPS

- Spine-1 advertises the GPU-15 prefix to Leaf-1 as 400G (Reduced/reflecting the failed link).
- Spine-2 still advertises the same 800G.
- So, Leaf-1 adjusts forwarding to a **1:2 ratio** (400G to Spine-1, 800G to Spine-2)
- Result: Improved traffic distribution, reduced congestion, and optimized performance.

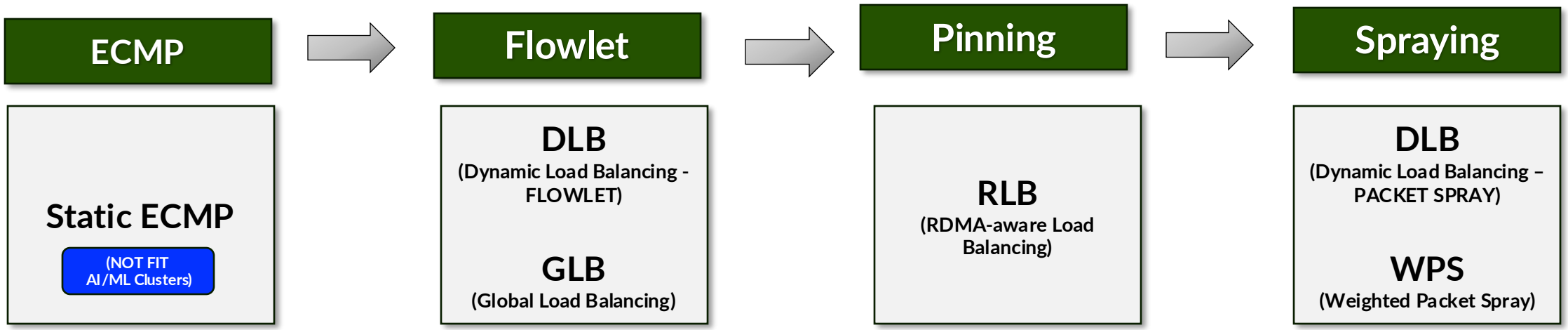
Weighted Packet Spraying - Link Failure Scenario



Caveats: Out-of-order packet reception capabilities are required at the NIC

HPE JUNIPER networking

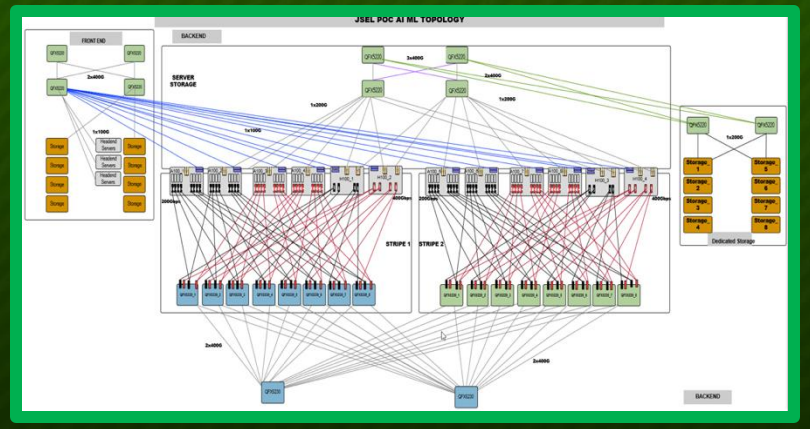
Summary : AI Load Balancing



	DLB-FLOWLET Dynamic Load Balancing	GLB Global Load Balancing	RLB RDMA-aware Load Balancing	Packet Spray / WPS Weighted Packet Spraying
Cluster Size	Any	Scaling limitation 64 Qual Profile (For now)	Medium to Large clusters	Any
Consideration	None	QFX5240 only Fabric and Upcoming TH6 based Fabric	Best for workloads that don't have a heavy All-To-All requirement (for now)	None
NIC Type	Any NICs	Any NICs	Any NICs	Advanced NICs (Need Re-ordering)

AI Optimized Ethernet: Put to the test

We achieve comparable results with InfiniBand using the MLPerf benchmarks



AI Cluster

- *64 A100 and 32 H100 GPU's, Rail optimized architecture*
- *High Performance 400G Networking with AI Optimized Ethernet*
- *Run LLM tests with Bert-Large, DLRM*

AI Optimized Ethernet: Matching InfiniBand Performance

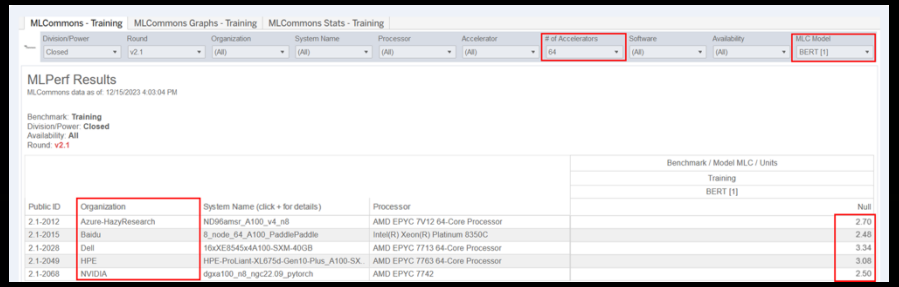
BERT-Large

2.52 min

Juniper Ethernet Training Time

2.5-3.3 min

Other benchmarks posted (including IB)



UEC & Load Balancing

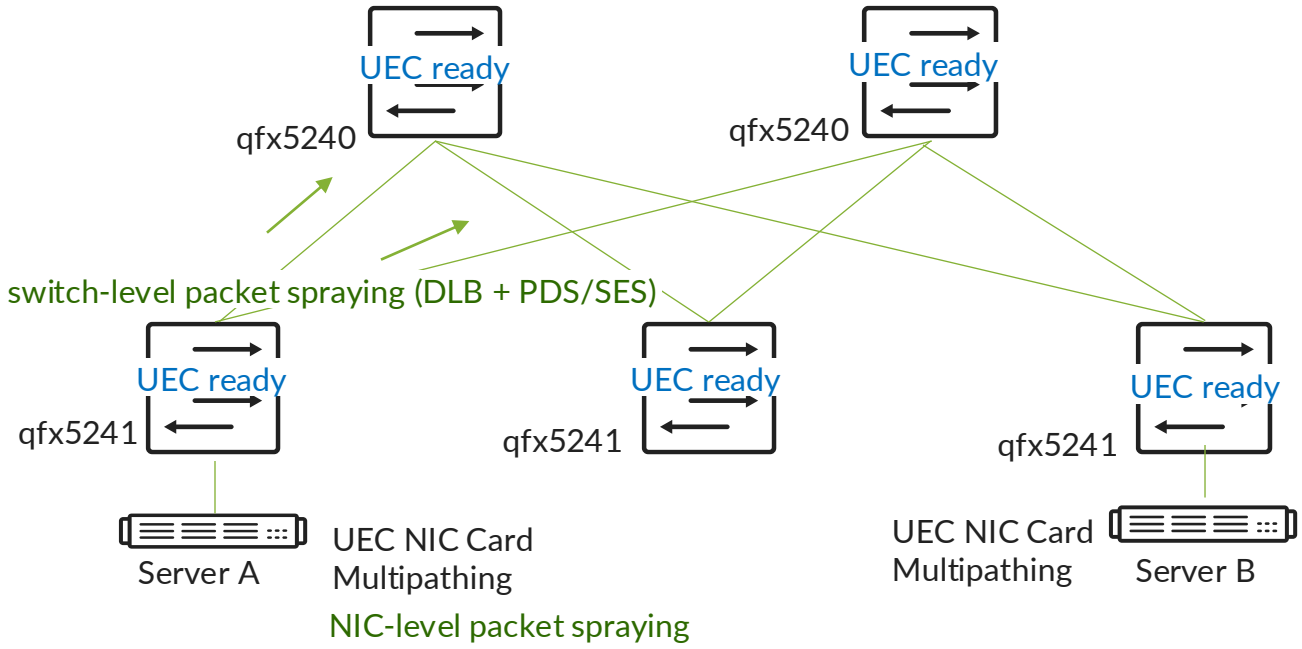
Ultra Ethernet Consortium

UEC – 10,000 Ft View

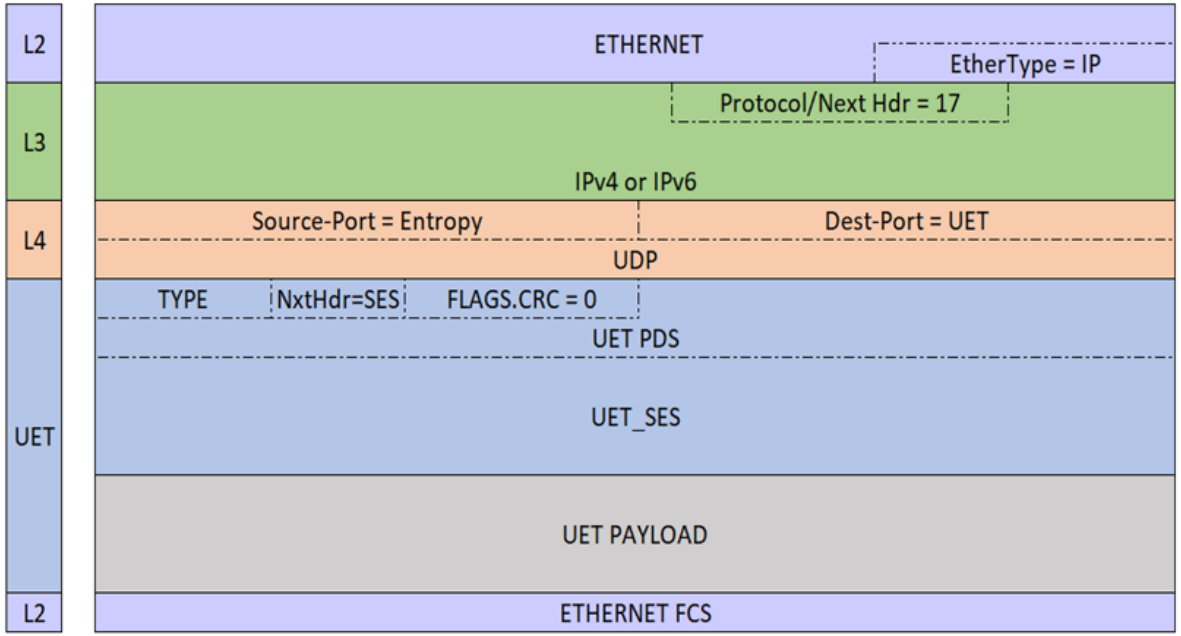
	InfiniBand	Lossless Ethernet Fabric HPE-JUNIPER	Ultra Ethernet Consortium
Congestion Avoidance Effective utilization of fabric capacity	Central control	• Distributed architecture • Dynamic Routing • Intelligent Load Balancing (DLB, GLB, sDLB etc.)	• Dynamic Routing • Packet Spraying • Flexible Re-ordering on NIC • In Network Collectives (INC)
Congestion Control Preventing packet drops (lossless)	Credit Request/Grant mechanism	DCQCN (PFC + ECN)	• Link-level credit mechanism (LLR) • Receiver credit mechanism (RCCC) • Sender Network Signal CC (NSCC) • Telemetry based CC (CSIG) • Packet Trimming (DCN)

UET forwarding & Load Balancing

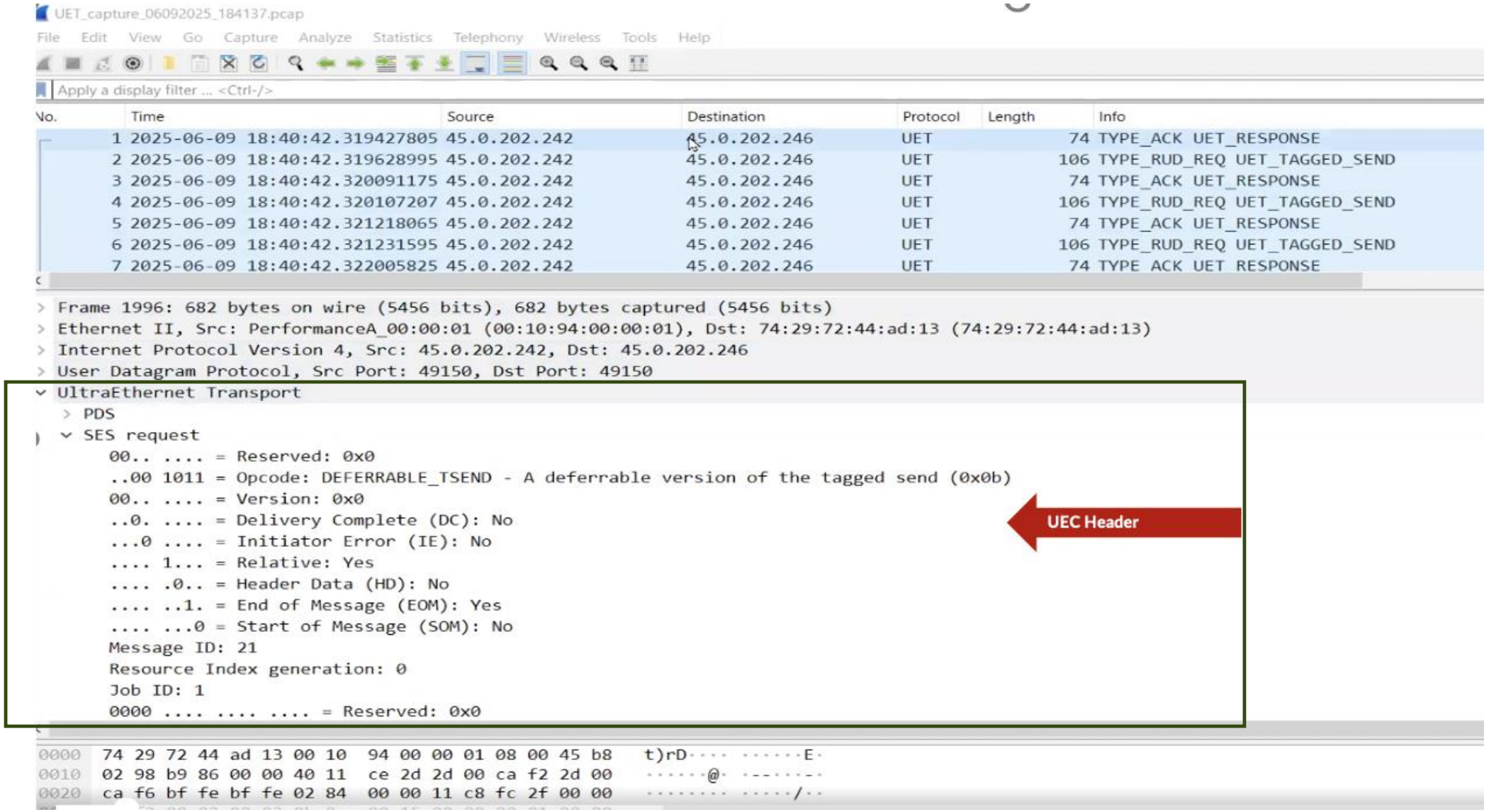
- The UET packet format is UDP encapsulated with PDS and SES header instead of BTH header in ROCEv2 – our QFX5240 and QFX5230 switches will be **UET-ready** to forward the UEC frames
- Sprayed and non-sprayed packets at the NIC card level to be **compatible with our DLB mechanisms – Spray in NIC and Spray in ToR Ethernet switch**



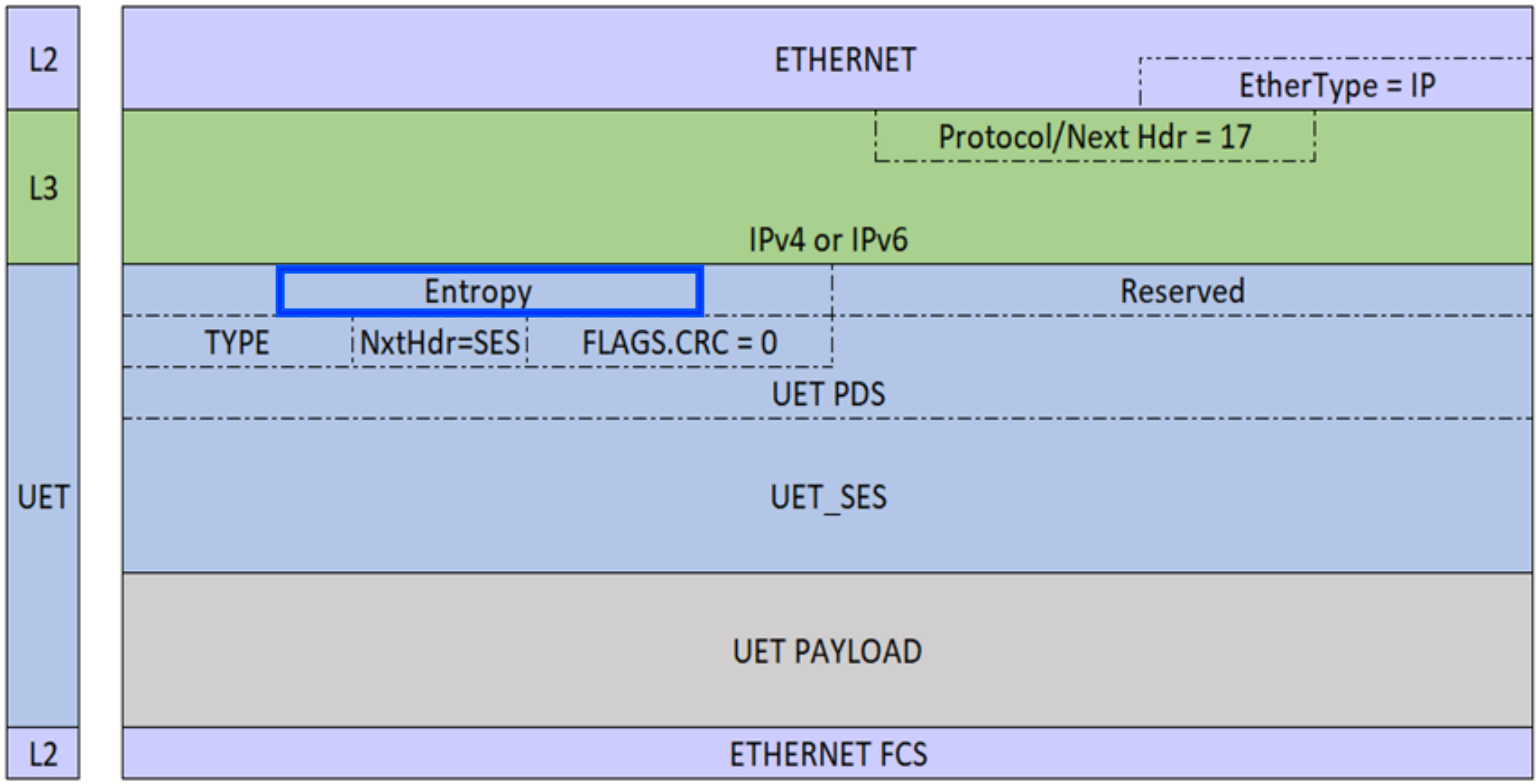
UEC – Transport packet encapsulation
unencrypted with UDP entropy



UEC – new packet format and load balancing



UET over IP encapsulation – new entropy field



- Here the UDP is not used so the overhead is reduced but then the new field ***entropy*** will be used for the packet spraying
- UET PDS will include the PSN (Sequence Number) ACK'd or NACK'd to make the packet delivery reliable
- If the UET uses the NSCC as congestion management, then the NACK will contain the entropy information to let the originator adapt its entropy accordingly – for example, by changing the source UDP

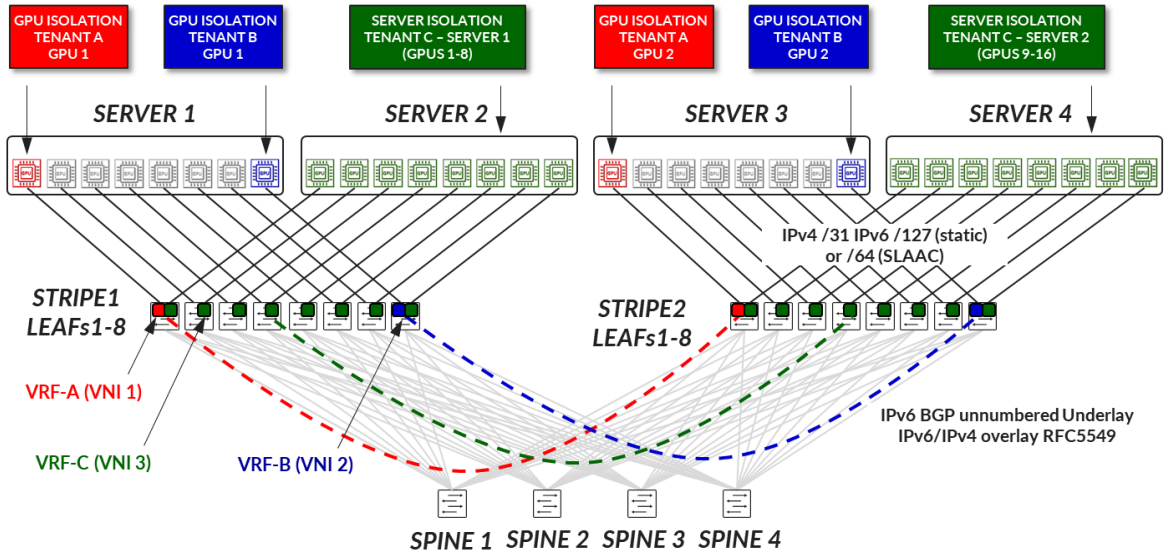
AI DC

Multi-tenancy

AI/ML – Why Multi-tenancy (MT) ?

Why do we need multi-tenancy in the AI/ML clusters?

1. **AI/ML Services on-demand deployments:**
   - GPUaaS
   - BMSaaS
2. **Fabric Security**
3. **Simplified fabric resources control (performance control)**



AI DC Multi-Tenancy (MT) types:

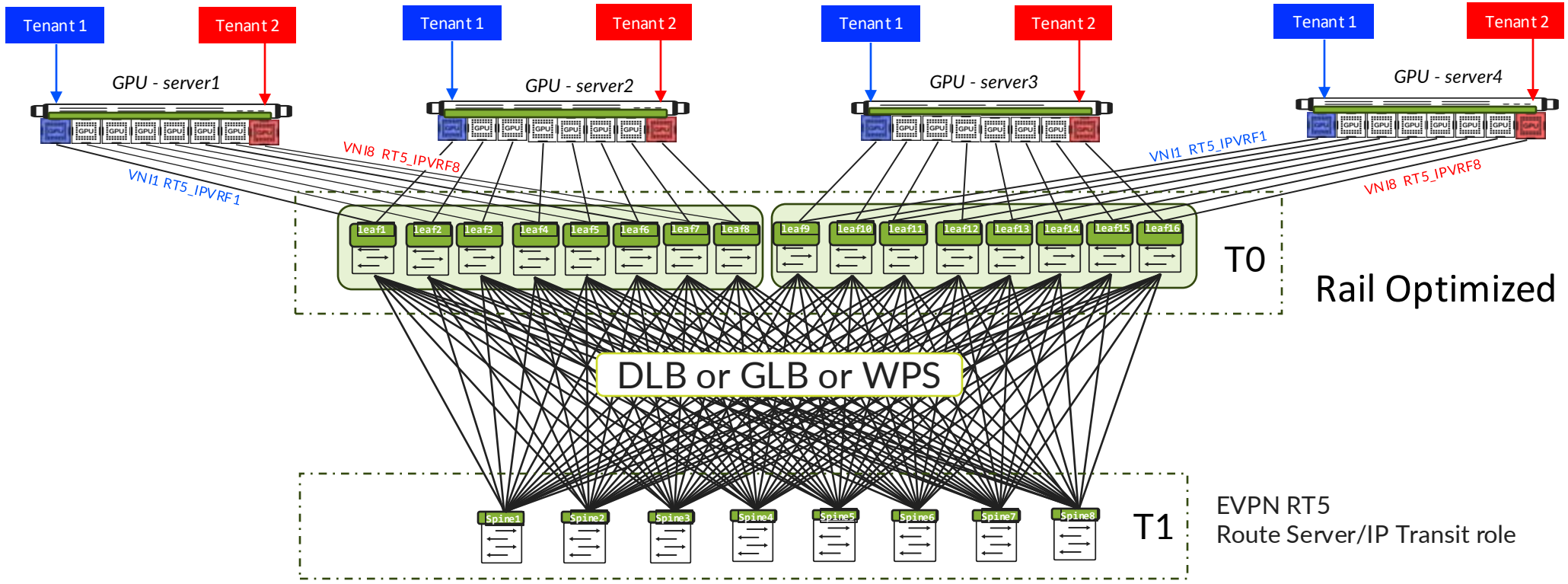
- Server-level Multi-tenancy
- GPU-level Multi-tenancy

AI DC architectures with MT:

- ROD (Rail Optimized Design)
- Non-Rail Optimized

HPE JUNIPER networking

GPU-level EVPN multi-tenancy & Load Balancing



Rail Optimized Design

EVPN RT5
Route Server/IP Transit role

- At the Leaf level:
  - With DLB/GLB:
    - QPAIR/SRC.UDP Hashing
    - Link Bandwidth
    - Queue Depth
  - WPS – Weighted Packet Spraying

- At the Spine level:
  - With DLB/GLB:
    - QPAIR/SRC.UDP Hashing
    - Link Bandwidth
    - Queue Depth
  - WPS – Weighted Packet Spraying

HPE JUNIPER networking

Thank you

HPE JUNIPER networking