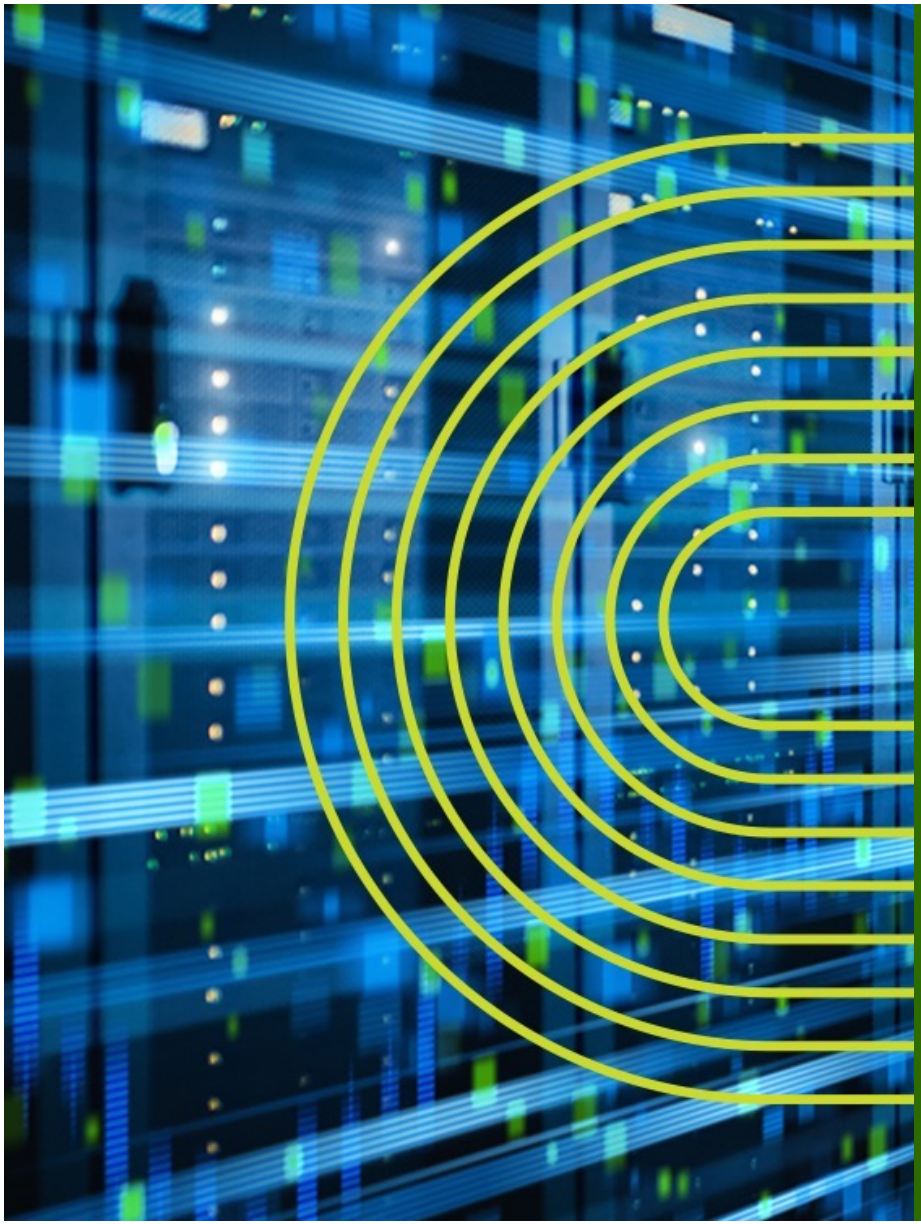
# 400G Ethernet PoC for AI Cluster

Juniper Networks | Driven by Experience™



# ROCEV2 VS INFINIBAND

Juniper Networks

# AI 성능측정

## 네트워크 세상에는 BMT: rfc2544 by 계측기

- AI 세상에는?

**MLperf**

internal

# MLCommons - MLPerf Round: v4.0

*https://mlcommons.org/benchmarks/training/*


Division/Power: Closed | Round: v4.0 | Organization: JuniperNetworks | System Name: (All) | Processor: AMD EPYC 7763 6... | Accelerator: (All) | Accelerators Per Node: 8 | Software: (All) | Availability: Available on-premise | MLC Model: (All)

MLPerf Results

MLCommons data as of: 6/12/2024 1:37:25 PM

Benchmark: **Training**
Division/Power: **Closed**
Availability: **Available on-premise**
Round: **v4.0**

**MLCommons** will help grow machine learning from a research field into a mature industry through benchmarks, public datasets and best practices.

| | | | | | | | | | Benchmark / Model MLC / Units (copy) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Training | | |
| | | | | | | | | | bert | dlrm_dcnv2 | llama2_70b_l.. |
| Public ID | Availability | Organization | System Name | Total Accelerators | Accelerator Model Name | Accelerators Per Node | Host Processor Model Name | Host Processors .. | Latency (In minutes) | Latency (In minutes) | Latency (In minutes) |
| 4.0-0040 | Available on-premise | JuniperNetworks | A100_n8 | 64 | NVIDIA A100-SXM4-80GB | 8 | AMD EPYC 7763 64-Core .. | 2 | | | 75.469 |
| 4.0-0039 | Available on-premise | JuniperNetworks | A100_n8 | 64 | NVIDIA A100-SXM4-80GB | 8 | AMD EPYC 7763 64-Core .. | 2 | 2.641 | | |
| 4.0-0038 | Available on-premise | JuniperNetworks | A100_n8 | 64 | NVIDIA A100-SXM4-80GB | 8 | AMD EPYC 7763 64-Core .. | 2 | | 2.859 | |
| 4.0-0044 | Available on-premise | JuniperNetworks | H100_n4 | 32 | NVIDIA H100-SXM5-80GB | 8 | AMD EPYC 7763 64-Core .. | 2 | | | 44.790 |
| 4.0-0043 | Available on-premise | JuniperNetworks | H100_n4 | 32 | NVIDIA H100-SXM5-80GB | 8 | AMD EPYC 7763 64-Core .. | 2 | 1.884 | | |
| 4.0-0042 | Available on-premise | JuniperNetworks | H100_n4 | 32 | NVIDIA H100-SXM5-80GB | 8 | AMD EPYC 7763 64-Core .. | 2 | | 2.879 | |
| 4.0-0041 | Available on-premise | JuniperNetworks | H100_n1 | 8 | NVIDIA H100-SXM5-80GB | 8 | AMD EPYC 7763 64-Core .. | 2 | | | 116.087 |

internal

# MLCommons - MLPerf Round: v4.0

*https://mlcommons.org/benchmarks/training/*

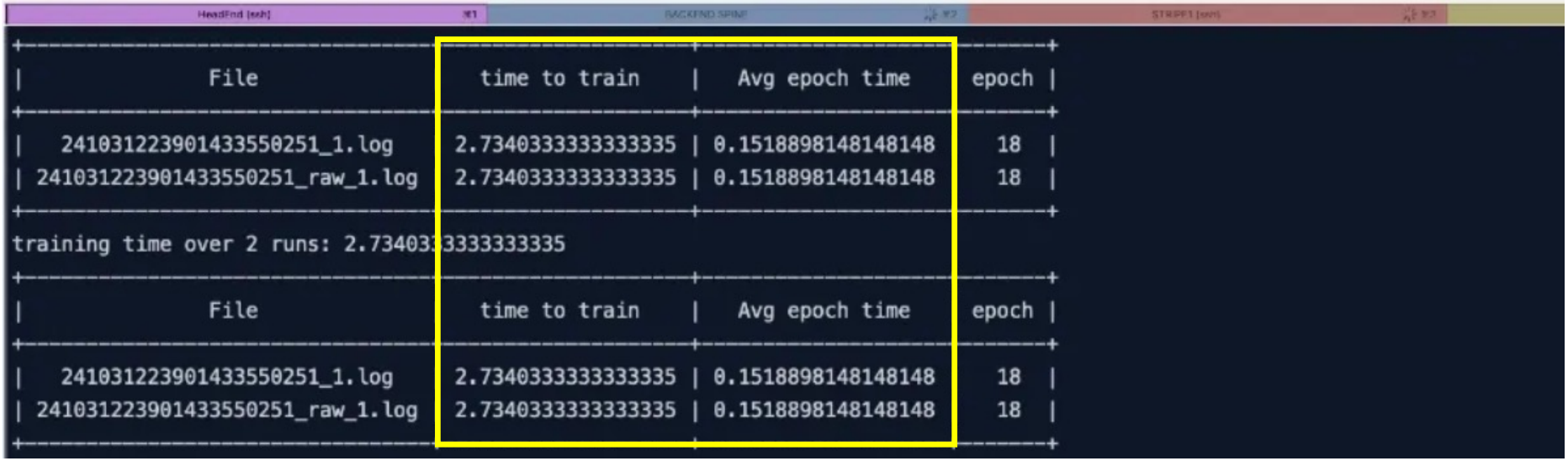
| Model | A100 (64 GPU) | | H100 (32 GPU) | | Comment |
|---|---|---|---|---|---|
| | Juniper | IB | Juniper | IB | |
| BERT$_{LARGE}$ | ~ 2.6 min | ~ 2.5-3.3 min | ~ 1.05 min | NA | |
| DLRM_v2 | ~ 2.8 min | NA | ~ 3.07 min | NA | |
| GPT-3 | ~ 33 hrs | NA | NA | NA | |
| | | | | | |
| IB | NVIDIA InfiniBand | | | | |
| NA | Does not exist or not validated | | | | |

internal

| Training Model | Workload Type | Model Data Rate | DCQCN (ECN/PFC) | DLB | Congestion (IXIA) | JCT (time to train) | Avr epoch time |
|---|---|---|---|---|---|---|---|
| DLRM | RoCEv2 | 30~45G | - | O | - | 2.734 min | 0.151 min |

## Result

- MLPerf Result



```
+-------------------------------------+--------------------+--------------------+-------+
|                File                 |    time to train   |   Avg epoch time   | epoch |
+-------------------------------------+--------------------+--------------------+-------+
|   241031223901433550251_1.log       | 2.7340333333333335 | 0.1518898148148148 |   18  |
| 241031223901433550251_raw_1.log     | 2.7340333333333335 | 0.1518898148148148 |   18  |
+-------------------------------------+--------------------+--------------------+-------+
training time over 2 runs: 2.7340333333333335
+-------------------------------------+--------------------+--------------------+-------+
|                File                 |    time to train   |   Avg epoch time   | epoch |
+-------------------------------------+--------------------+--------------------+-------+
|   241031223901433550251_1.log       | 2.7340333333333335 | 0.1518898148148148 |   18  |
| 241031223901433550251_raw_1.log     | 2.7340333333333335 | 0.1518898148148148 |   18  |
+-------------------------------------+--------------------+--------------------+-------+
```

# AI 네트워크 KPI = JCT

**Job Completion Time**

2022년 Meta AI/ML Fabric 보고서에 따르면,

33% AI 워크로드 처리시간이 network 을 통해 소비되고,

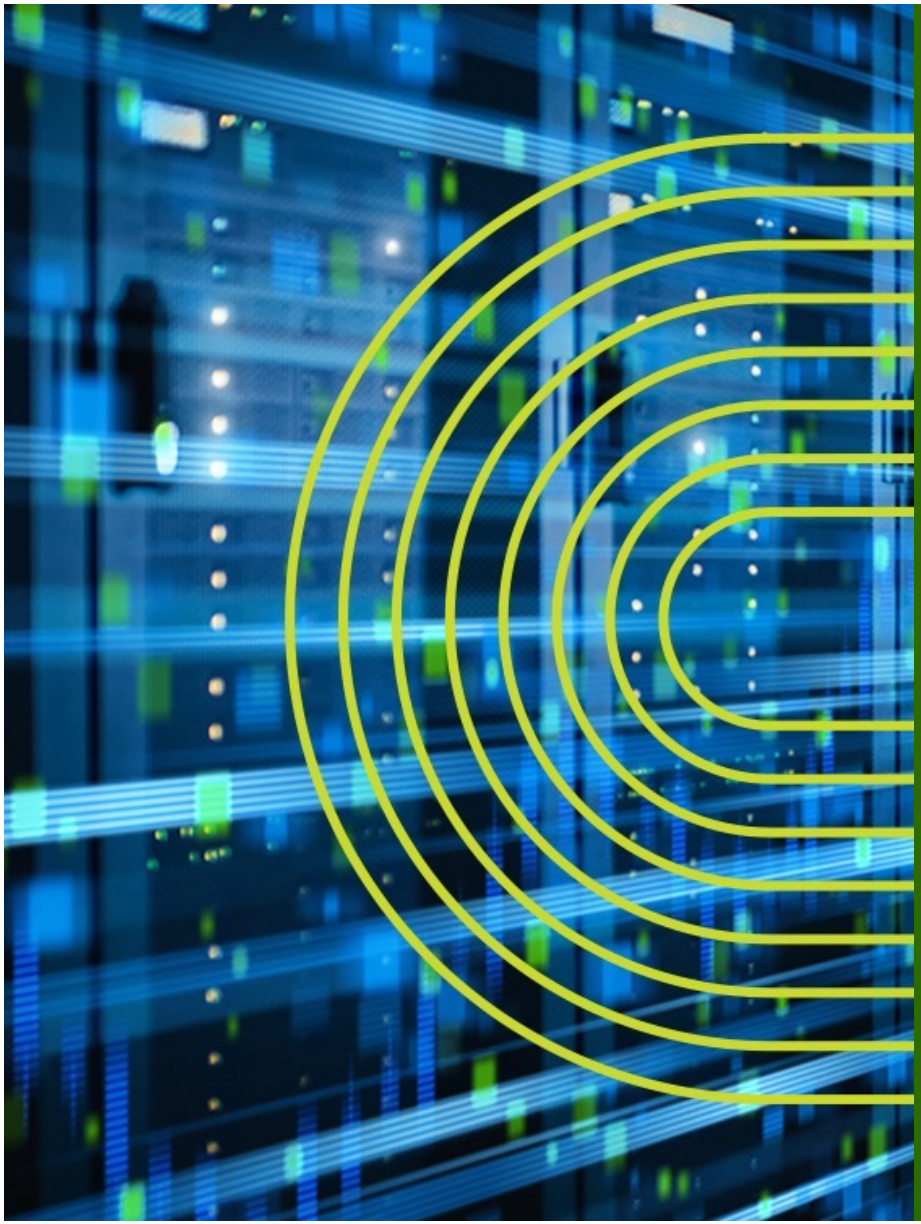
네트워크에 의한 waiting 시간은 Cost 낭비를 의미 (GPU Idle time 이 그 만큼 길어진다는 것임)

그러므로, 네트워크에서 이를 최소화 해 주어야함.

↑ High Performance

↓ Low Latency

0 Loss “0”



# AUTOMATION FOR ROCEV2

Juniper Networks

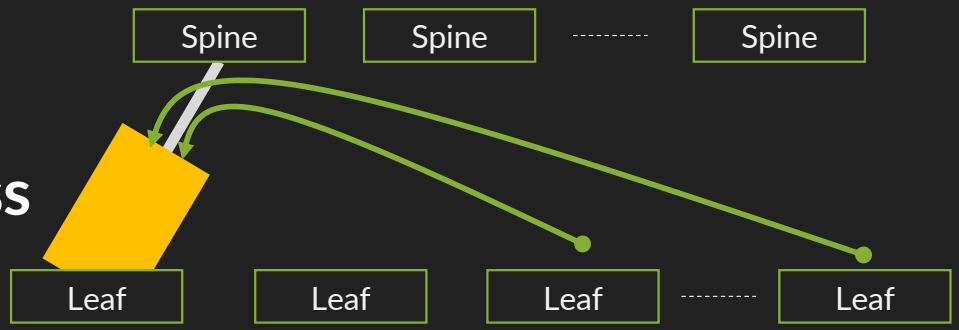
# In-Cast Congestion

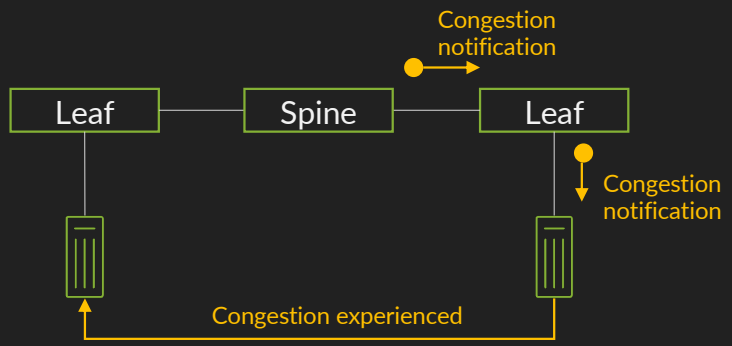
internal

# 0

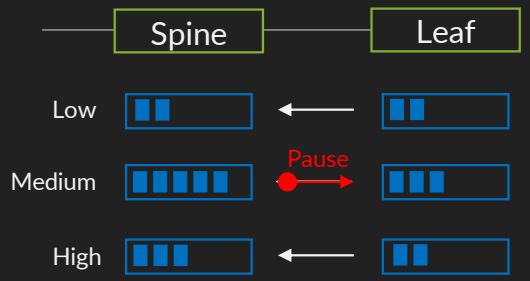
# LossLess Congestion Control

- N to 1
- Congestion = Loss




Explicit Congestion Notification (ECN)

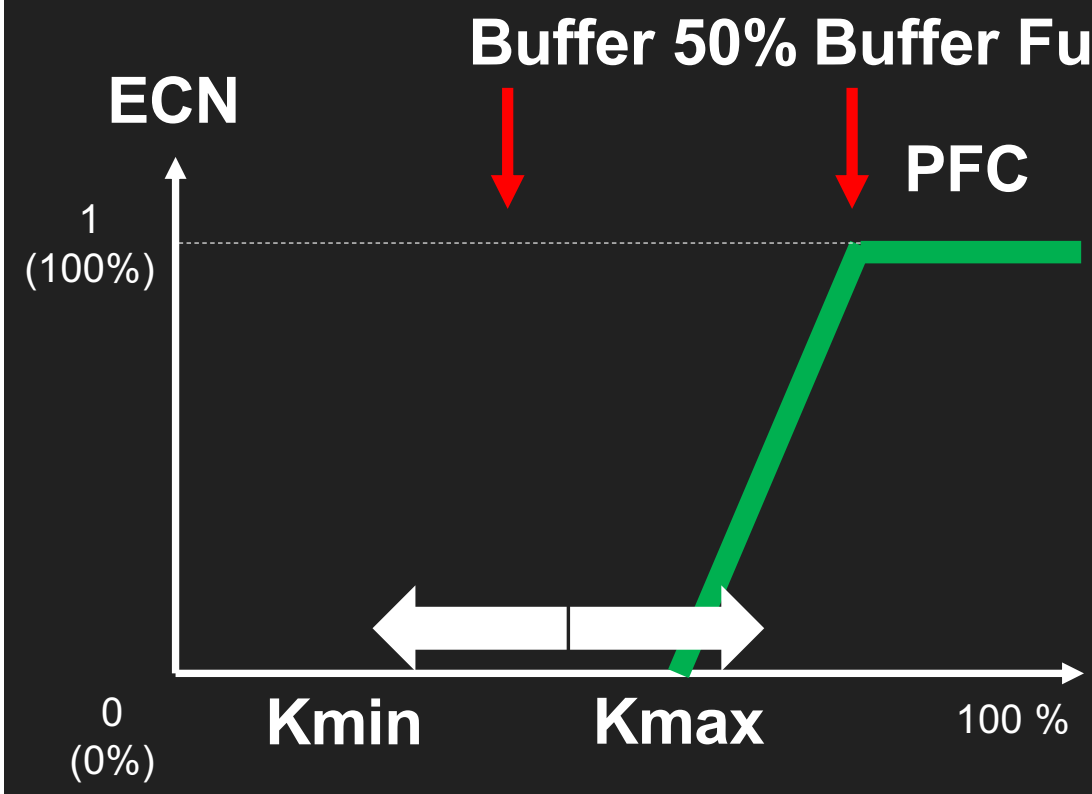


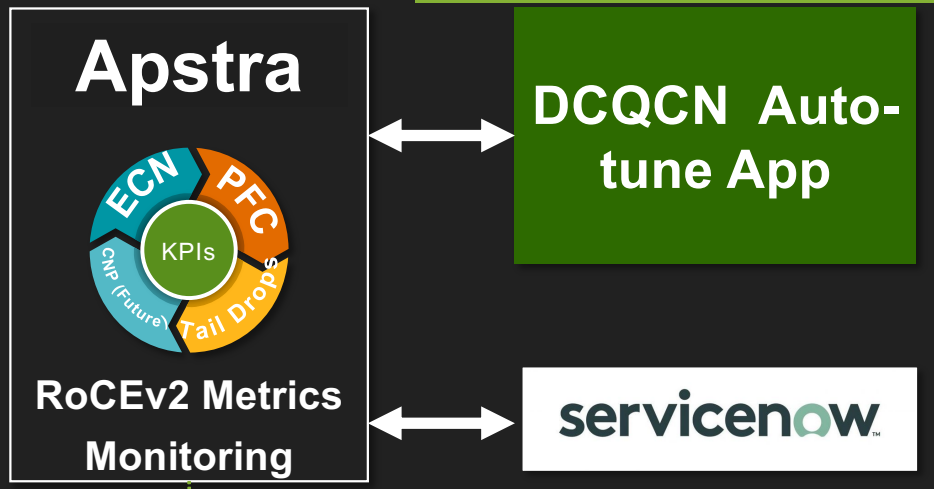

Priority Flow Control (PFC)

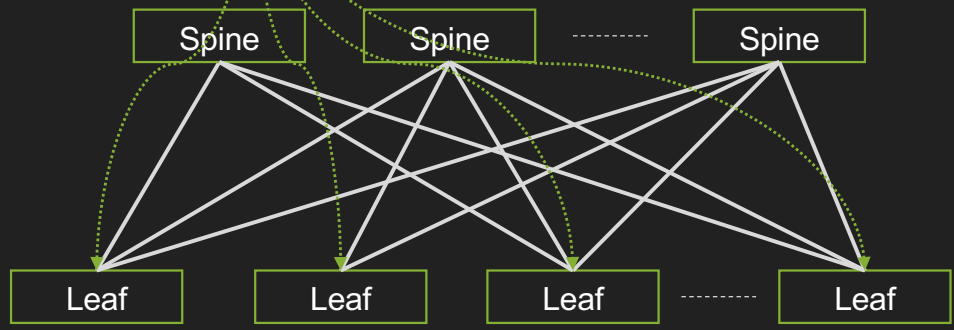
DC Quantized Congestion Notification (DCQCN)

internal

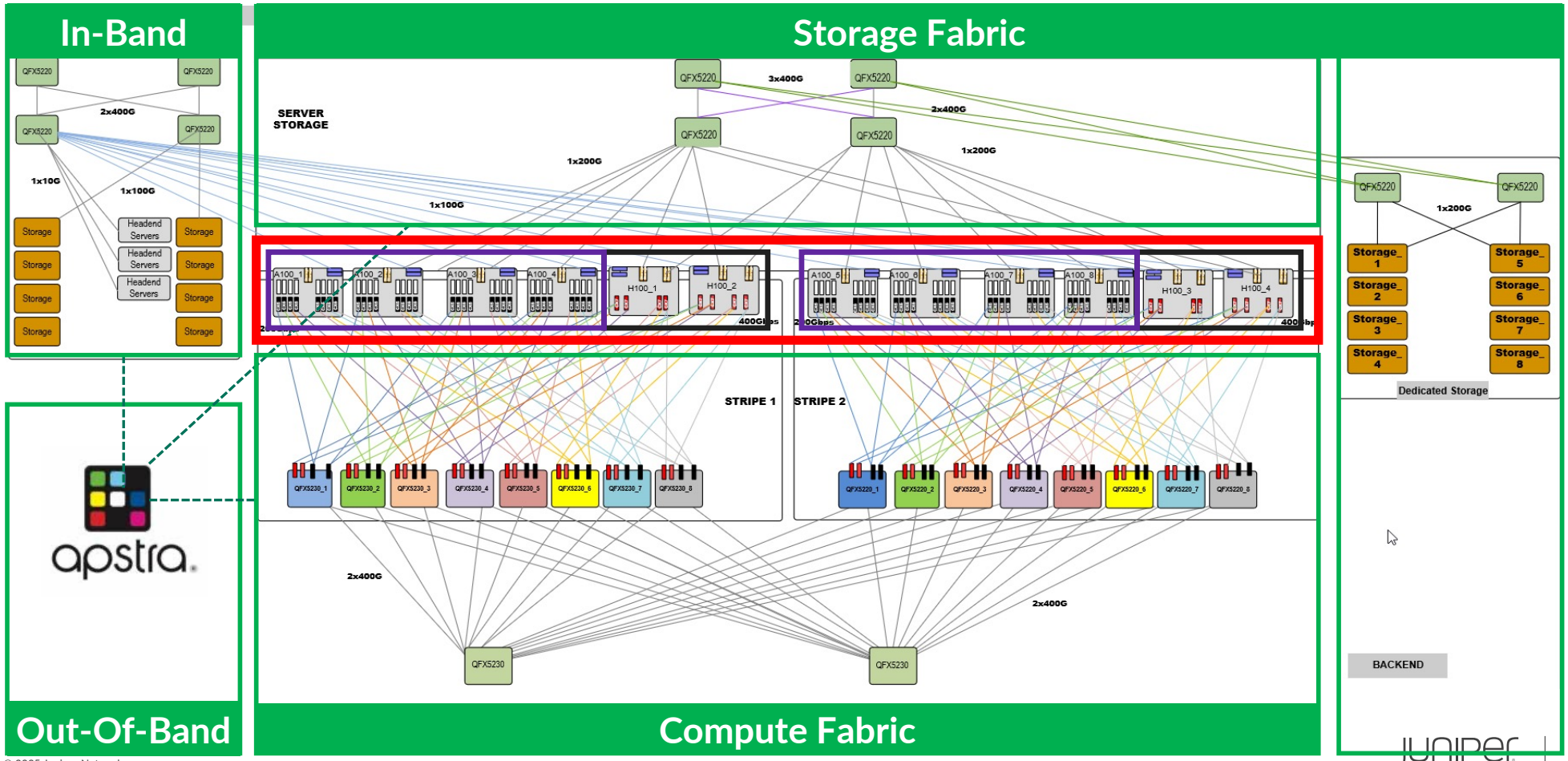
# “JNPR” Auto-Tune DCQCN



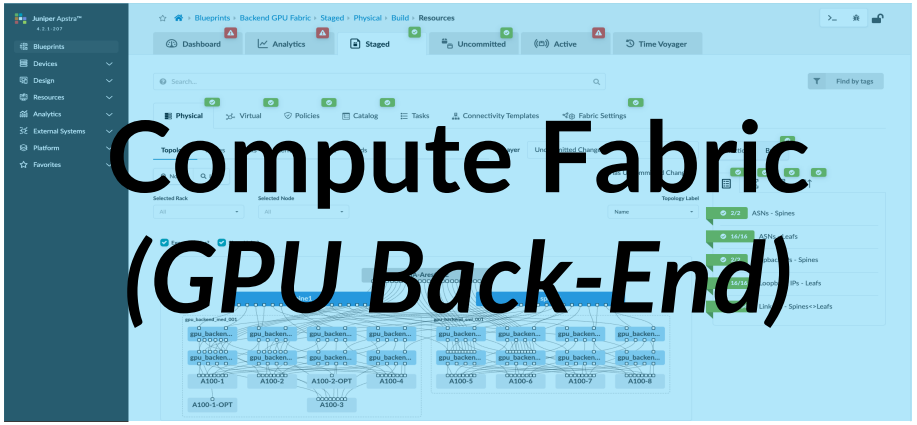



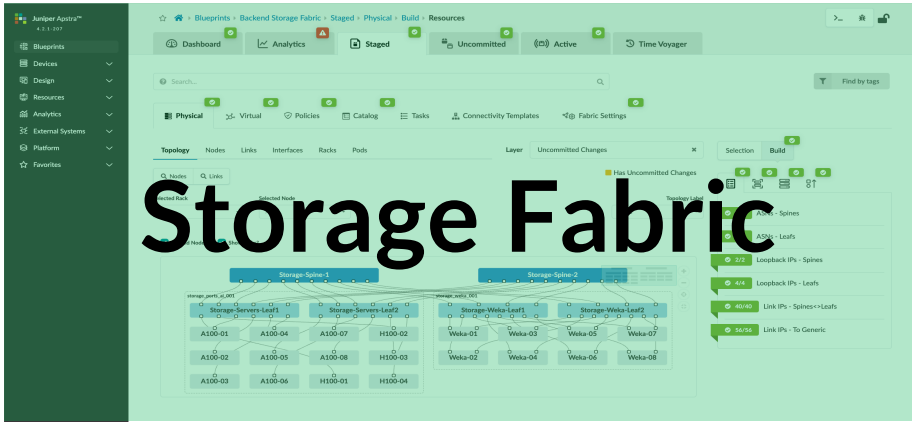
# Nvidia's Reference Architecture



internal

# APSTRA - AI ML Cluster Blueprints

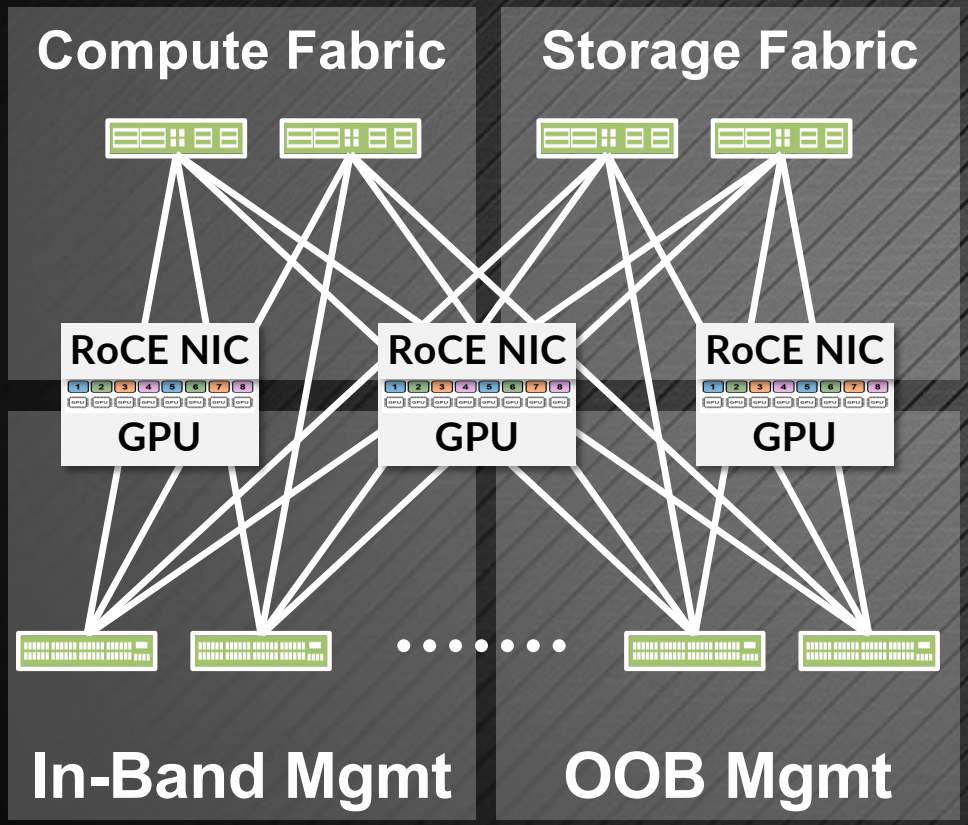




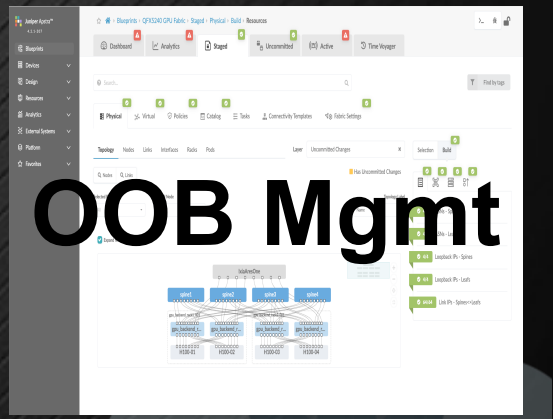
internal

# JNPR Compute Fabric (GPU Back-End) + Others



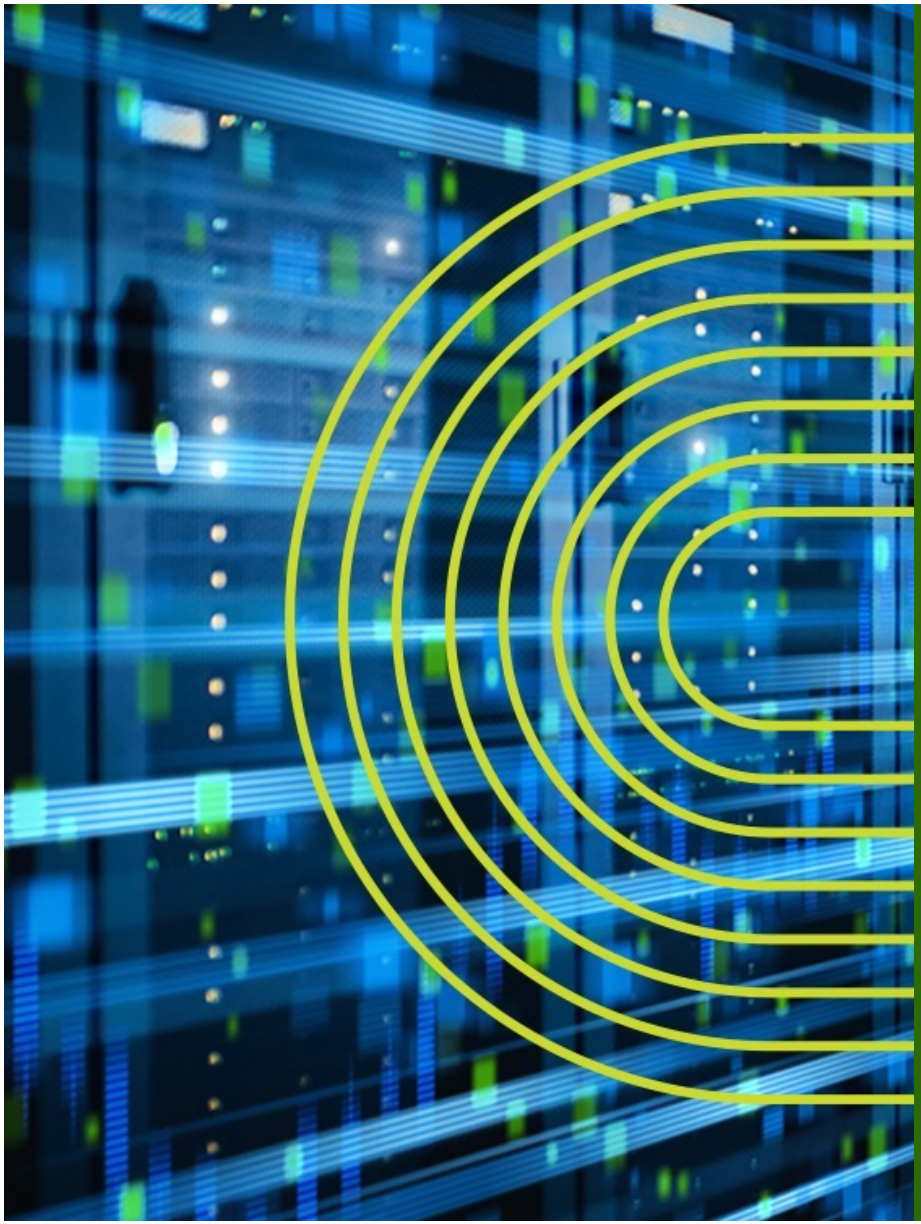









juniper networks



# TEST PLAN

Juniper Networks

# PoC Items

This test plan is designed to test and verify all important aspects of Juniper's AI ML cluster solution using **[QFX5220-32CD/QFX5230-64CD/PTX10K-LC1201]** solution.

Test areas include:

- **[Load Balancing] : Static vs Dynamic LB**
- **[MLPerf Benchmark Operations] : DLRM, BERT, Customer Model**
- **[Telemetry] : Fabric & GPU Nodes**
- **[DCQCN] : Auto-Tune DCQCN**

# PoC Conditions

| Training Model | Workload Type | Model Data Rate | DCQCN (ECN/PFC) | DLB | Congestion (IXIA) |
|---|---|---|---|---|---|
| DLRM | RoCEv2 | 30~45G | - | - | - |
| DLRM | RoCEv2 | 30~45G | - | O | - |
| DLRM | RoCEv2 | 30~45G | - | O | O |
| DLRM | RoCEv2 | 30~45G | O | O | O |

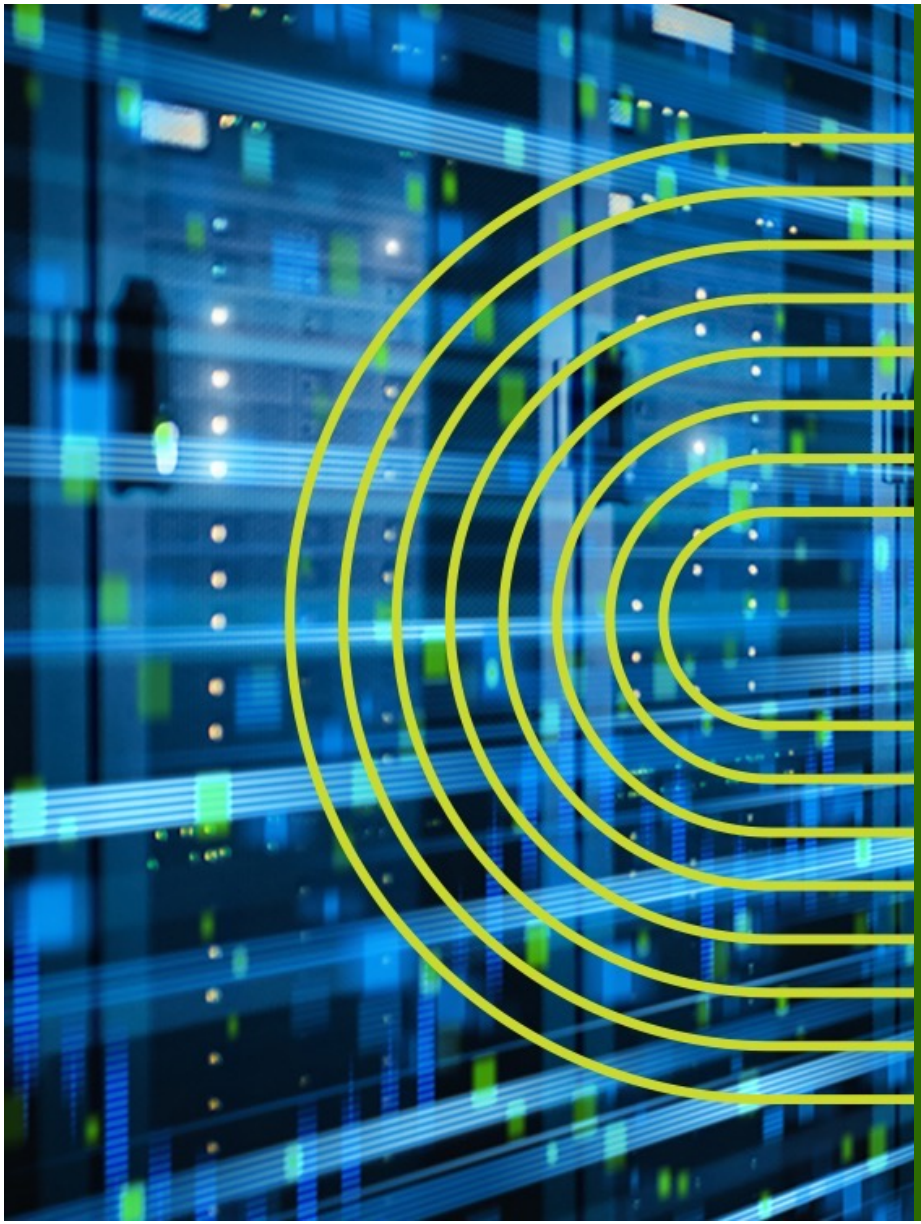
# AI Data Center Network with Juniper Apstra, NVIDIA GPUs, and WEKA Storage—Juniper Validated Design (JVD)

Description:

- Discover Juniper's AI cluster design automated with Apstra & Terraform for NVIDIA GPUs and WEKA Storage. Achieve fast innovation, flexibility, and optimized AI fabric networks with minimized job completion time.

**Resources:**

- **JVD**
- **Solution Overview**
- **Test Report Brief**



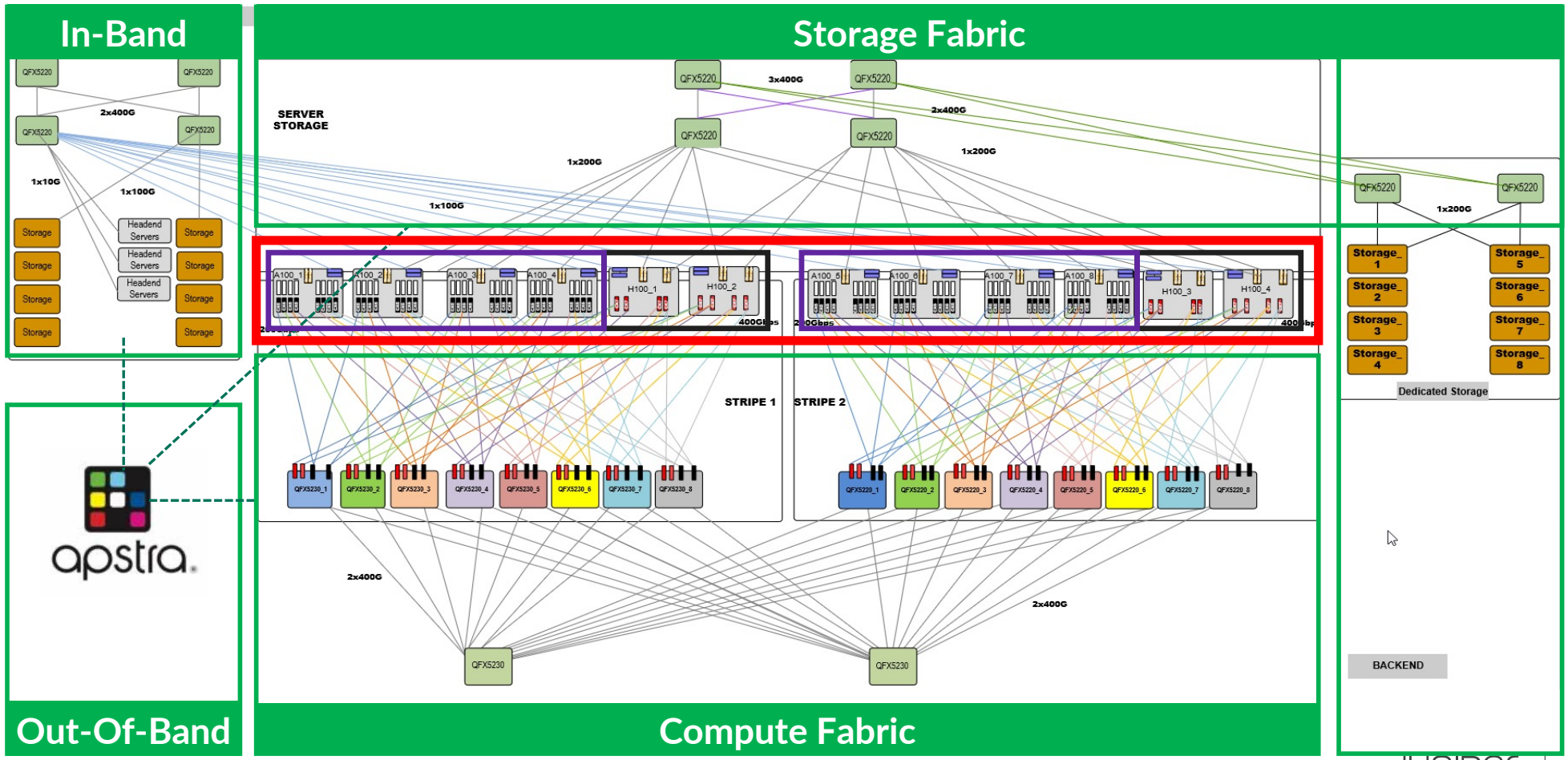
# POC

# TOPOLOGY & DUT

JUNIPER NETWORKS

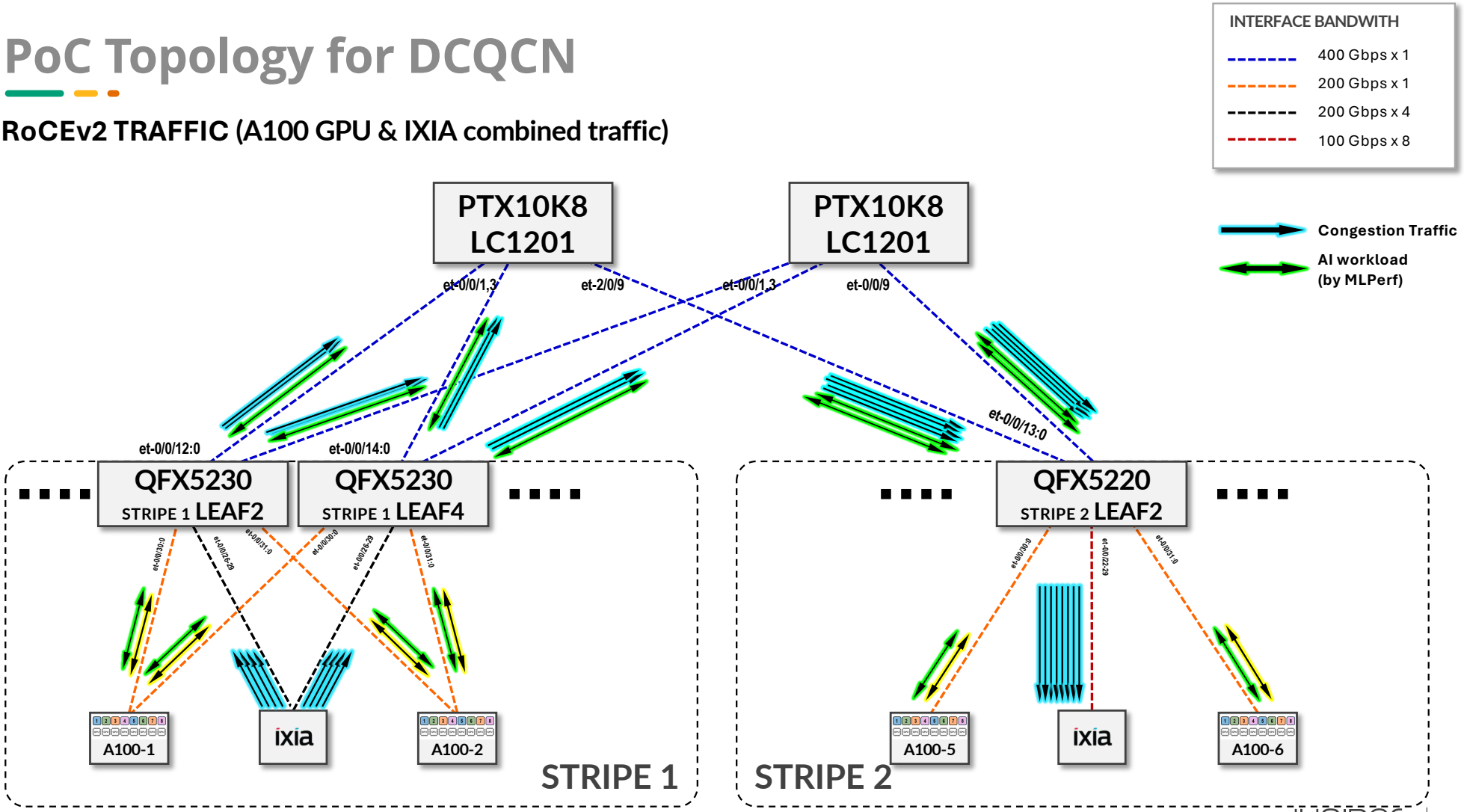
# POC Topology

# PoC Topology



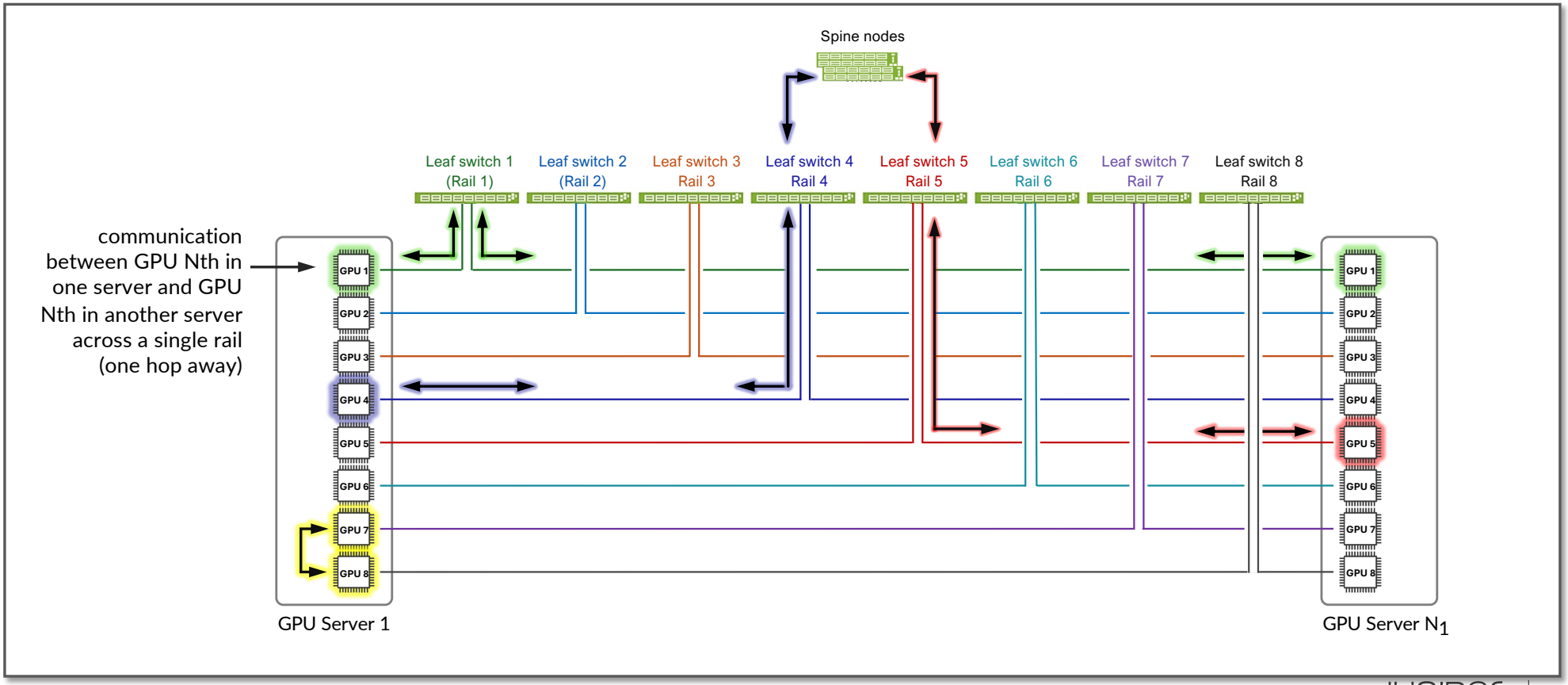
# PoC Topology for DCQCN

**RoCEv2 TRAFFIC (A100 GPU & IXIA combined traffic)**



# PoC Topology for Rail-Optimized

## Rail-Optimized Design



# POC DUTs

* DUT: Device Under Test

# PoC DUTs (1/4)

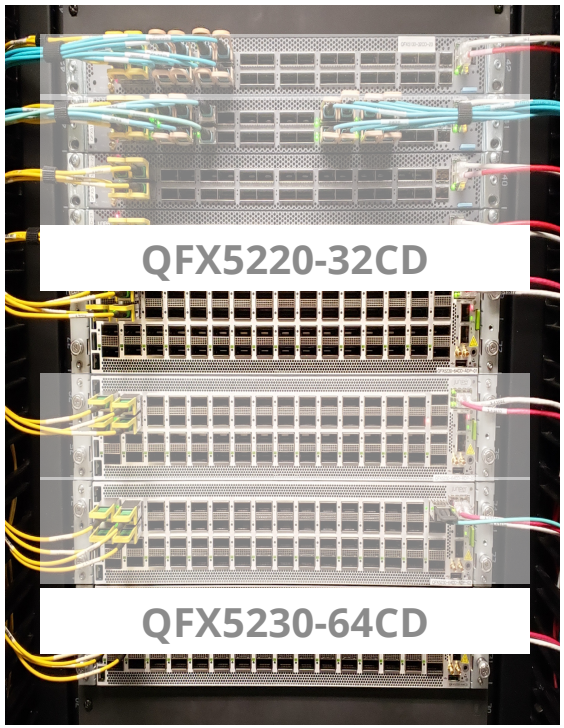
## Sunnyvale PoC Lab

# PoC DUTs (2/4) – Leaf/Spine switch

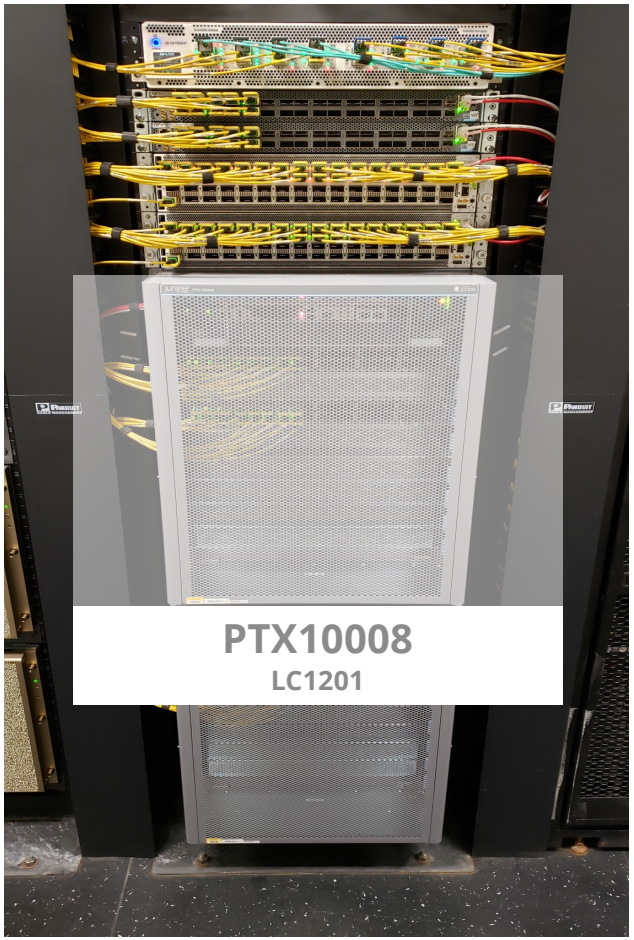
**PTX10K8 as Spine, QFX5220-32CD/QFX5230-64CD as Leaves, IXIA**

## IXIA
traffic generator


## Leaf
QFX5220-32CD/QFXS523-64CD

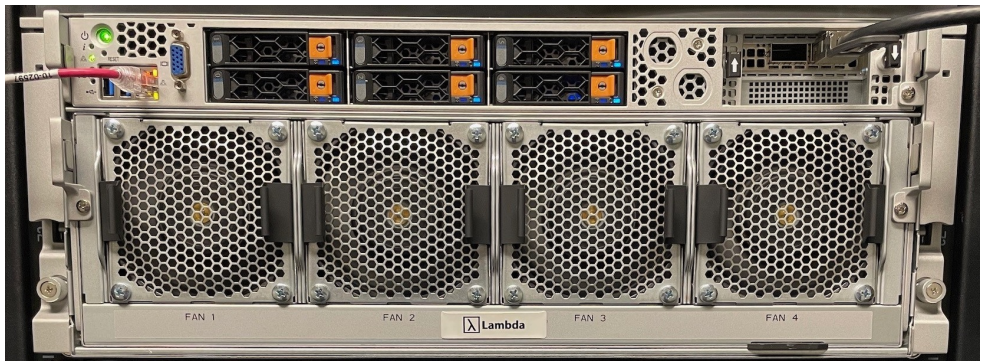



## Spine
PTX10008 w/LC1201



# PoC DUTs (3/4) – GPU Node

## NVIDIA A100 GPU Node




***SPECS:***

**Operating system:** Ubuntu 22.04: Includes Lambda Stack for managing TensorFlow, PyTorch, CUDA, cuDNN, etc.

**Processor:** 2x AMD EPYC 7763: 64 cores, 2.45~3.5GHz, 256MB cache, PCIe 4.0

**GPUs: 8x NVIDIA A100 (80GB) SXM4:**

**HGX platform with NVLink and NVSwitch fabric**

**System memory:** 2048 GB: DDR4-3200 ECC RDIMM

**OS drives:** 2x 1.92 TB M.2 NVMe: Data center SSD, 1 DWPD, PCIe 4.0

**Data drives:** 6x 3.84 TB U.2 NVMe: Data center SSD, 1 DWPD, PCIe 4.0

Standard networking: 1x NVIDIA ConnectX-6 Dx adapter card, 100GbE, dual-port QSFP28, AIOM PCIe 4.0 x16

**Storage Networking: 1x 200 Gbps NVIDIA ConnectX-6 VPI NIC**: Dual-port QSFP56, HDR InfiniBand/Ethernet

**GPUDirect RDMA Networking: 8x NVIDIA ConnectX-7**

**Adapter Card 200Gb/s Single-port QSFP PCIe 4.0 x16**

**Warranty & support:** Hardware & System Support - 3 Year Standard: Three-year hardware warranty with Lambda technical support covering hardware and software issues including Lambda Stack, ML frameworks, drivers, OS and BIOS

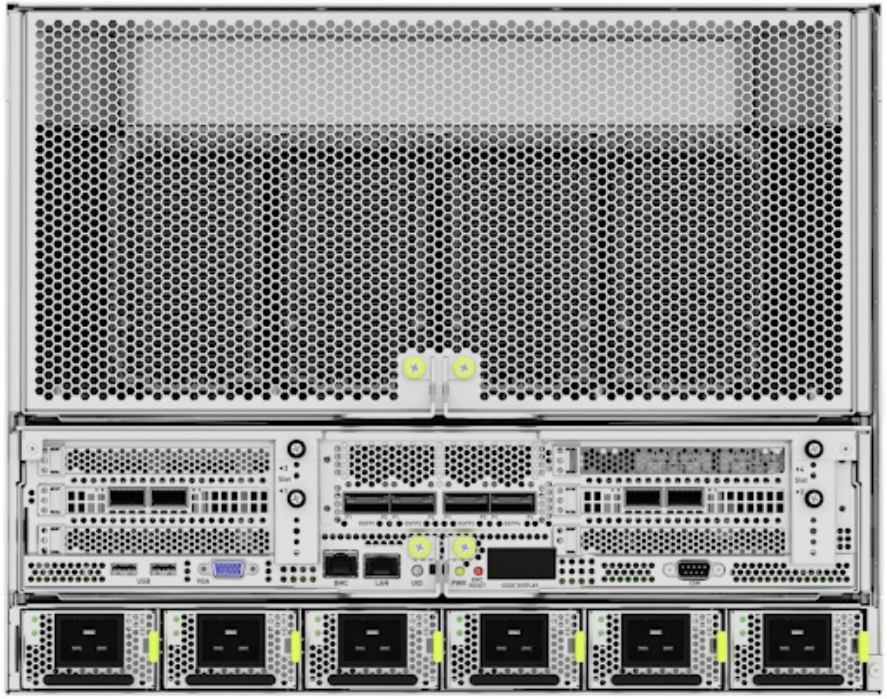
**Power:** 2+2 redundant 3000W PSUs, 4x C20 inlets, 200-240Vac input

**Physical:** 4U rackmount, 6.9 x 17.6 x 35.4 in (174 x 446 x 900 mm, HxWxD), 166 lb (75.3 kg) system, 225 lb (102.1 kg) shipping

internal

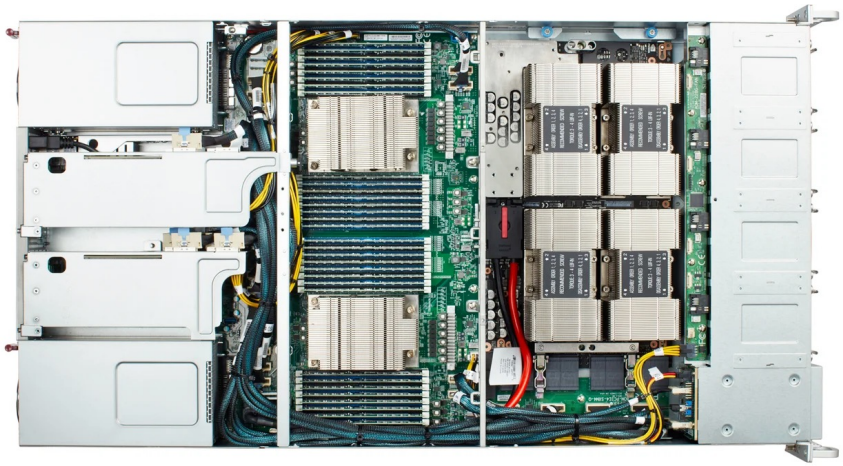
# PoC DUTs (3/4) – GPU Node

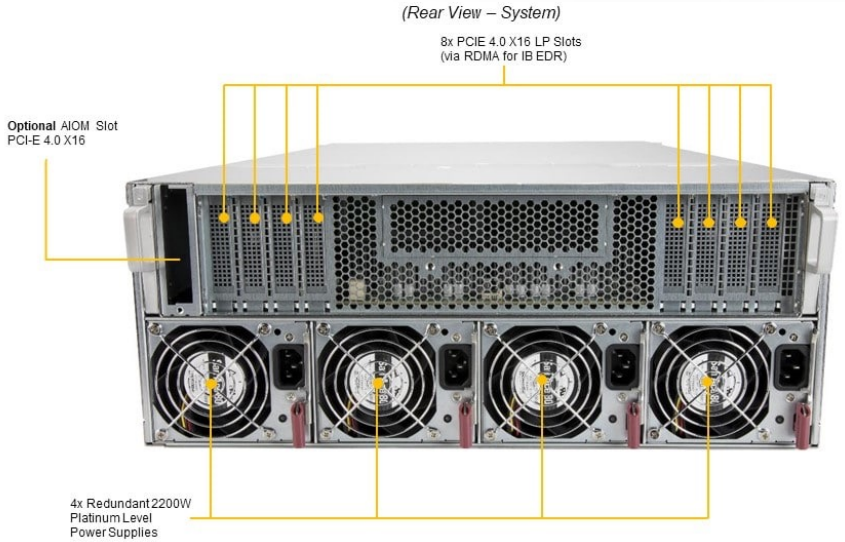
## Four of the NVIDIA DGX H100



# PoC DUTs (3/4) – GPU Node

## NVIDIA A100 GPU Node



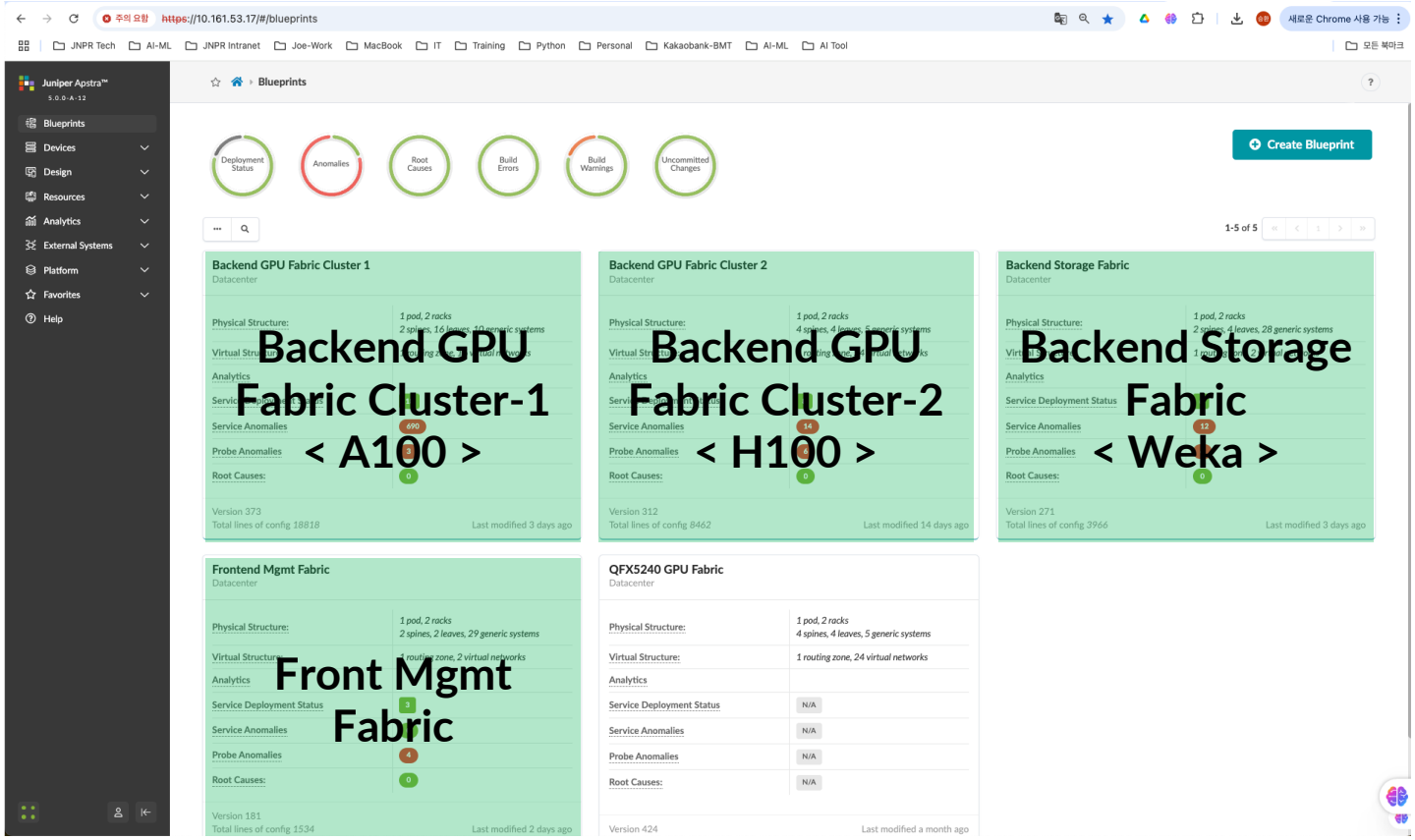
# PoC DUTs (3/4) – GPU Node

**NVIDIA-SMIU, CUDA, Driver version**

```
jnpr@A100-01:~/scripts$ nvidia-smi
+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 535.161.08             Driver Version: 535.161.08   CUDA Version: 12.2     |
|-----------------------------------------+----------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|=========================================+======================+======================|
|   0  NVIDIA A100-SXM4-80GB           On | 00000000:07:00.0 Off |                    0 |
| N/A   35C    P0              73W / 400W |   1916MiB / 81920MiB |      0%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+
|   1  NVIDIA A100-SXM4-80GB           On | 00000000:0A:00.0 Off |                    0 |
| N/A   32C    P0              73W / 400W |   2442MiB / 81920MiB |      0%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+
~~~ omitted
+-----------------------------------------+----------------------+----------------------+
|   7  NVIDIA A100-SXM4-80GB           On | 00000000:C4:00.0 Off |                    0 |
| N/A   34C    P0              71W / 400W |    500MiB / 81920MiB |      0%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+
```

# PoC DUTs (4/4) – Apstra mgmt solution

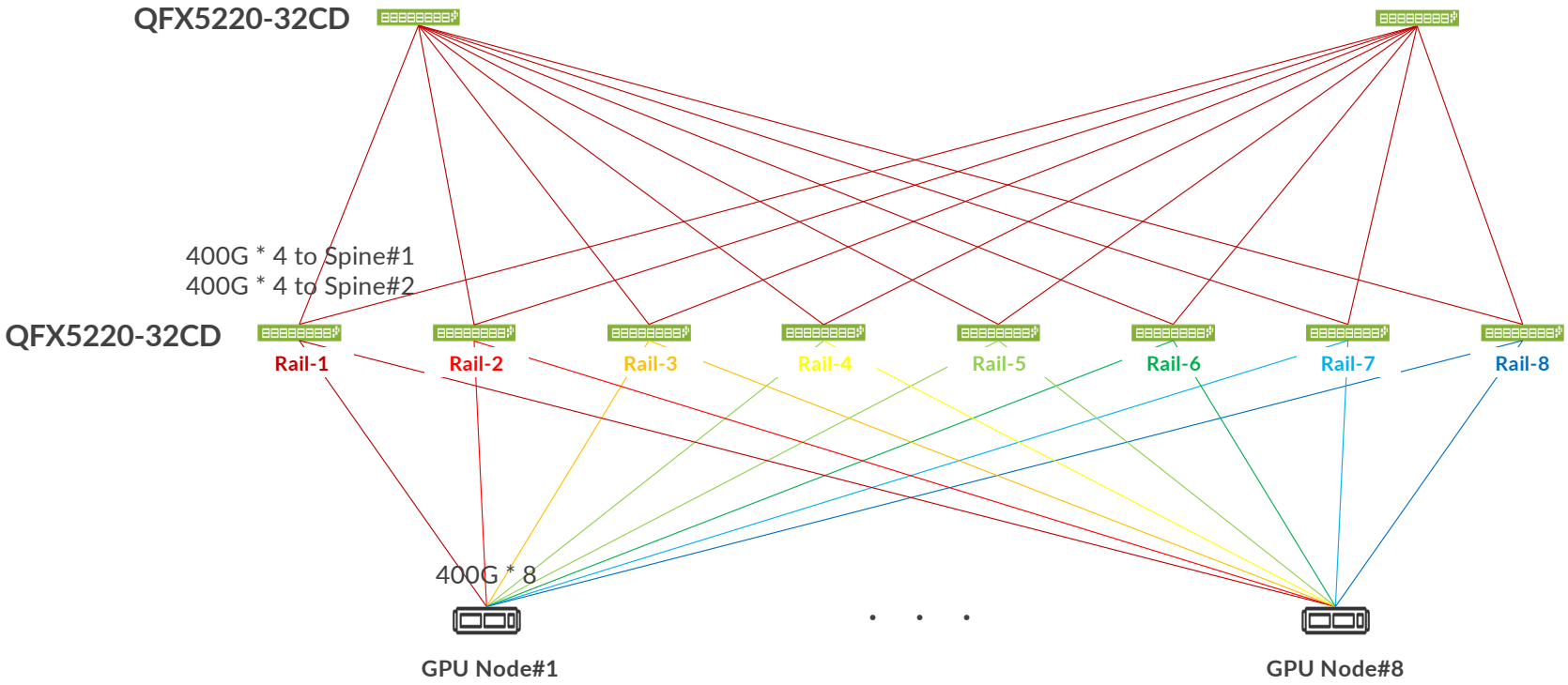
## Apstra for Auto-mangement



# Expected Diagram(Phase 1 – Trial Biz)

- Spine PTX10008(CL1201) * 2EA
- Leaf QFX5220-32CD * 8EA
- DC Automation : Apstra

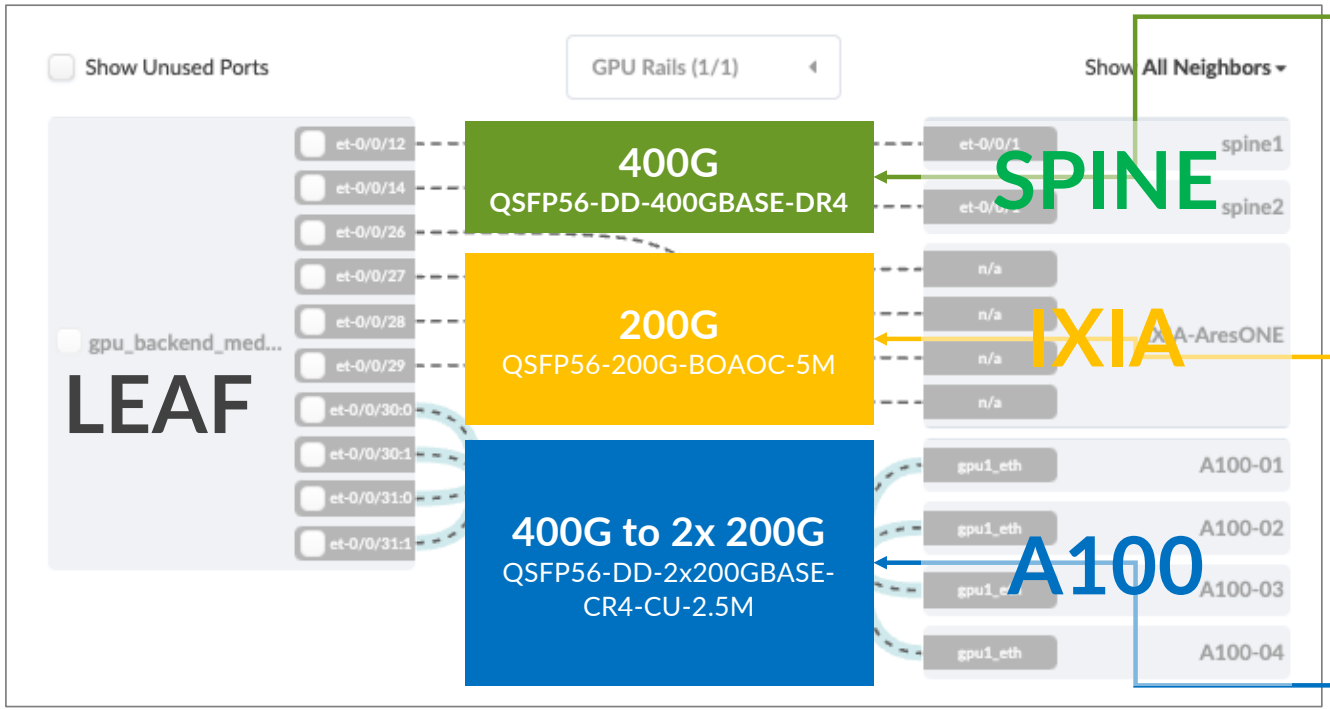
## Phase1(Trial) (GPU Server 8 nodes)

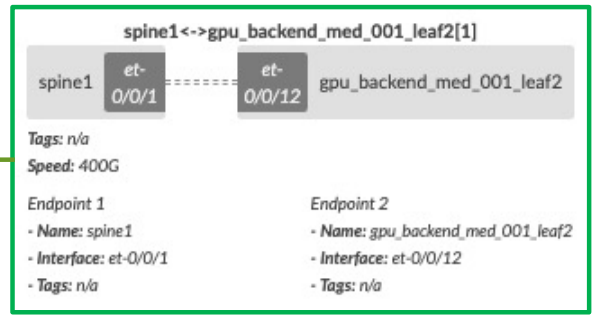


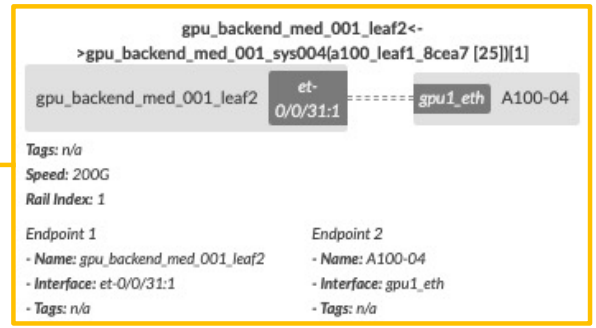
# Transceiver(Optic)

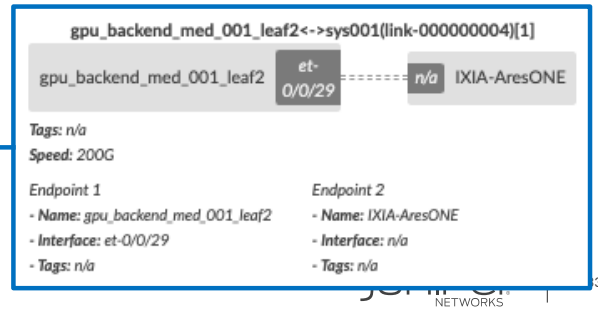
# PoC Transceiver Info.

**200G / 400G(B/O) to GPU & IXIA**



# reference> 400G optic Interop

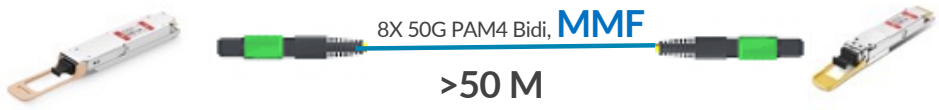
## 400G Optics

| Vendor | Optic Pair | Product Number (SKU) | Form Factor | Description | Electrical Lane | Optical Line | Electrical Modulation | Optical Modulation | Fiber Type | Connector | Length | Fiber Pair | Optical wave per Fiber | Wavelength |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| JNPR | 호환 | QDD-400G-VR4 (25년 2월 출시) | QSFP56-DD | QSFP56-DD-400GBADSE-VR4 | 8x50G (QSFP56) | 4x100G | 50G PAM-4 | 100G PAM-4 | MMF | MPO-12 APC | 50m | 4 | 1 | 850nm |
| NVIDIA | | NVIDIA MMA4Z00-NS400 | OSFP | MMA4Z00-NS400 400Gb/s SR4 | 4x100G (QSFP112) | 4x100G | 100G PAM-4 | 100G PAM-4 | MMF | MPO-12 APC | 50m | 4 | 1 | 850nm |
| JNPR | 호환 | QDD-400G-DR4 | QSFP56-DD | QSFP56-DD-400GBASE-DR4 | 8x50G (QSFP56) | 4x100G | 50G PAM-4 | 100G PAM-4 | SMF | MPO-12 APC | 500m | 4 | 1 | 1310nm |
| NVIDIA | | NVIDIA MMS4X00-NS400 | OSFP | NVIDIA MMS4X00-NS400 DR4 | 4x100G (QSFP112) | 4x100G | 100G PAM-4 | 100G PAM-4 | SMF | MPO-12 APC | 100m | 4 | 1 | 1310nm |

1x400G OSFP NIC Card


OSFP-400G-DR4

QDD-400G-DR4

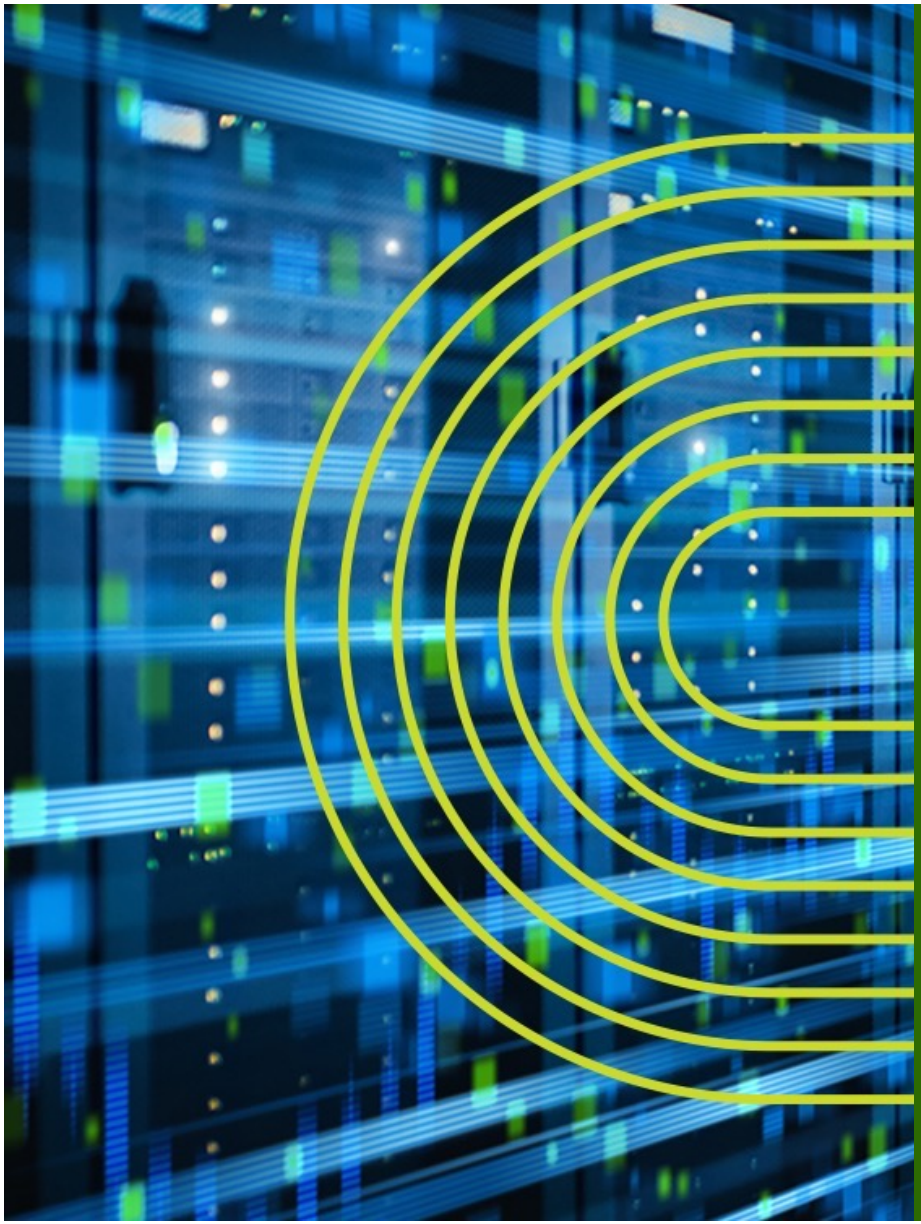



OSFP-400G-SR4

QDD-400G-VR4

QFX5220-32CD

QFX5230-64CD



# JUNIPER AI/ML

# SOLUTION PORTFOLIO

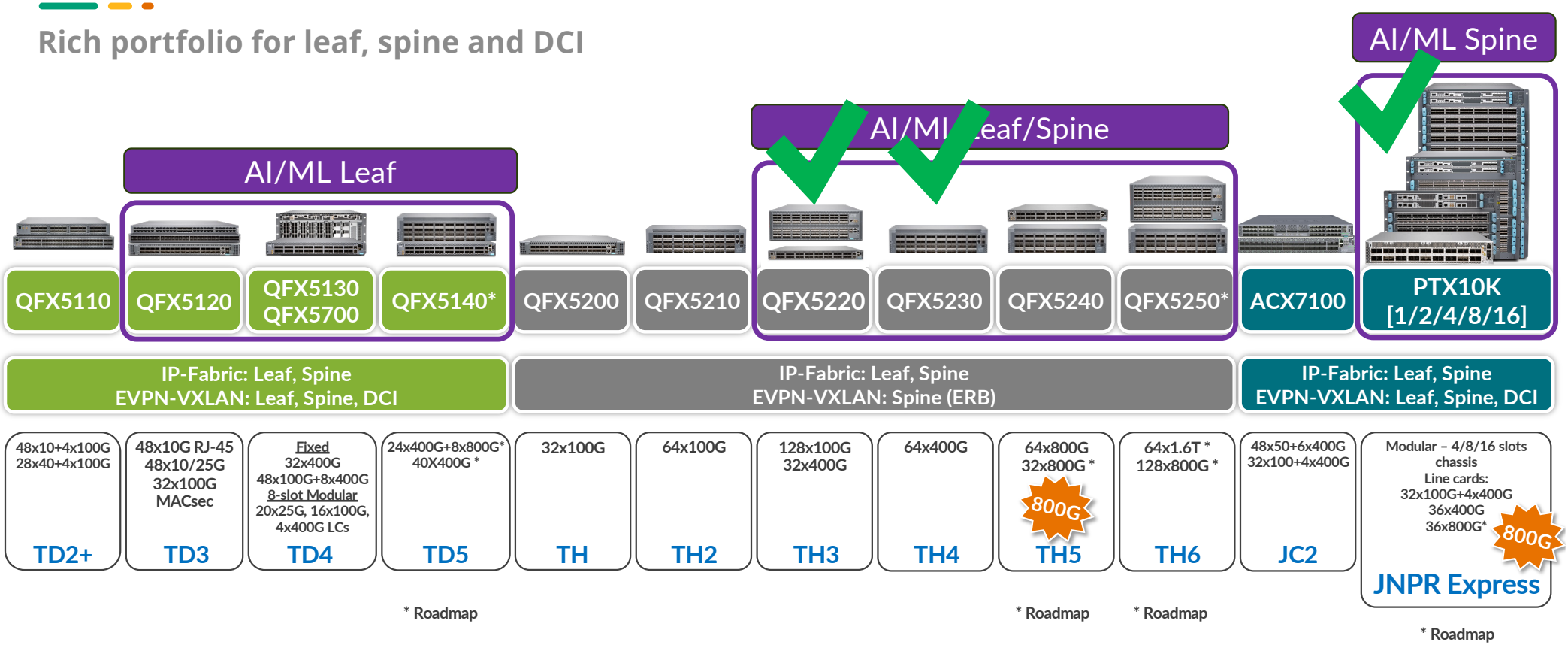
Juniper Networks

# Switch Platform

PoC Focused

# Data center switching portfolio

## Rich portfolio for leaf, spine and DCI



# Leaf - QFX5220-32CD (12.8 Tbps)

## PRODUCT SPECS

- 1 RU chassis
- Broadcom Tomahawk 3 forwarding ASIC
- 32 x 400GE – optional support per port 100GE, 4x100GE, 40GE or 4x25GE speeds
- 12.8 Tbps throughput
- Broadwell-DE Quad Core, 1.6 GHz CPU
- 16 GB (2 x 8GB) DDR4 SDRAM
- 2 x 50 GB SSD
- 1600W PSU (1+1 redundancy)
- Front-Back / Back-Front cooling
- 2 x SFP+/SFP ports for In-band network management
- 1x RS-232 Console port
- 1x USB 2.0 port
- 1x RJ45 for ToD


32 x 40G/100G/400G Ports

**QFX5220-32CD Front View**


**QFX5220-32CD AFO Rear View**

# Leaf - QFX5230-64CD (12.8 Tbps)

## PRODUCT SPECS

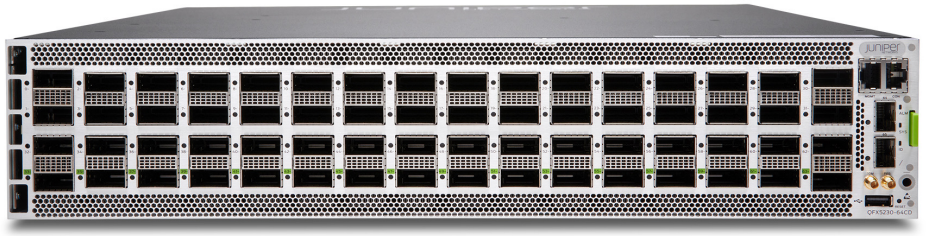
- 2RU
- 25.6T capacity (unidirectional)
- 32GB DDR4 RAM
- 2x100G SSD storage
- 2xPSUs
- Ports-to-FRUs (AFO) and FRUs-to-ports (AFI) cooling
- Redundant (3x2) +(1x2) hot-pluggable fan modules
- System configuration:

64x400G / 128x200G / 256x100G

- EVPN-VXLAN (ESI-LAG):

VxLAN L3 GW

Type5—to—Type5 stitching



QFX5230-64CD Front View

QFX5230-64CD AFO Rear View

# Spine – PTX10008(LC1201)

8 slot chassis

**PTX10008**

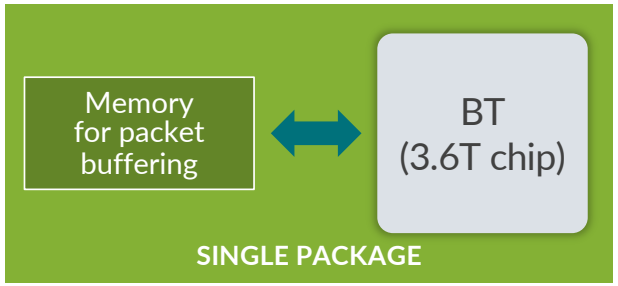
36x400G LC 144x100G
32x100G + 4X400G LC

4 slot chassis

**PTX10004***

36x400G LC 144x100G
32x100G + 4X400G LC

## Juniper ASIC - BT Chips

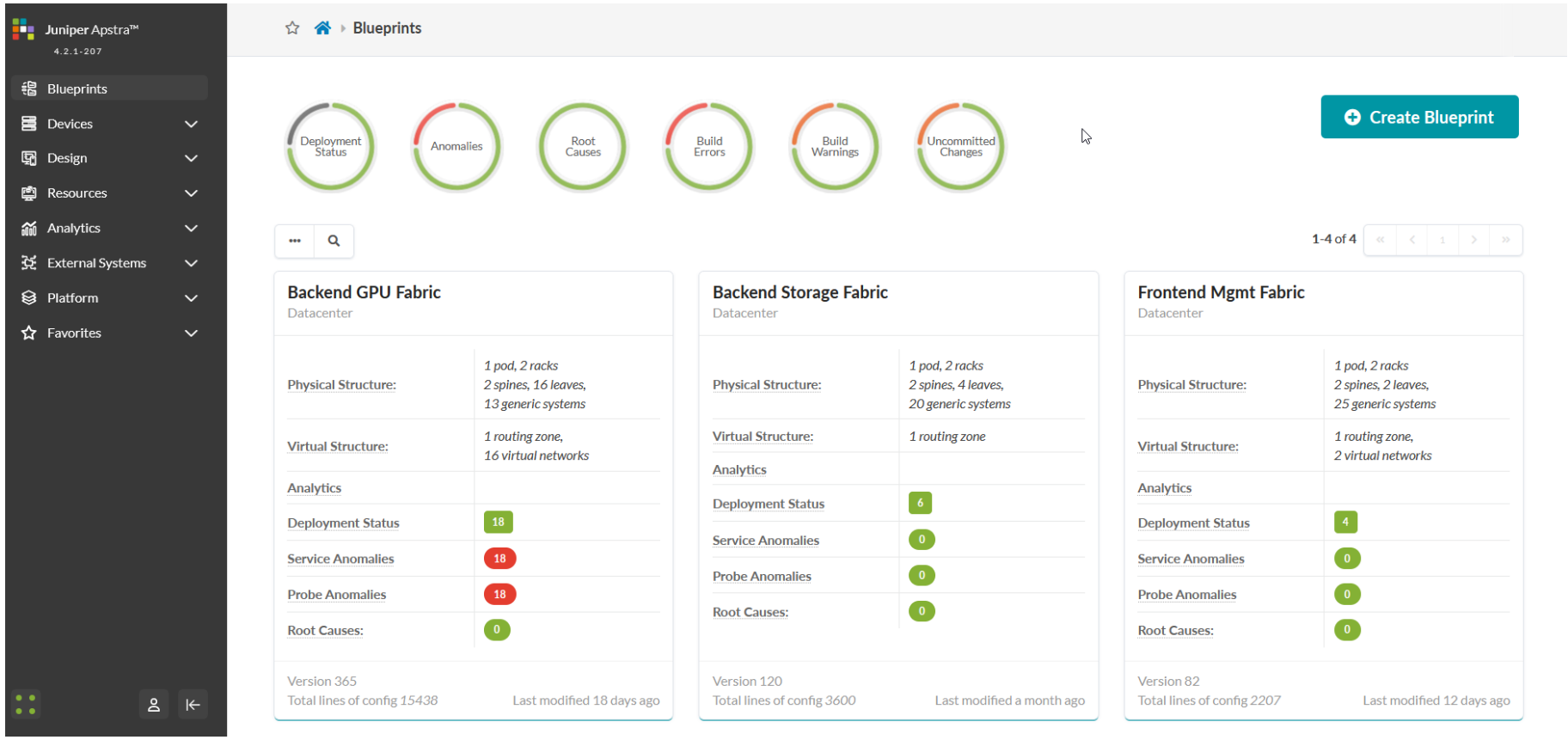



- 25G/50G support
- 400G support
- MACsec capable (400G)
- 2M on-chip FIB
- 14nm technology

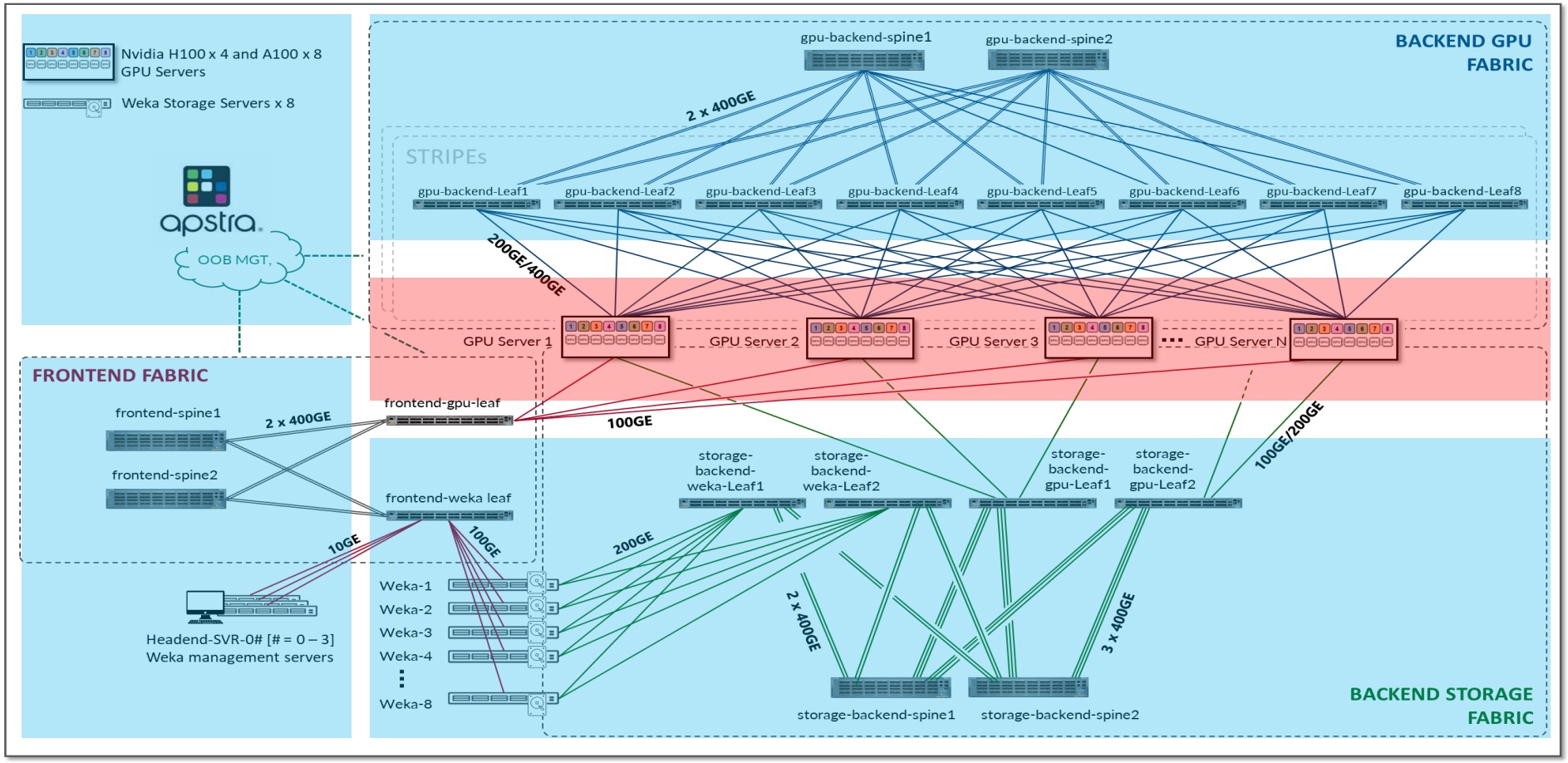

**LC1201**
**36 x 400G (14.4T)**

# Management Software

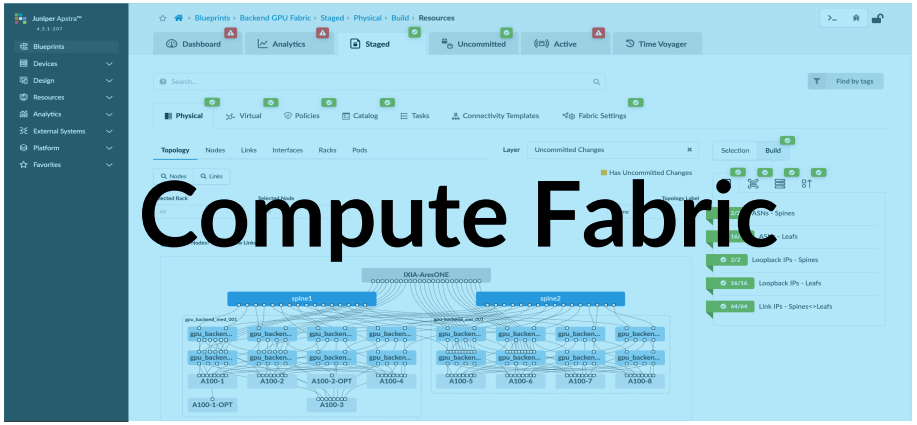
# APSTRA - AI ML Cluster Blueprints

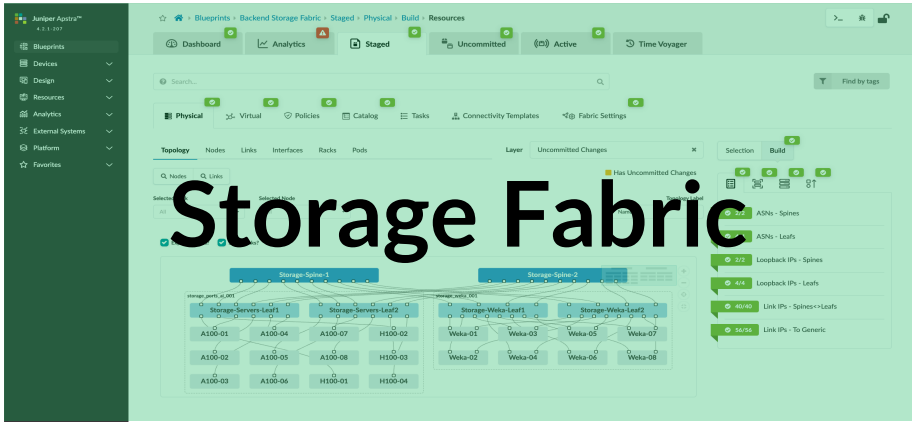


# AI Innovation Lab - Fabrics

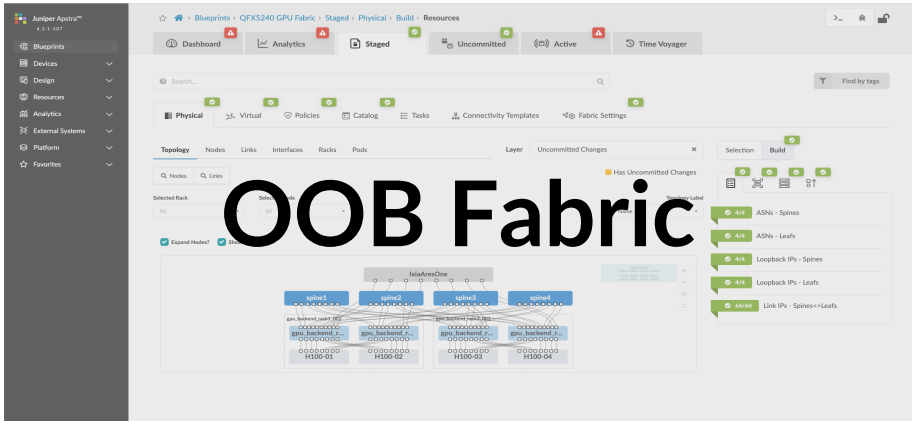


# APSTRA - AI ML Cluster Blueprints







# THANK YOU

JUNIPER NETWORKS | Driven by Experience™