



400G Ethernet PoC for AI Cluster



ROCEV2 VS INFINIBAND

AI 성능측정

네트워크 세상에는 BMT: rfc2544 by 계측기

- AI 세상에는?

MLperf

MLCommons - MLPerf Round: v4.0

<https://mlcommons.org/benchmarks/training/>

internal

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

Public ID	Availability	Organization	System Name	Total Accelerators	Accelerator Model Name	Accelerators Per Node	Host Processor Model Name	Host Processors	Benchmark / Model MLC / Units (copy)		
									Training		
									bert	dlrm_dcnv2	llama2_70b_l1.
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

MLCommons - **MLPerf** Round: v4.0

<https://mlcommons.org/benchmarks/training/>

internal

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

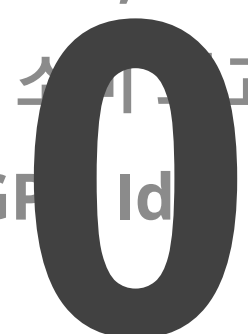
- MLPerf Result

```
HeadEnd [snp]  0.01  (BACKEND SPINE)  0.02  (STRIP1 [snp])  0.02
+-----+-----+-----+-----+
|           File           | time to train | Avg epoch time | epoch |
+-----+-----+-----+-----+
| 241031223901433550251_1.log | 2.734033333333335 | 0.1518898148148148 | 18 |
| 241031223901433550251_raw_1.log | 2.734033333333335 | 0.1518898148148148 | 18 |
+-----+-----+-----+-----+
training time over 2 runs: 2.734033333333335
+-----+-----+-----+-----+
|           File           | time to train | Avg epoch time | epoch |
+-----+-----+-----+-----+
| 241031223901433550251_1.log | 2.734033333333335 | 0.1518898148148148 | 18 |
| 241031223901433550251_raw_1.log | 2.734033333333335 | 0.1518898148148148 | 18 |
+-----+-----+-----+-----+
```

AI 네트워크 KPI = JCT

Job Completion Time

2022년 Meta AI/ML Fabric 보고서에 따르면,
33%의 네트워크로드 처리시간이 network 을 통해 소모되고,
네트워크에 의한 waiting 시간은 Cost 낭비를 의미 (GPU Idle time 이
그 만큼 같아 떨어진다는 것임)



그러므로, 네트워크에서 이를 최소화 해 주어야함.

**High
Performance**

**Low
Latency**

**Loss
“0”**

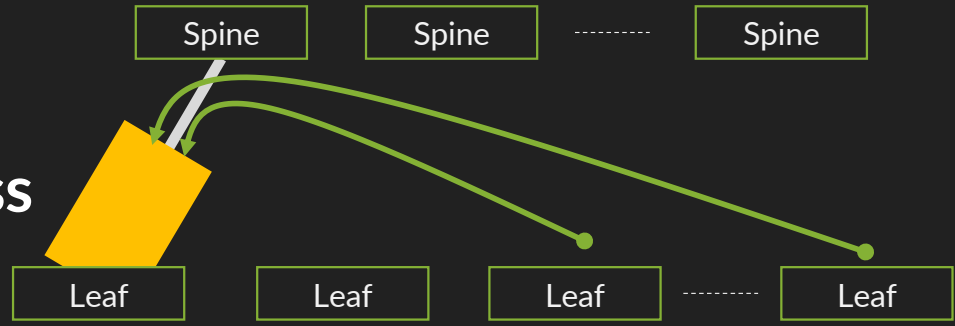


AUTOMATION FOR ROCEV2

In-Cast Congestion

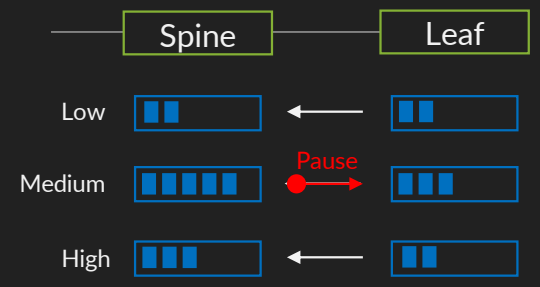
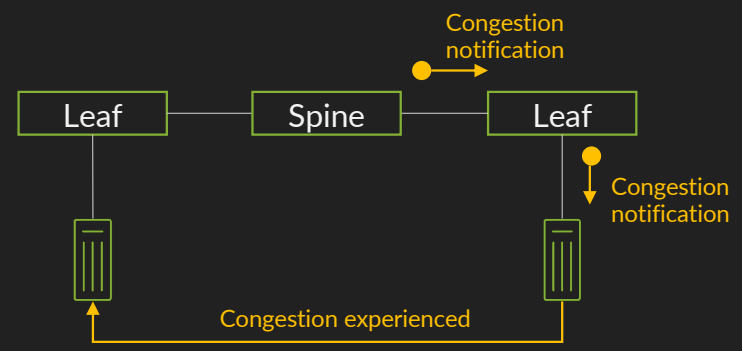
internal

- N to 1
- Congestion = Loss



0

LossLess Congestion Control



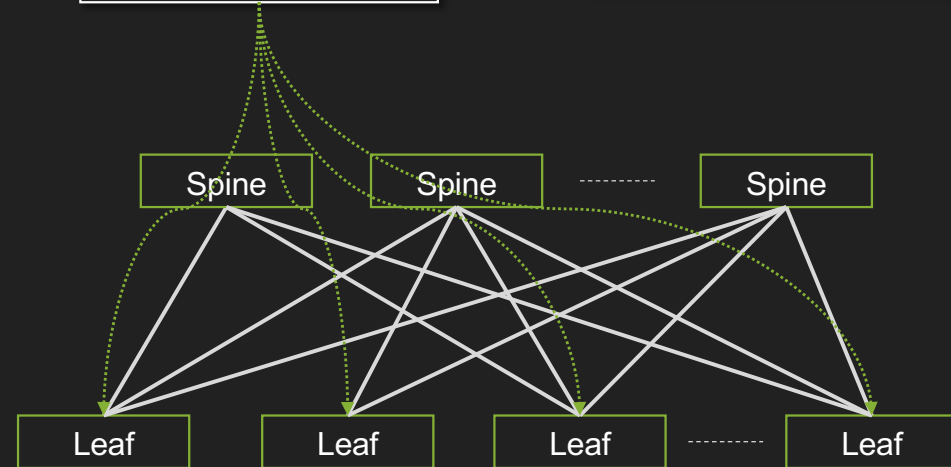
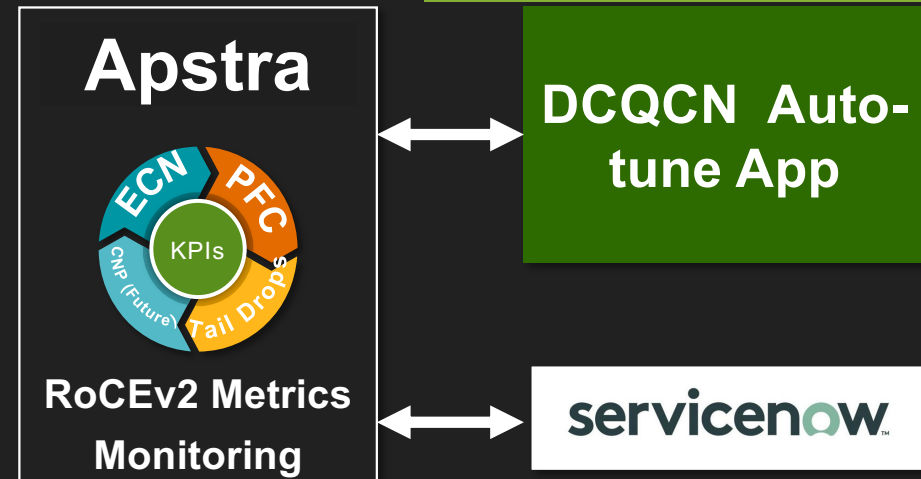
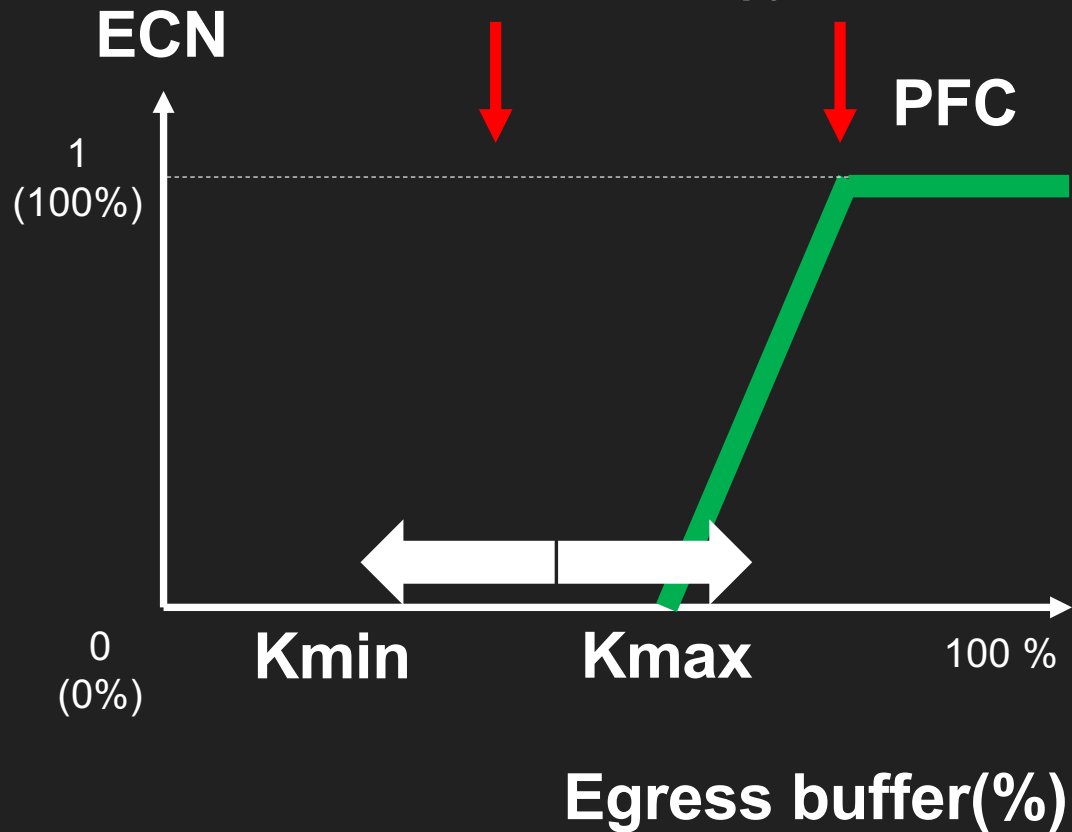
Explicit Congestion Notification (ECN)

Priority Flow Control (PFC)

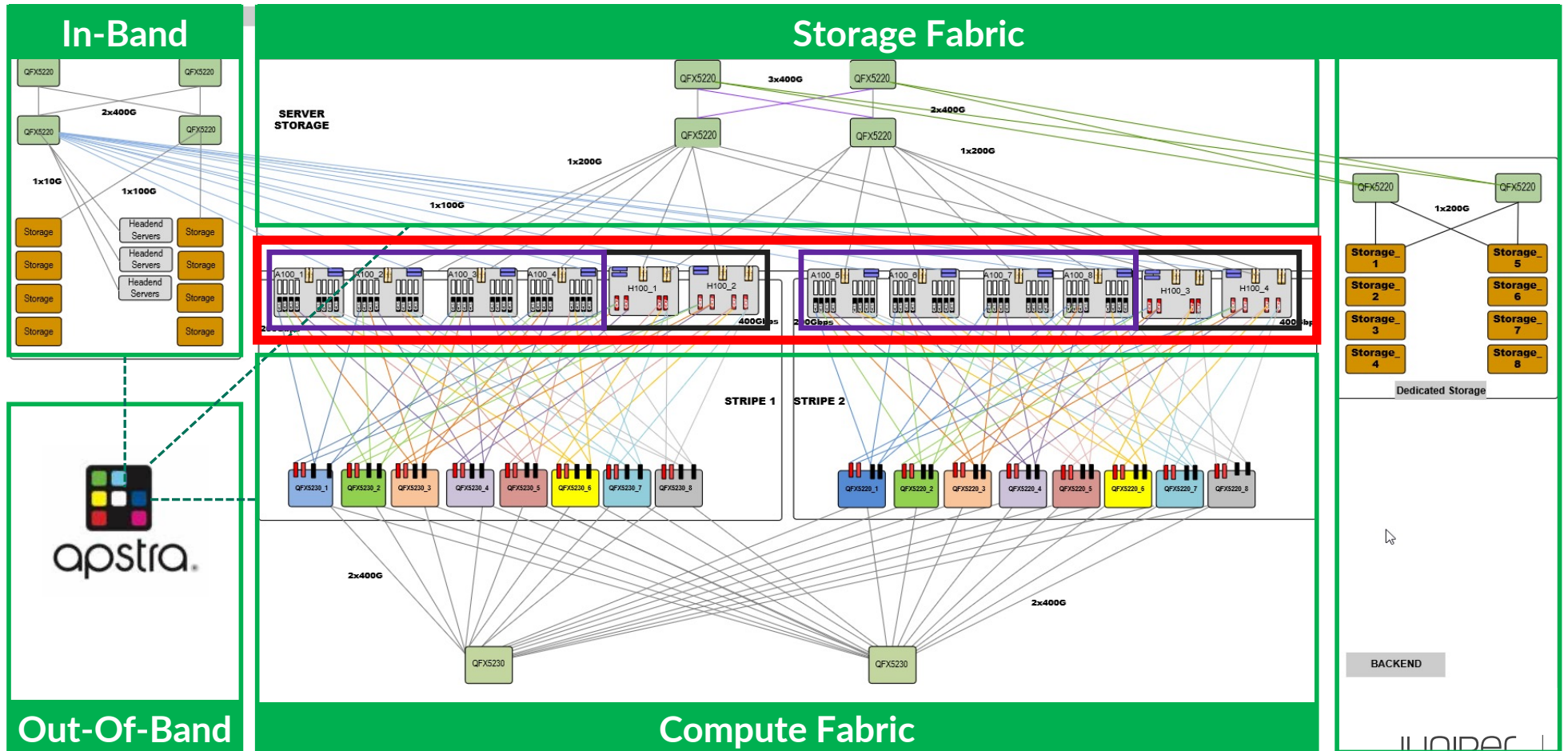
DC Quantized Congestion Notification (DCQCN)

“JNPR” Auto-Tune DCQCN

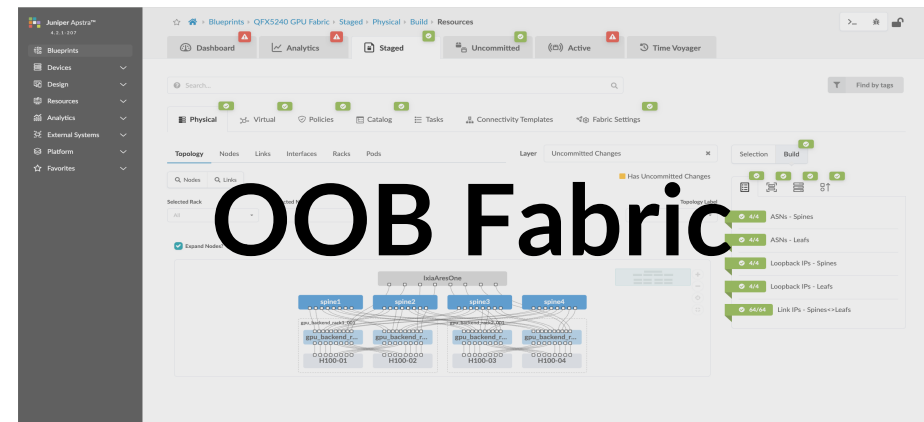
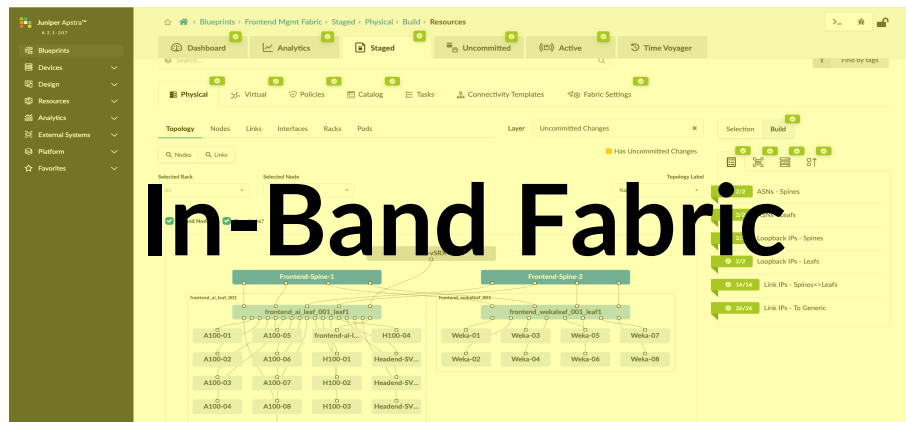
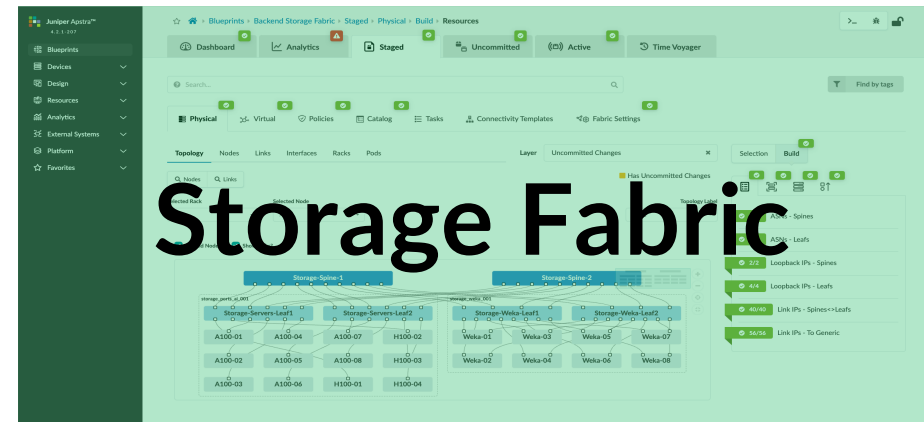
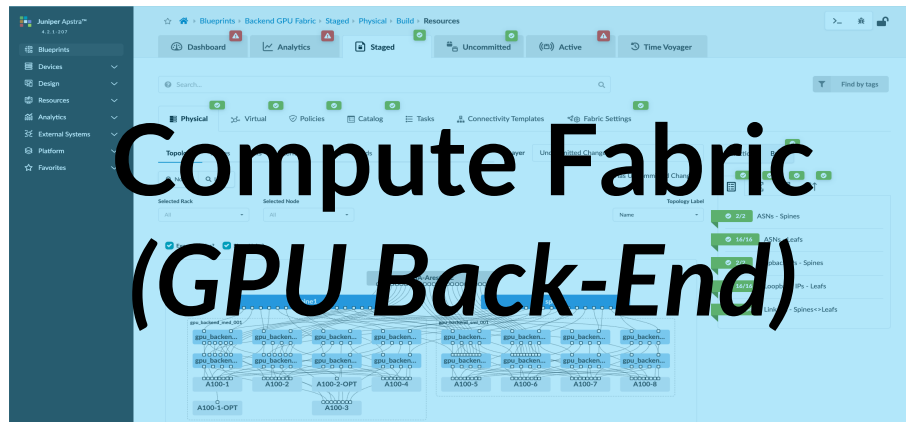
Buffer 50% Buffer Full



Nvidia's Reference Architecture



APSTRA - AI ML Cluster Blueprints

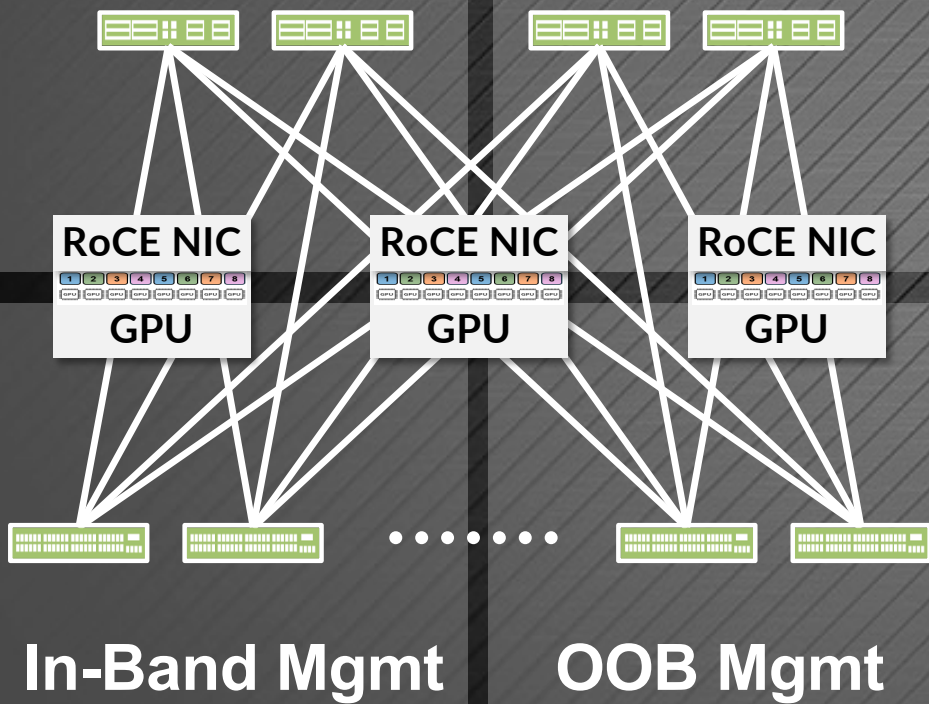


internal

JNPR Compute Fabric_(GPU Back-End) + Others

Compute Fabric

Storage Fabric



Compute Fabric

Storage Fabric

In-Band Mgmt

OOB Mgmt



TEST PLAN

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](#)

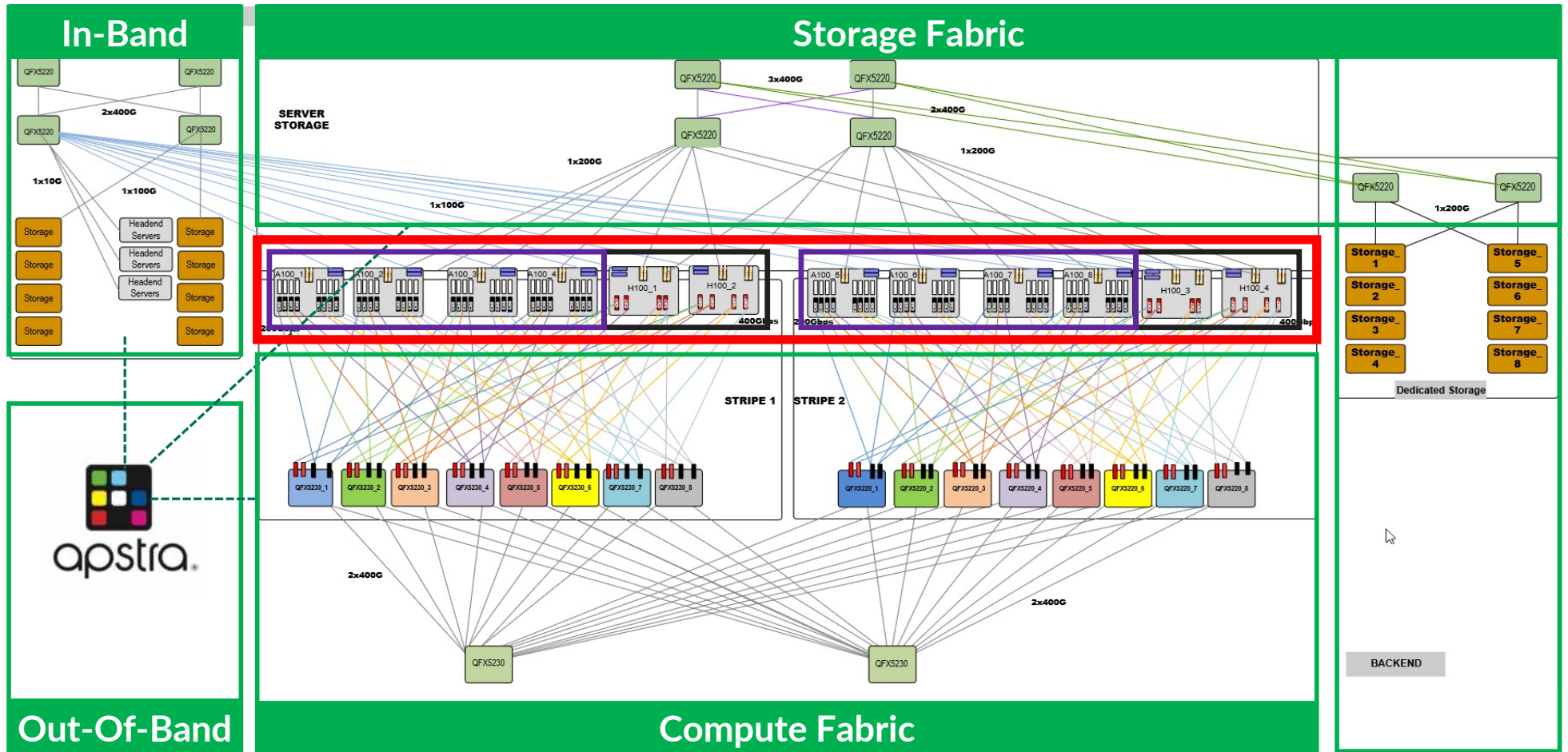


POC

TOPOLOGY & DUT

POC Topology

PoC Topology



PoC Topology for DCQCN

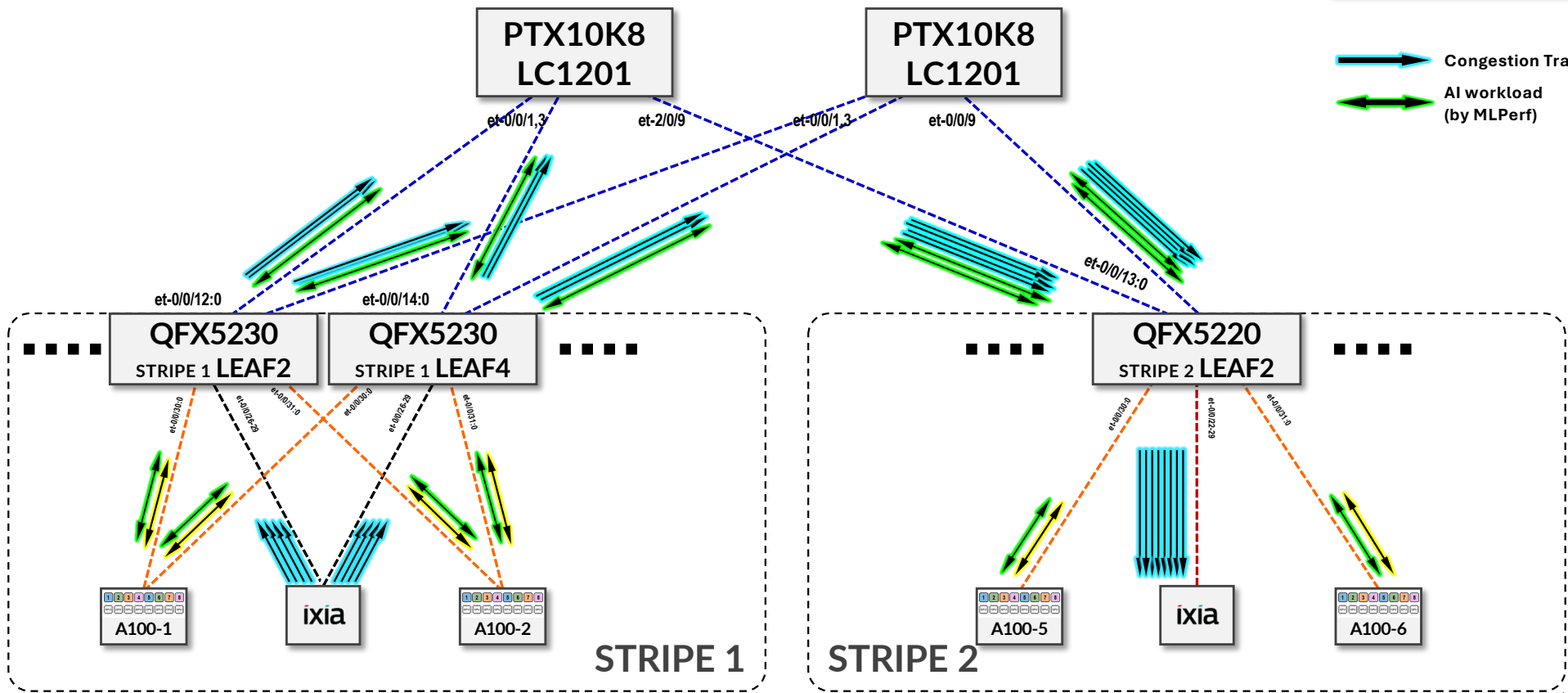
RoCEv2 TRAFFIC (A100 GPU & IXIA combined traffic)

INTERFACE BANDWIDTH

- - - - 400 Gbps x 1
- - - - 200 Gbps x 1
- - - - 200 Gbps x 4
- - - - 100 Gbps x 8

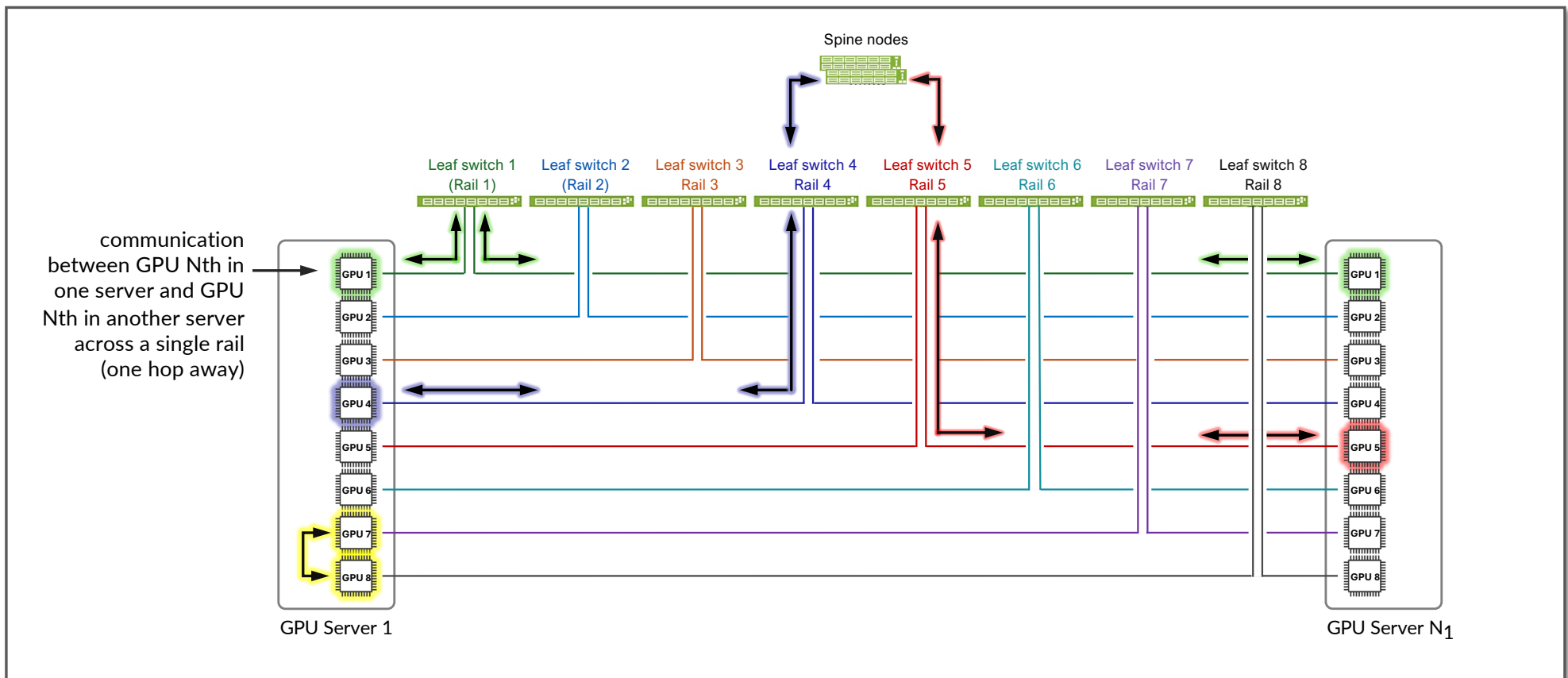
Legend:

- Congestion Traffic
- ↔ AI workload (by MLPerf)



PoC Topology for Rail-Optimized

Rail-Optimized Design



POC DUTs

* DUT: Device Under Test

PoC DUTs (1/4)

Sunnyvale PoC Lab



© 2025 Juniper Networks



Juniper Business Use Only

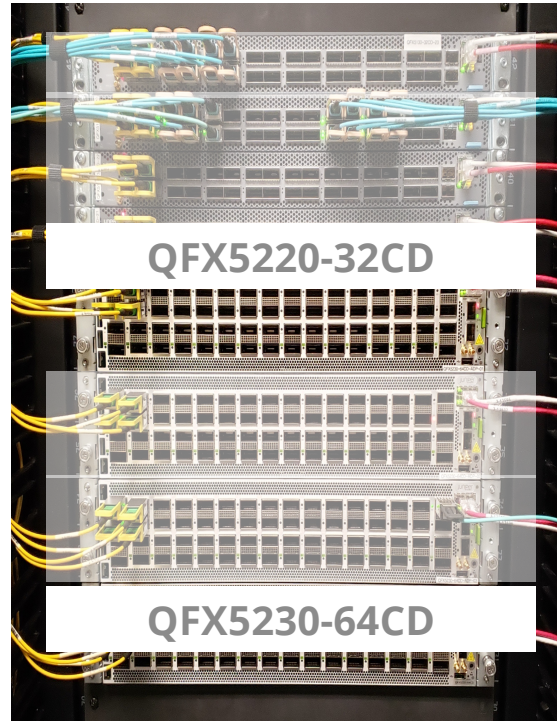
PoC DUTs (2/4) – Leaf/Spine switch

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

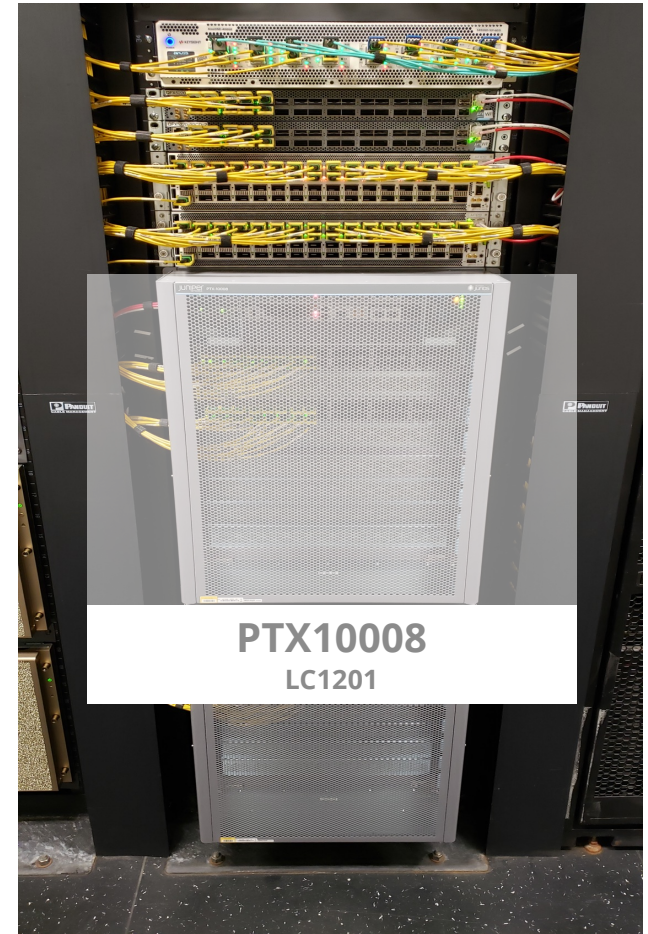
IXIA
traffic generator



Leaf
QFX5220-32CD/QFX5230-64CD

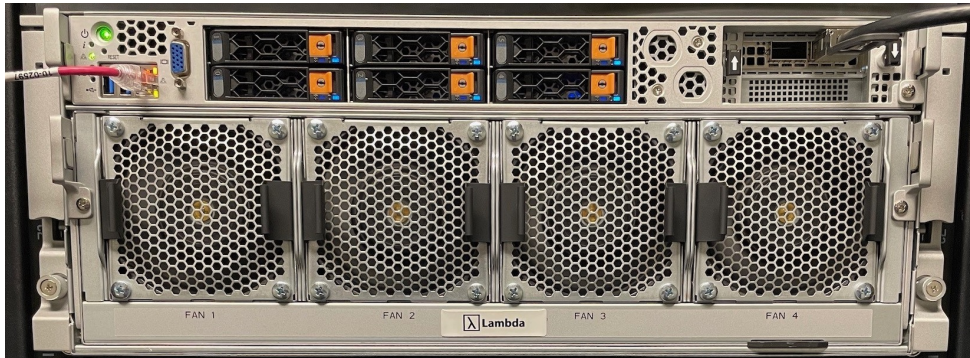


Spine
PTX10008 w/LC1201



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

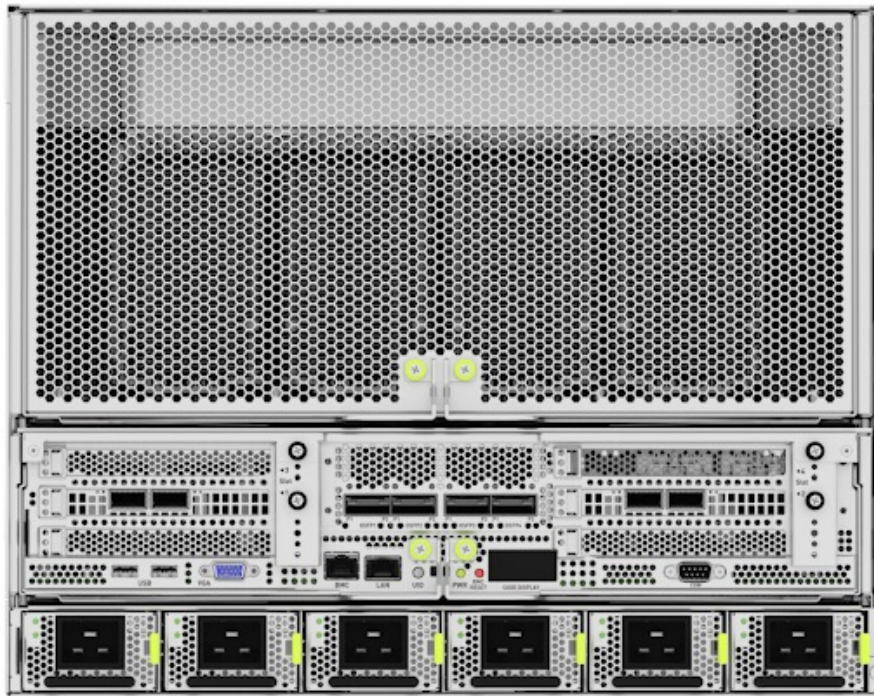
Power: 2+2 redundant 3000W PSUs, 4x C20 inlets, 200-240Vac input

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

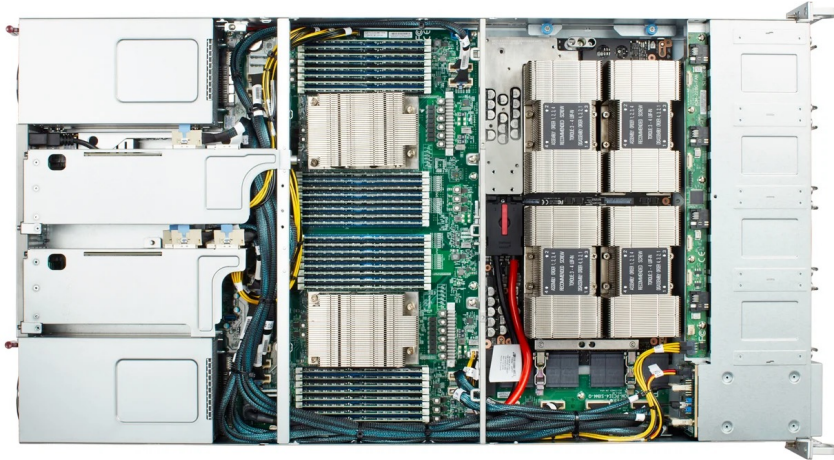
PoC DUTS (3/4) - GPU Node

Four of the NVIDIA DGX H100



PoC DUTs (3/4) – GPU Node

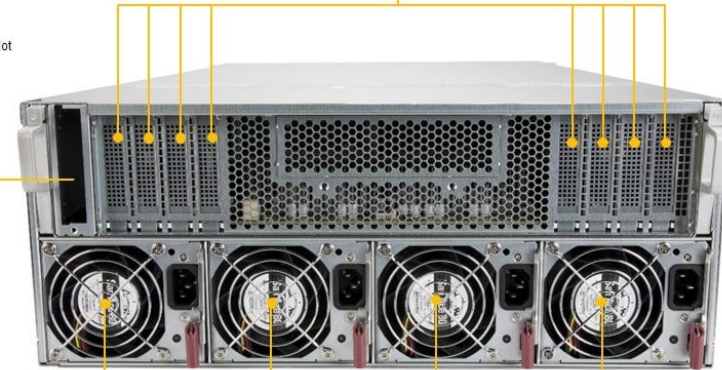
NVIDIA A100 GPU Node



(Rear View – System)

8x PCIe 4.0 X16 LP Slots
(via RDMA for IB EDR)

Optional AIOM Slot
PCI-E 4.0 X16



4x Redundant 2200W
Platinum Level
Power Supplies

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. |
|====+=====+====+=====+=====+=====+
|   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

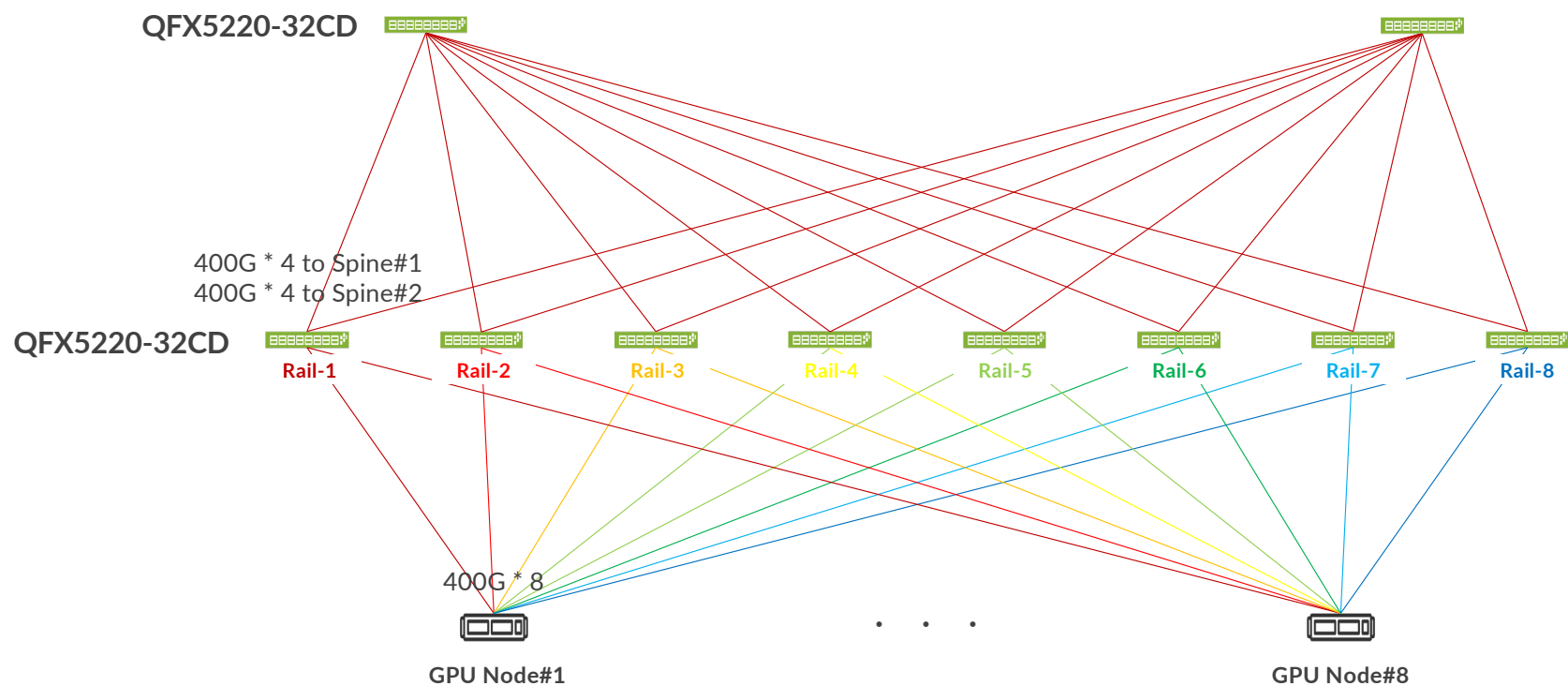
The screenshot displays the Juniper Apstra management console interface. The left sidebar contains navigation options: Blueprints, Devices, Design, Resources, Analytics, External Systems, Platform, Favorites, and Help. The main content area shows a 'Blueprints' overview with several circular status indicators (Deployment Status, Anomalies, Root Causes, Build Errors, Build Warnings, Uncommitted Changes) and a 'Create Blueprint' button. Below this, five blueprint cards are visible, each representing a different fabric cluster:

- Backend GPU Fabric Cluster 1** (Datacenter): Physical Structure (1 pod, 2 racks; 2 spines, 16 leaves, 10 generic systems); Virtual Structure (1 routing zone, 2 virtual networks); Analytics (3); Service Deployment Status (3); Service Anomalies (390); Probe Anomalies (31); Root Causes (3). Version 373, Total lines of config 18818, Last modified 2 days ago.
- Backend GPU Fabric Cluster 2** (Datacenter): Physical Structure (1 pod, 2 racks; 4 spines, 4 leaves, 6 generic systems); Virtual Structure (1 routing zone, 2 virtual networks); Analytics (14); Service Deployment Status (14); Service Anomalies (14); Probe Anomalies (14); Root Causes (14). Version 312, Total lines of config 8462, Last modified 14 days ago.
- Backend Storage Fabric** (Datacenter): Physical Structure (1 pod, 2 racks; 2 spines, 4 leaves, 28 generic systems); Virtual Structure (1 routing zone, 2 virtual networks); Analytics (12); Service Deployment Status (12); Service Anomalies (12); Probe Anomalies (12); Root Causes (12). Version 271, Total lines of config 3966, Last modified 3 days ago.
- Frontend Mgmt Fabric** (Datacenter): Physical Structure (1 pod, 2 racks; 2 spines, 2 leaves, 29 generic systems); Virtual Structure (1 routing zone, 2 virtual networks); Analytics (3); Service Deployment Status (3); Service Anomalies (3); Probe Anomalies (3); Root Causes (3). Version 181, Total lines of config 1534, Last modified 2 days ago.
- QFX5240 GPU Fabric** (Datacenter): Physical Structure (1 pod, 2 racks; 4 spines, 4 leaves, 5 generic systems); Virtual Structure (1 routing zone, 24 virtual networks); Analytics (N/A); Service Deployment Status (N/A); Service Anomalies (N/A); Probe Anomalies (N/A); Root Causes (N/A). Version 424, Last modified a month ago.

Expected Diagram(Phase 1 - Trial Biz)

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

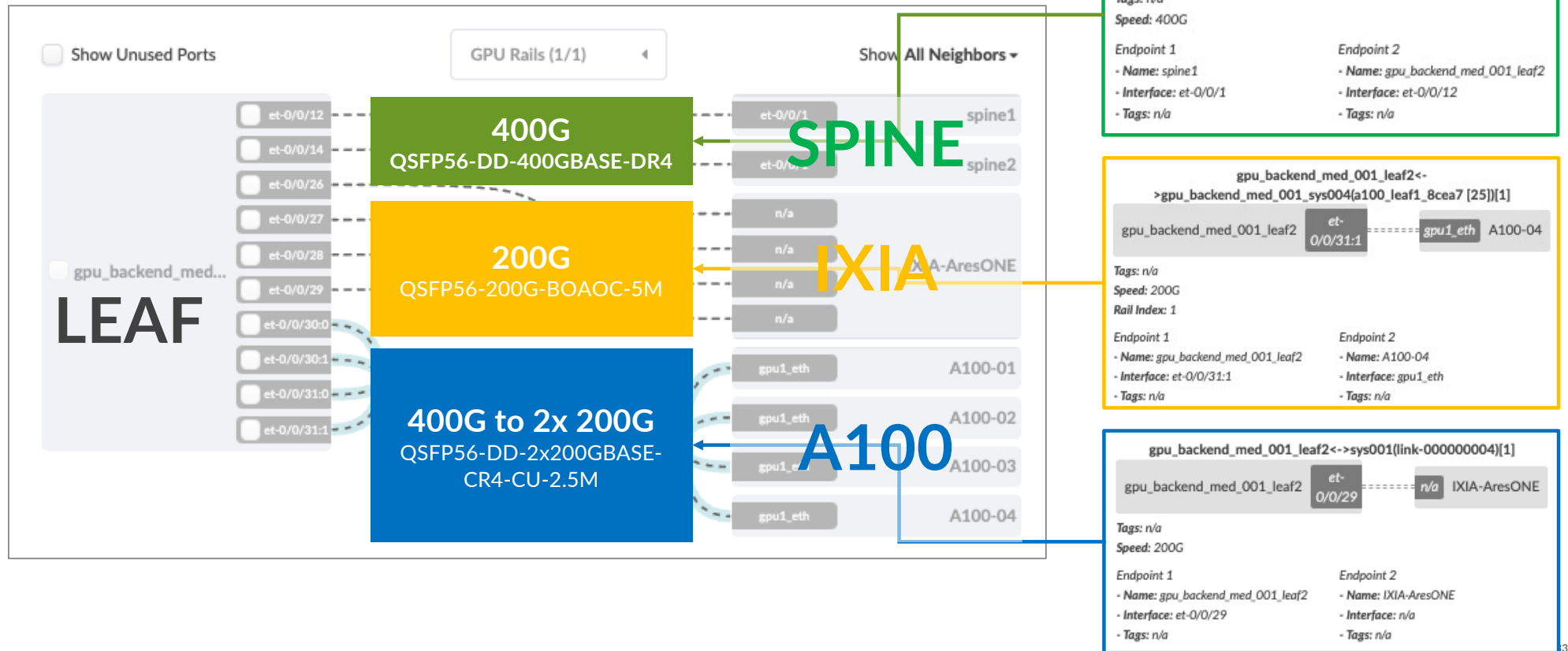
Phase1(Trial) (GPU Server 8 nodes)



Transceiver(Optic)

PoC Transceiver Info.

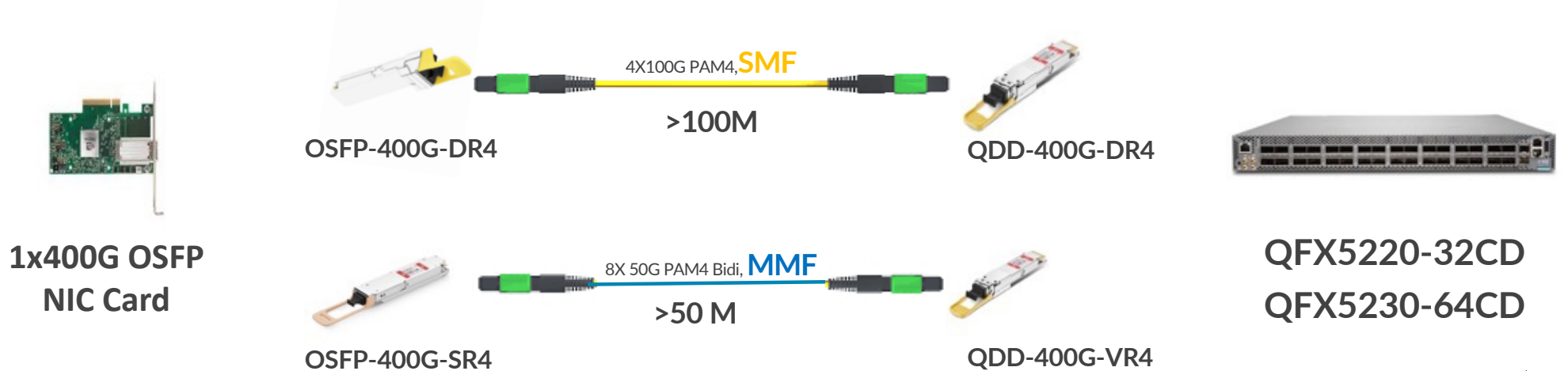
200G / 400G(B/O) to GPU & IXIA



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





JUNIPER AI/ML SOLUTION PORTFOLIO

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



5+1 Fan FRU

1600W PSU FRU 1+1 Redundancy

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

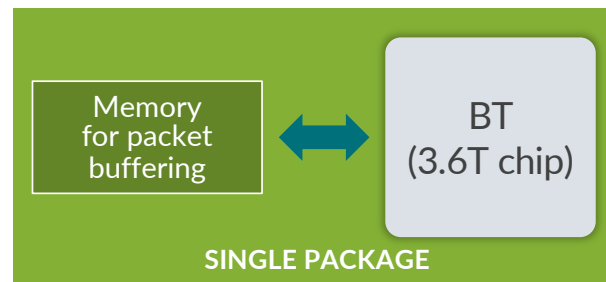


4 slot chassis

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

Juniper Apstra™
4.2.1-207

- Blueprints
- Devices
- Design
- Resources
- Analytics
- External Systems
- Platform
- Favorites

☆ 🏠 > Blueprints

Deployment Status

Anomalies

Root Causes

Build Errors

Build Warnings

Uncommitted Changes

➕ Create Blueprint

⋮

1-4 of 4
« < 1 > »

Backend GPU Fabric
Datacenter

Physical Structure:	1 pod, 2 racks 2 spines, 16 leaves, 13 generic systems
Virtual Structure:	1 routing zone, 16 virtual networks
Analytics	
Deployment Status	18
Service Anomalies	18
Probe Anomalies	18
Root Causes:	0

Version 365
Total lines of config 15438 Last modified 18 days ago

Backend Storage Fabric
Datacenter

Physical Structure:	1 pod, 2 racks 2 spines, 4 leaves, 20 generic systems
Virtual Structure:	1 routing zone
Analytics	
Deployment Status	6
Service Anomalies	0
Probe Anomalies	0
Root Causes:	0

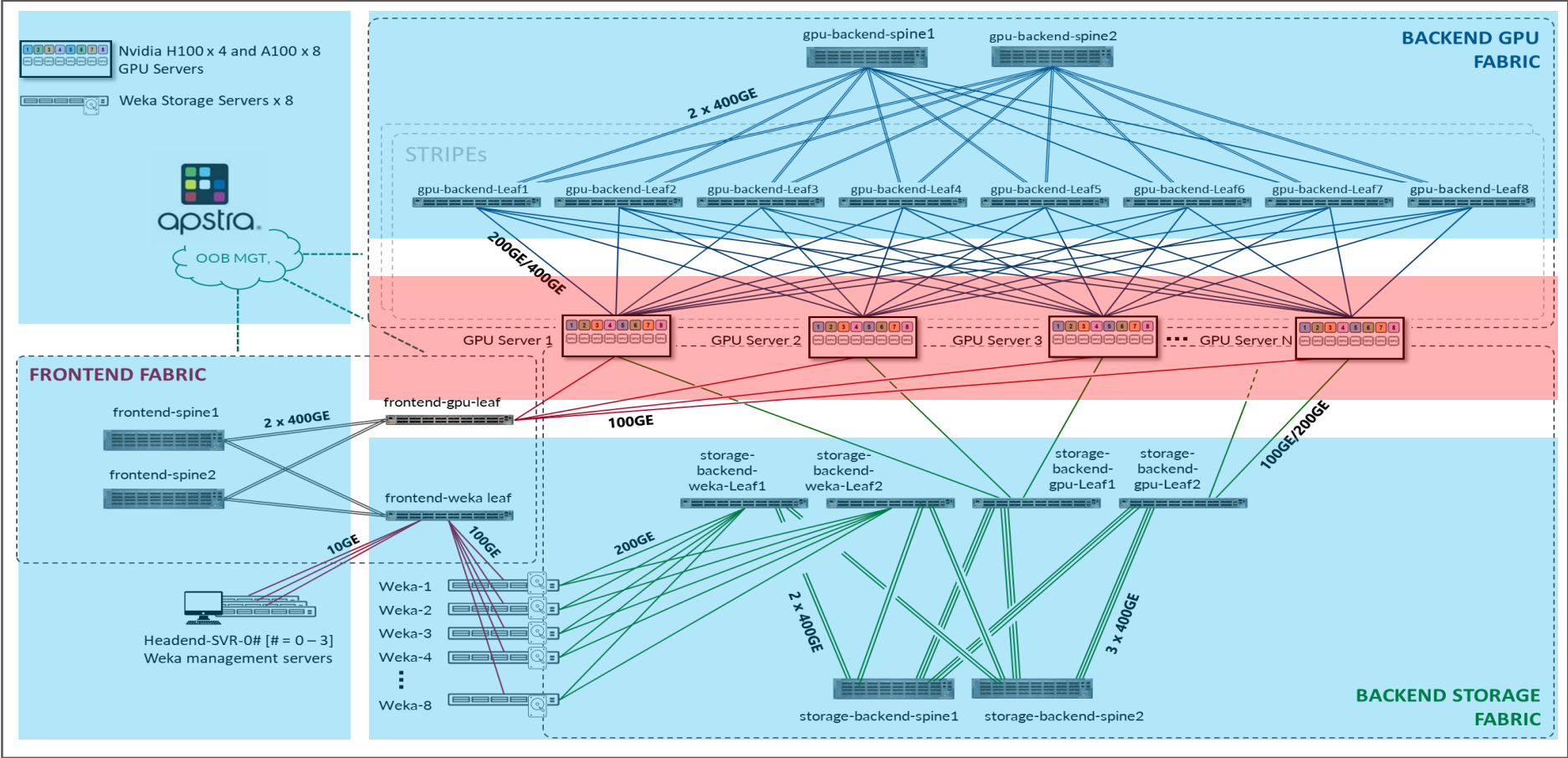
Version 120
Total lines of config 3600 Last modified a month ago

Frontend Mgmt Fabric
Datacenter

Physical Structure:	1 pod, 2 racks 2 spines, 2 leaves, 25 generic systems
Virtual Structure:	1 routing zone, 2 virtual networks
Analytics	
Deployment Status	4
Service Anomalies	0
Probe Anomalies	0
Root Causes:	0

Version 82
Total lines of config 2207 Last modified 12 days ago

AI Innovation Lab - Fabrics



APSTRA - AI ML Cluster Blueprints

The screenshot shows the Juniper Apstra interface for the 'Compute Fabric' blueprint. The interface is primarily blue and white. The main area displays a network topology diagram with various nodes and connections. A large, bold, black text overlay 'Compute Fabric' is centered on the screen. The left sidebar contains navigation options like 'Blueprints', 'Design', 'Resources', 'External Systems', and 'Platform'. The top navigation bar includes 'Dashboard', 'Analytics', 'Staged', 'Uncommitted', 'Active', and 'Time Voyager'.

The screenshot shows the Juniper Apstra interface for the 'Storage Fabric' blueprint. The interface is primarily green and white. The main area displays a network topology diagram with nodes and connections. A large, bold, black text overlay 'Storage Fabric' is centered on the screen. The left sidebar and top navigation bar are identical to the Compute Fabric screenshot.

The screenshot shows the Juniper Apstra interface for the 'In-Band Fabric' blueprint. The interface is primarily yellow and white. The main area displays a network topology diagram with nodes and connections. A large, bold, black text overlay 'In-Band Fabric' is centered on the screen. The left sidebar and top navigation bar are identical to the other screenshots.

The screenshot shows the Juniper Apstra interface for the 'OOB Fabric' blueprint. The interface is primarily grey and white. The main area displays a network topology diagram with nodes and connections. A large, bold, black text overlay 'OOB Fabric' is centered on the screen. The left sidebar and top navigation bar are identical to the other screenshots.



THANK YOU

JUNIPER
NETWORKS | Driven by
Experience™