Juniper® Validated Design

JVD Test Report Brief: AI Data Center Network with Juniper Apstra, AMD GPUs, and Vast Storage

test-report-brief- jvd-aiclusterdc-aiml-AMD-03-03

Introduction

This document outlines the qualification testing of the AI ML Cluster solution for effective congestion control with AMD MI300 GPUs, utilizing an IP-Clos architecture with QFX5240-64CD/OD and QFX5241-64CD switches as DUTs and Apstra as the control point.

The AI ML cluster includes (as depicted in Figure 1):

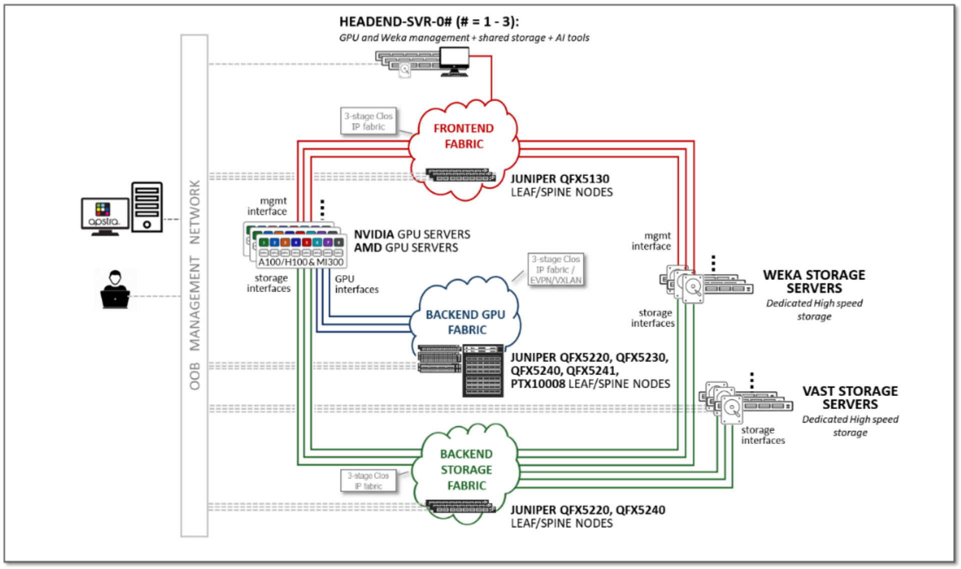
1. Front-end 3-clos IP fabric with QFX5220 switches as spine and leaf nodes connecting the headend servers with the AMD GPU servers for job management and weka storage devices

2. Dedicated storage backend 3-clos IP Fabric with QFX5220s or QFX5240s connecting the Vast storage servers and AMD GPU servers.

3. Backend GPU (compute) 3-clos IP fabric as described in the Test Topology section.

4. The Cluster solution is orchestrated by Juniper Apstra and hence is dependent on a Juniper Apstra server and Apstra flow deployment for configurations and telemetry.

Figure 1. AI ML cluster



Test Topology

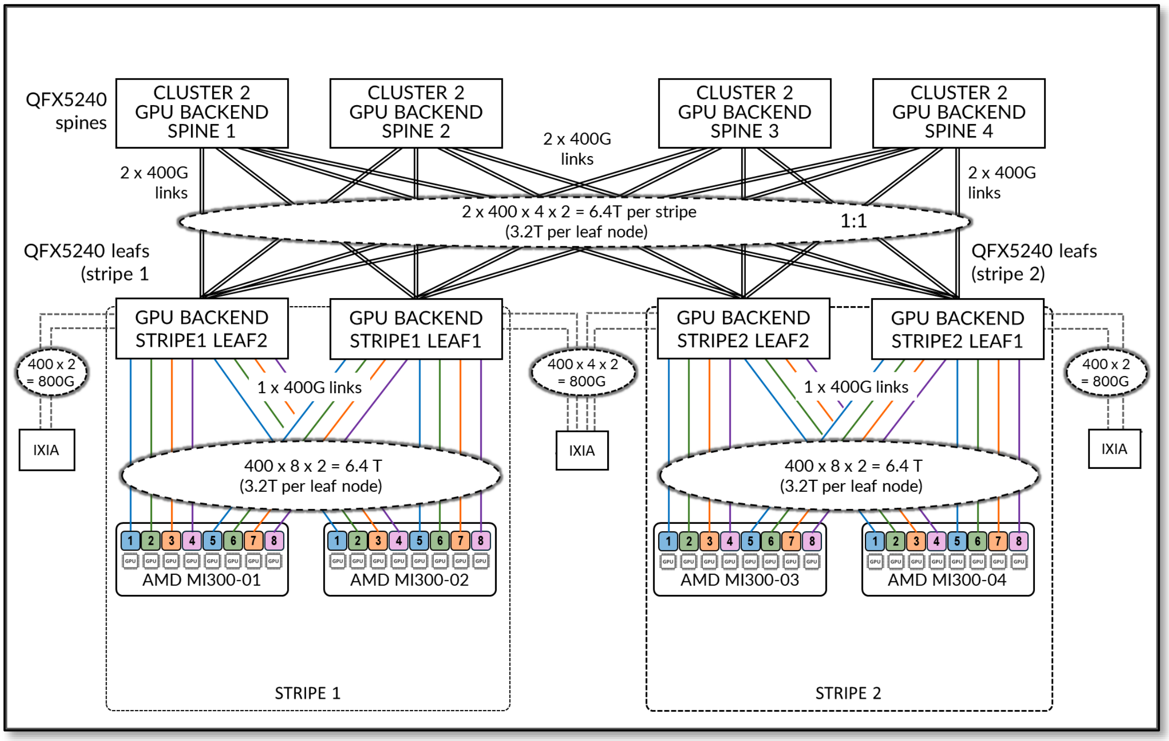
The Backend GPU (compute) 3-clos IP fabric includes the following components, as depicted in Figures 2 and 3):

LEAF NODES	QFX5230/QFX5240/QFX5241
SPINE NODES	QFX5240/QFX5241
GPU Servers	AMD MI300 GPUs with: Option 1: Broadcom Thor2 and Option 2: AMD Pollara

Leaf to Spine links can be either:

- Native QSFP-800G/400G/100G
- Channelized OSFP 2x400G, 8X100G

Figure 2: Reference Topology



QFX5240 spines
CLUSTER 2 GPU BACKEND SPINE 1
CLUSTER 2 GPU BACKEND SPINE 2
CLUSTER 2 GPU BACKEND SPINE 3
CLUSTER 2 GPU BACKEND SPINE 4
2 x 400G links
2 x 400G links
2 x 400G links
2 x 400 x 4 x 2 = 6.4T per stripe (3.2T per leaf node)
1:1
QFX5240 leafs (stripe 1
QFX5240 leafs (stripe 2)
GPU BACKEND STRIPE1 LEAF2
GPU BACKEND STRIPE1 LEAF1
GPU BACKEND STRIPE2 LEAF2
GPU BACKEND STRIPE2 LEAF1
400 x 2 = 800G
400 x 4 x 2 = 800G
400 x 2 = 800G
IXIA
IXIA
IXIA
1 x 400G links
1 x 400G links
400 x 8 x 2 = 6.4 T (3.2T per leaf node)
400 x 8 x 2 = 6.4 T (3.2T per leaf node)
1 2 3 4 5 6 7 8
GPU
AMD MI300-01
AMD MI300-02
AMD MI300-03
AMD MI300-04
STRIPE 1
STRIPE 2

If you have questions about this Juniper Validated Design, contact your Juniper representative.

Test Approach

The testing plan focused on validating a set of congestion control scenarios with Junos 23.4X100-D20, implementing congestion control and ECMP DLB flow-let mode for lossless RoCEv2 traffic forwarding in the AI Data Center Network with Juniper Apstra, AMD GPUs, and VAST Storage Juniper Validated Design (JVD) solution, when running RoCEv2 traffic flows.

These congestion scenarios include fine-tuning dependent parameters (shared buffer allocation, drop-profiles, DLB for a lossless fabric, and more). The overall goal is to establish a lossless RoCEv2 network with ECMP DLB flow-let mode enabled.

Job Completion Time (JCT) values are compared against the MLCommons benchmarks.

Test Goals and Non-Goals

The goal of this phase is to arrive at a better controlled performance of an AI ML storage cluster compared to previously validated values of congestion threshold parameters with the fine-tuning of additional knobs supported in the Junos OS Evolved 23.4X100-D30 release for improved throughput, latency and JCT values with AMD MI300 GPU servers.

Additional goals include:

1. Determine an optimal Alpha-per-Queue setting for a profile experiencing varying congestion levels with dynamic alpha setting possible per queue level.

2. Establish a PFC XON limit for the ingress ports PG shared buffer threshold setting at which a peer resumes transmitting the packets after a brief PAUSE because of the PFC sent by this node.

3. Monitor ECN-marked packets per congested queue at the CLI level, rather than at the interface level, to enable congestion control specific to each congested queue.

4. Determine optimal values for PFC Watchdog parameters to detect and mitigate PFC pause storms for a recovery from the congested scenario. While this feature itself avoids PFC propagating through the network due to back-pressure halting the traffic, this also ensures the PFC deadlock does not happen in case of link/device failures.

Non-goals for this validation include:

1. Validation of base protocols like EBGP and BFD; this validation focused on congestion threshold parameters.

Platforms Tested

Table 1: Platforms, Controllers, and Roles

Role	Platform	OS
Leaf1	QFX5240/5230	Junos OS Evolved 23.4X100-D30
Leaf2	QFX5240/5230	Junos OS Evolved 23.4X100-D30
Leaf3	QFX5240/5230	Junos OS Evolved 23.4X100-D30
Leaf4	QFX5240/5230	Junos OS Evolved 23.4X100-D30
Spine1	QFX5240/5230	Junos OS Evolved 23.4X100-D30
Spine2	QFX5240/5230	Junos OS Evolved 23.4X100-D30
Spine3	QFX5240/5230	Junos OS Evolved 23.4X100-D30
Spine4	QFX5240/5230	Junos OS Evolved 23.4X100-D30
TGen	Ixia	IxOS 10.25
Control Point	Apstra	5.0

ROLE	PLATFORM	VERSION
LEAF	QFX5230-64CD	Junos OS Evolved 23.4X100-D30
	QFX5240-64OD	Junos OS Evolved 23.4X100-D30
	QFX5241-64OD	Junos OS Evolved Release 23.4X100-D42
SPINE	QFX5240-64OD	Junos OS Evolved 23.4X100-D30
	QFX5241-64OD	Junos OS Evolved Release 23.4X100-D42
	PTX10008 with JNP10K-LC1201 linecard	Junos OS Evolved Release 23.4R2-S3.10
CLUSTER DEPLOYMENT/MONITORING	Apstra	5.0

Version Qualification History

This JVD has been qualified in Junos OS Evolved Release 23.4X100-D30.

Test Environment

Buffer management values (default):

- Shared buffer – lossless 80%, headroom 10%, lossy 10%
- Dynamic threshold – 7 (default)
- ECN fill level – 55%

DLB values (default):

- flowlet-table-size 256
- flowlet inactivity-timer 256us
- flowlet sampling-rate 62500/s
- flowset table-size 512
- flowlet egress-quantization min 20
- flowlet egress-quantization max 50
- flowlet egress-quantization rate-weightage 50
- flowlet reassignment disabled

Traffic Flows:

- Ixia RoCEv2 Tx traffic: 50% (400G), 75%(600G) & 100%(800G) traffic load sent to all leaf's.
- 16 QPs per port from IXIA + Model to the Leaf's as ingress traffic

Performance Data for Cluster-2 (with MI-300 GPUs)

Table 2: LLAMA3 model Job Completion time (JCT) test results with 4 spines

Traffic Profile	Tuned parameters in leaf/spine (all other parameters as per defaults above)	JCT [min]
Only Model	sampling rate 1000000 flowset-table-size 2048	40.23
Only Model	reassignment prob-threshold 3 reassignment quality-delta 6 sampling rate 1000000 flowset-table-size 2048	36.44
Only Model	reassignment prob-threshold 3 reassignment quality-delta 6 sampling rate 1000000 flowset-table-size 2048 egress-quantization rate-weightage 80	36.57
Only Model	reassignment prob-threshold 3 reassignment quality-delta 3 sampling rate 1000000 flowset-table-size 2048	54.35
Only Model	reassignment prob-threshold 3 reassignment quality-delta 6 sampling rate 1000000 flowset-table-size 2048 inactivity-interval 512	39.08
Model + Ixia (50% traffic)	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	36.34
Model + Ixia (75% traffic)	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	36.35
Model + Ixia (100% traffic)	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	39.11

Table 3: LLAMA3 model Job Completion time (JCT) congestion test results with 2 spines

Traffic Profile	Tuned parameters in leaf/spine (all other parameters as per defaults above)	JCT [min]
Model + Ixia (50% traffic)	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	39.7
Model + Ixia (75% traffic)	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	39.5

Traffic Profile	Tuned parameters in leaf/spine (all other parameters as per defaults above)	JCT [min]
Model + Ixia (75% traffic)	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6 flowlet rate-weightage 80 flowlet egress-quantization min 10 flowlet egress-quantization min 90 dynamic-threshold 9	42.02

Table 4: Broadcom Thor2 NIC RCCL Bandwidth tests (with 4 spines and 32 GPUs)

Traffic Profile	Tuned parameters in leaf/spine (all other parameters as per defaults above)	Bandwidth (GB/s)
RCCL all-reduce	Default	343.86
RCCL all-reduce	COS disabled	336.9
RCCL all-reduce	DLB disabled	336.72
RCCL all-reduce	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	347.9
RCCL all-reduce	flowset-table-size 2048 sampling rate 1000000 flowlet rate-weightage 80 flowlet egress-quantization min 10 flowlet egress-quantization min 90	347.14
RCCL all-reduce	flowset-table-size 2048 sampling rate 1000000 flowlet rate-weightage 60 flowlet egress-quantization min 10 flowlet egress-quantization min 90	347.25
RCCL all-to-all	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	45.75

Table 5: Broadcom Thor2 NIC RCCL Bandwidth congestion tests (with 2 spines and 32 GPUs)

Traffic Profile	Tuned parameters in leaf/spine (all other parameters as per defaults above)	Bandwidth (GB/s)
RCCL all-reduce	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	148.8

Traffic Profile	Tuned parameters in leaf/spine (all other parameters as per defaults above)	Bandwidth (GB/s)
RCCL all-reduce	flowset-table-size 2048 sampling rate 1000000	331.77
RCCL all-reduce	DLB disabled	317.5
RCCL all-reduce + Ixia (50% traffic)	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	92.08
RCCL all-reduce + Ixia (50% traffic)	flowset-table-size 2048 sampling rate 1000000	317.1
RCCL all-reduce + Ixia (100% traffic)	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	119.1
RCCL all-reduce + Ixia (100% traffic)	flowset-table-size 2048 sampling rate 1000000	315.5
RCCL all-reduce + Ixia (100% traffic)	flowset-table-size 2048 sampling rate 1000000 flowlet rate-weightage 60	314.92
RCCL all-reduce + Ixia (100% traffic)	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 255 reassignment quality-delta 6	108.27
RCCL all-to-all + Ixia (50% traffic)	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	11.34
RCCL all-to-all + Ixia (50% traffic)	flowset-table-size 2048 sampling rate 1000000	9.9

The following table shows the results of the AMD Pollara 400 NIC RCCL bandwidth tests with 4 spines and 16 GPUs. These tests use the same AI/ML workloads as the previous tests.

Table 6: AMD Pollara NIC RCCL Bandwidth Tests (with 4 spines and 16 GPUs)

Traffic Profile	Tuned parameters in leaf/spine (all other parameters as per defaults above)	Bandwidth (GB/s)
RCCL all-reduce	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	346.8
RCCL all-reduce + Ixia (50% traffic)	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	345..8

Traffic Profile	Tuned parameters in leaf/spine (all other parameters as per defaults above)	Bandwidth (GB/s)
RCCL all-reduce + Ixia (100% traffic)	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	343.9
RCCL all-to-all	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	83.04
RCCL all-to-all + Ixia (50% traffic)	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	82.8
RCCL all-to-all + Ixia (100% traffic)	flowset-table-size 2048 sampling rate 1000000 reassignment prob-threshold 3 reassignment quality-delta 6	83.3

Corporate and Sales Headquarters
Juniper Networks, Inc.
1133 Innovation Way
Sunnyvale, CA 94089 USA
Phone: 888.JUNIPER (888.586.4737)
or +1.408.745.2000
Fax: +1.408.745.2100
www.juniper.net

APAC and EMEA Headquarters
Juniper Networks International B.V.
Boeing Avenue 240
1119 PZ Schiphol-Rijk
Amsterdam, The Netherlands
Phone: +31.207.125.700
Fax: +31.207.125.701