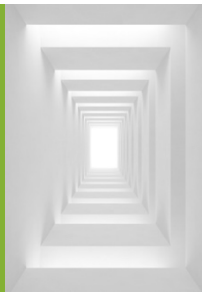
Juniper® Validated Design

JVD Test Report Brief: AI Data Center Multitenancy with EVPN/VXLAN



test-report-brief-JVD-AICLUSTERDC-EVPNType5-01-04

Introduction

This test report brief contains qualification test results for the AI Data Center Multitenancy with EVPN/VXLAN—Juniper Validated Design (JVD) GPU Backend fabric which provides the infrastructure for GPU to GPU communication using RDMA over Converged Ethernet (RoCEv2).

The GPU Backend fabric is designed to provide a near lossless fabric, achieving maximum throughput, minimal latency, and minimal network interference for AI traffic flows. Customers looking to implement GPU as a Service (GPUaaS) can follow this JVD.

Two types of GPU multitenancy are part of the testing: Server Isolation, where entire servers are dedicated to single tenants, and GPU Isolation, where individual GPUs within a server are assigned to different tenants.

Testing focused on validating the components of the EVPN/VXLAN solution, including connectivity between GPU servers and leaf nodes, and fabric operations, and congestion management, and load balancing ensuring lossless transmission of RDMA.

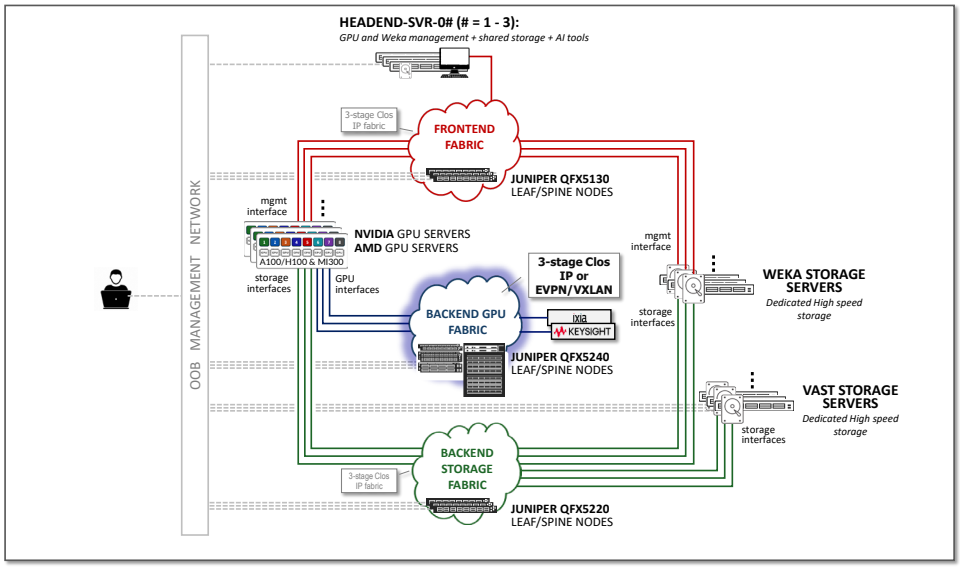
Key tests included:

- Validating IPv6 underlay BGP operation using BGP auto-discovery (unnumbered) peering.
- Validating IPv6 SLAAC (Stateless Address Autoconfiguration) to provide dynamic IPv6 address assignment to the GPU servers interfaces.
- Validating IPv6 overlay BGP operation
- Validating IPv6/IPv4 BGP overlay with family EVPN signaling.
- Verifying EVPN-VXLAN control plane with Type-5 routes (L3 IP-VRFs).
- Verifying EVPN-VXLAN forwarding plane by ensuring IP reachability between IP-VRF instances.
- Verifying Pure Type-5 route adveritising and proper installation for each Tenants' IP-VRF
- Validating per Tenant end-to-end RoCEv2 traffic.
- Validating Congestion Management using DCQCN.
- Validating load balancing using DLB.
- Conducting fault tolerance testing (link/node failures, failover behavior).

Test Topology

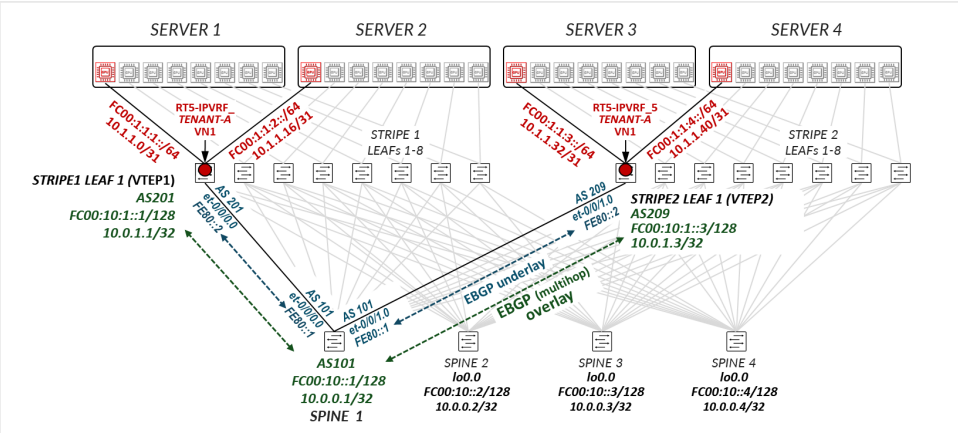
The AI cluster Lab topology consists of 3 Fabrics: Frontend (IP fabric), Backend Storage (IP fabric), and Backend GPU (IP fabric or EVPN/VXLAN).

Figure 1: Test Topology



Testing for this JVD was limited to the Backend GPU with EVPN/VXLAN and EVPN BGP type 5 routes. With IPv6 BGP unnumbered underlay. Both IPv4 (RFC5549) and IPv6 overlay were tested.

Figure 2: Test GPU Backend Topology



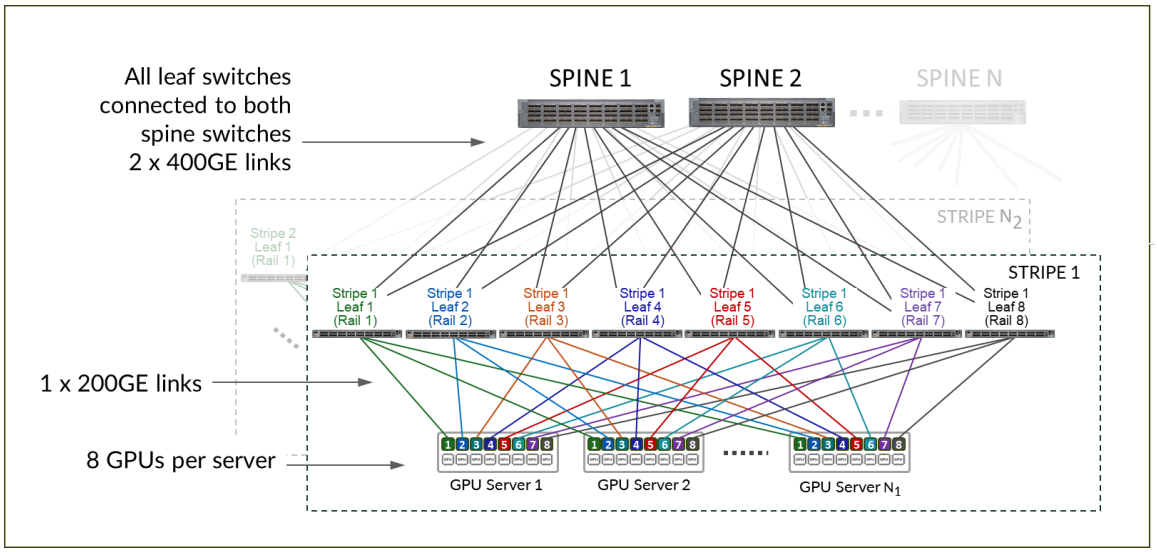
Functionality testing was performed with RoCEv2 traffic generated using IXIA traffic generator. Both QFX5240 and QFX5230 were tested in spine node roles, while QFX5130, QFX5230, and QFX5240 were tested in the leaf role.

Performance testing was completed with a combination of AMD and NVIDIA GPU servers, and RoCEv2 traffic generated using IXIA. With QFX5240s in the leaf and spine node roles.

Table 1: Topology and Devices

Layer	Devices
Spine Layer	QFX5240-64QD QFX5230-64QD
Leaf Layer	QFX5240-64OD QFX5230-64CD 3x QFX5130-32CD

The links between the leaf and spine nodes are 400G x 2. The links between the leaf nodes and the GPU servers are 1 x 200G connected following a rail optimized architecture.



Connectivity between the servers and the leaf nodes was tested with static IPv4 addresses, static IPv6 addresses, and dynamicaly assigned IPv6 addresses via SLAAC.

Figure 2: Logical Topology

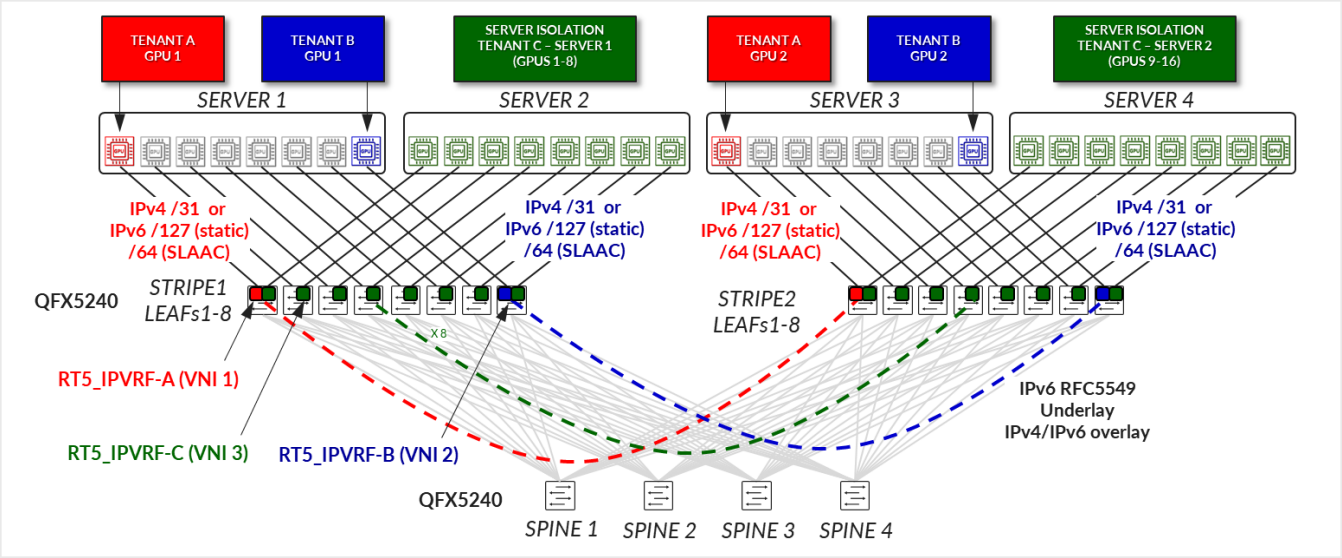


Table 2: Platforms, Controllers, and Roles

Tag	Role	Model	OS	Linecard	RE	SW Fabric	Virtial Chassis	Helper/ DUT
R0	Leaf1	QFX5240-64OD	JUNOS EVO 23.4X100-D31	NA	NA	NA	NA	DUT
R1	Leaf2	QFX5230-64CD	JUNOS EVO 23.4X100-D31	NA	NA	NA	NA	DUT
R2	Leaf3	QFX5130-32CD	JUNOS EVO 23.4X100-D31	NA	NA	NA	NA	DUT
R3	Leaf4	QFX5130-32CD	JUNOS EVO 23.4X100-D31	NA	NA	NA	NA	DUT
R4	Leaf5	QFX5130-32QD	JUNOS EVO 23.4X100-D31	NA	NA	NA	NA	DUT
R5	Spine1	QFX5240-64QD	JUNOS EVO 23.4X100-D31	NA	NA	NA	NA	DUT
R6	Spine2	QFX5240-64QD	JUNOS EVO 23.4X100-D31	NA	NA	NA	NA	DUT

R7	Spine3	QFX5230-64CD	JUNOS EVO 23.4X100-D31	NA	NA	NA	NA	DUT
R8	Spine4	QFX5230-64CD	JUNOS EVO 23.4X100-D31	NA	NA	NA	NA	DUT
RT0	TGen	Ixia	IxOS 10.25	NA	NA	NA	NA	Helper
H100-01 – H100-04	GPU server	Nvidia H100 GPU servers	CUDA 12.6 (V12.6.77) NCCL 2.23.4-1+cuda12.6	NA	NA	NA	NA	DUT
MI300-01 – MI300-04	GPU server	AMD Mi300x GPU servers	RCCL 2.21.5.60303-74~22.04 ROCm 6.3.3-74	NA	NA	NA	NA	DUT

Test Plan Goals

- Validate IPv6-only underlay, with BGP unnumbered auto-discovery peering.
- Validate EVPN-VXLAN IPv4 overlay route advertisements using IPv6 next-hops (RFC 5549).
- Validated Support for IPv4 and IPv6 RoCEv2 Flows while ensuring all features work for both IPv4 and IPv6 RoCEv2 flows.
- Validate latest mechanisms to improve performance of the AIML Cluster and monitor operations compared to previously validated congestion threshold parameters. This will be achieved by fine-tuning additional knobs supported in the 23.4R2 release:
    - Optimal Alpha-per-Queue Setting: To profile congestion levels with dynamic alpha settings at the queue level. PFC XON Limit for Ingress Ports: Setting the threshold at which the peer resumes transmitting packets after a brief pause due to PFC sent by the node.
      New CLI Stats for ECN Marked Packets: To monitor ECN-marked packets per congested queue for better congestion control.Optimal PFC Watchdog Parameters: To detect and mitigate PFC pause storms and ensure recovery from congestion scenarios.
- Validate latest DLB (dynamic Load Balancing) optimizations for RoCEv2 traffic:
    - DLB HashBucketSize: To enable better distribution of flows across ECMP links, preventing congestion on loaded link.
    - Selective DLB for DSCP Flows: To prioritize certain RoCEv2 flows based on RDMA opcode using Firewall Filters and egress-quantization para|meters.

The aim is to perform congestion scenario tests, measure thresholds, and identify the optimal values for congestion management and load balancing that will result in maximum job performance in the AIML Cluster Network.

How DCQCN (Data Center Quantized Congestion Notification) was tested:

Network Setup:

- Sender (S): RDMA-enabled host (VTEP-1).
- Leaf-1 & Spine-1/2: Part of the VXLAN EVPN fabric.
- Receiver (R): RDMA-enabled host (VTEP-2).

Step-by-Step Traffic Flow with Header Changes

Step 1: RDMA Packet from Sender (S)

- The sender S (VTEP-1) starts sending an RDMA WRITE request to Receiver R (VTEP-2).
- Since RoCEv2 runs on UDP/IP, the RDMA frame is encapsulated in UDP and Ethernet.

Table 3: Header at the Sender (S)

Header	Details
Ethernet Header	Src: S MAC, Dst: Leaf-1 MAC
IP Header (IPv4/IPv6)	Src: S IP, Dst: R IP
UDP Header (Dst Port: 4791)	UDP-based RoCEv2
RDMA Base Transport Header	—

Header	Details
RDMA WRITE Payload	—

- RoCEv2 traffic is UDP/IP-based, making it routable across Layer 3 networks.
- It reaches Leaf-1, which handles VXLAN encapsulation.

Step 2: VXLAN Encapsulation at Leaf-1

- Leaf-1 encapsulates the original RDMA packet into a VXLAN tunnel.
- The VXLAN VNI (Virtual Network Identifier) is added.
- The new destination MAC/IP is set to Spine-1 (next-hop VTEP).

Table 4: Header at Leaf-1 (Encapsulated in VXLAN)

Header	Details
Outer Ethernet Header	Src: Leaf-1 MAC, Dst: Spine-1 MAC
Outer IP Header (IPv4/IPv6)	Src: Leaf-1 IP, Dst: Leaf-2 IP
Outer UDP Header (VXLAN)	Dst Port: 4789 for VXLAN
VXLAN Header (VNI: 5001)	Identifies RDMA VXLAN segment
Inner Ethernet Header	Src: S MAC, Dst: R MAC
Inner IP Header	Src: S IP, Dst: R IP
Inner UDP Header (RoCEv2)	Dst Port: 4791
RDMA Base Transport Header	—
RDMA WRITE Payload	—

- The original RDMA packet is now inside VXLAN.
- The Leaf-1 to Spine-1 communication happens over the underlay network.

Step 3: VXLAN Packet Forwarded via Spine-1

- The VXLAN packet travels across the spine-leaf fabric.
- Spine-1 does a VXLAN lookup based on VNI 5001.
- Header remains unchanged except for minor routing updates (TTL decrement).

Step 4: VXLAN Decapsulation at Leaf-2

- Leaf-2 receives the VXLAN packet and removes the VXLAN header.
- The original RDMA frame is restored.

Table 5: Header at Leaf-2 (After VXLAN Decapsulation)

Header	Details
Ethernet Header	Src: Leaf-2 MAC, Dst: R MAC
IP Header (IPv4/IPv6)	Src: S IP, Dst: R IP
UDP Header (Dst Port: 4791)	UDP-based RoCEv2
RDMA Base Transport Header	—
RDMA WRITE Payload	—

- Leaf-2 forwards this packet to Receiver R (VTEP-2).
- The receiver processes the RDMA WRITE operation.

How DCQCN was tested:

Now let's see how DCQCN handles congestion at each level.

Scenario: Congestion at Leaf-2

1. Leaf-2 detects congestion in its buffer.
2. ECN marking is applied in the IP header of the RDMA packet.
3. Receiver R sees the ECN mark and sends a Congestion Notification Packet (CNP) back to Sender S.
4. Leaf-2 encapsulates the CNP inside VXLAN to send it back to Sender S.

- Receiver (R) sends a CNP back to Sender S to slow down transmission.
- Leaf-2 VXLAN encapsulates the CNP and sends it back to Leaf-1 via Spine-1.
- Leaf-1 VXLAN decapsulates the CNP and delivers it to Sender(S).

Table 6: CNP Header at Leaf-2 (Encapsulated in VXLAN)

Header	Details
Outer Ethernet Header	Src: Leaf-2 MAC, Dst: Spine-1 MAC
Outer IP Header (IPv4/IPv6)	Src: Leaf-2 IP, Dst: Leaf-1 IP
Outer UDP Header (VXLAN)	Dst Port: 4789 for VXLAN
VXLAN Header (VNI: 5001)	Same VNI as RDMA traffic
Inner Ethernet Header	Src: R MAC, Dst: S MAC
Inner IP Header	Src: R IP, Dst: S IP
Inner UDP Header (CNP)	Congestion Notification

The Sender S receives the CNP and reduces its transmission rate.

Table 7: Summary

Step	Action	Header Change
1	Sender (S) sends RDMA traffic	Normal RDMA over UDP/IP
2	Leaf-1 encapsulates in VXLAN	VXLAN header added
3	Spine-1 forwards VXLAN traffic	No change (only TTL decrement)
4	Leaf-2 decapsulates VXLAN	VXLAN header removed, original RDMA packet forwarded
5	Receiver detects congestion, sends CNP	Normal CNP over UDP/IP
6	Leaf-2 encapsulates CNP in VXLAN	VXLAN header added
7	Spine-1 forwards CNP VXLAN packet	No change
8	Leaf-1 decapsulates CNP, sends to Sender	VXLAN header removed, CNP delivered
9	Sender (S) slows down transmission	No header change, but rate adjusted

Table 8: Congestion Management in the Fabric

	Congestion 1 (Leaf1)	Congestion 2 (Spine)	Congestion 3 (Leaf2)
PFC	1. Leaf1 sends PFC to IXIA	1. Spine sends PFC to Leaf1 2. Leaf1 sends PFC to IXIA	1. Leaf2 sends PFC to Spine 2. Spine sends PFC to Leaf1 3. Leaf1 sends PFC to IXIA
ECN	1. Leaf1 marking to CE in inner ECN bits	1. Spine marking bits to CE in outer ECN bits 2. Leaf2 copies ECN bits from outer to inner header	1. Leaf2 marking to CE in inner ECN bits

Table 9: Show Commands

Event	Monitor CLI Command	Indicator	Action/ Expected Outcome
Validate Fabric Interfaces	`show interfaces <int>`	Interface is up and operational	Ensures the fabric interfaces are active.
Validate Router Advertisements	`show ipv6 router-advertisement`	RAs are sent and received on fabric links	• Confirms IPv6 neighbor discovery is functional. • Confirms Leaf node advertising IPv6 prefix to GPU servers for SLAAC. • Confirms Leaf and Spine node are advertising each other's link local address for BGP autodiscovery.

Event	Monitor CLI Command	Indicator	Action/ Expected Outcome
Verify IPv6 Neighbor Discovery	`show ipv6 neighbors`	All neighbors are in "reachable" state	• Confirms that IPv6 neighbor devices are properly discovered. • On the spine nodes, the IPv6 leaf nodes link local addresses must be present with their corresponding MAC addresses. • On the leaf nodes, the IPv6 link-local addresses of the leaf nodes with their corresponding MAC addresses must be present. • Also, on the leaf nodes, the link-local and global (autoconfigured) IPv6 addresses of the GPU servers interface's and IXIA's interfaces must be present.
Verify Link-Local Connectivity	`ping fe80::<peer-addr> interface <int>`	Successful ICMP response	Confirms link-local IPv6 address connectivity between GPU servers and Leaf nodes, and between Leaf and Spine nodes.
Verify global IPv6 address connectivity	`ping <prefix>::<peer-addr> <tenant-vrf-name>`	Successful ICMP response	Confirms global IPv6 address connectivity between GPU servers and Leaf nodes.
Verify BGP Peering	`show bgp summary`	All BGP sessions in "Established" state	Ensures successful establishment of underlay auto-discovered BGP peering.- underlay. Ensures successful establishment of overaly BGP peering.

Verify BGP Route Exchange	`show route protocolbgp`
	
	`show route advertising-protocol BGP table <Tenant>`
	
	`show route receive-protocol BGP table <Tenant>`
		Loopback routes of peers are visible	• Confirms advertisement of spine and leaf loopback IPv4/IPv6 addresses. Routes should be placed in inet.0/inet6.0.
			• Confirms advertisement of IPv6 addresses of links between GPU servers and Leaf node. Routes should be placed in the corresponding Tenant VRF.

Verify ECMP Load Balancing	`show route <dest-ipv6> detail`	Multiple next-hops present	Confirms equal-cost multipath (ECMP) forwarding.
Verify Fabric Forwarding	`traceroute no-resolve <dest-ipv6>`	Expected underlay path is followed	Ensures proper data plane connectivity.
ECN marked packets	`show interfaces <int> extensive`	Review "Output Errors" and "ECN Marked Packets"	This indicates that the buffer is utilized to the fill level value, and the switch requests senders to reduce the transmit rate. If this prematurely impacts application performance, tune as indicated in the Action field. Increase the drop profile fill level value until PFC occurs. Then, reduce by decrements of 5 until PFC stops. ECN marking occurs later in the buffer utilization, reducing the frequency of traffic throttling.

Event	Monitor CLI Command	Indicator	Action/ Expected Outcome
PFC Pause Frames	`show interfaces <int> extensive`	Review “Priority Flow Control Statistics”	Indicates traffic is coming in at a rate greater than the shared input buffer. Reduce drop profile fill level value until PFC no longer occurs. ECN occurs earlier in buffer. utilization to mitigate the Pause Frame and interruption of traffic transmission.
Tail Drops & Egress queue peak buffer occupancy	`show interfaces <int> extensive`	Review “Egress Queues” and “Dropped Packets”	Indicates packets are being dropped based on WRED profile. Reduce the drop profile fill level values until ECN occurs before drops/PFC. Traffic reduction should happen earlier, allowing queues to clear without drops.

Input drops & Ingress priority-group buffer occupancy	`show interfaces <int> extensive`
	`show interfaces queue <int>`
	`show interfaces queue buffer-occupancy <int>`
		Review “Drops” and “Resource Errors”	The ingress shared buffer is being exceeded and unable to store the incoming packets. This event is common when PFCs are generated. Increase shared buffer > ingress > buffer-partition <%>. For "Resource Errors", increase the "ingress buffer-partition lossless-headroom" percentage. If the ingress shared buffer partition is at the expected value, reduce the fill level values until ECN occurs before drops/PFC. More memory is allocated to the specified buffer if needed. No or minimal input drops were expected.

SLAAC	`show ipv6 router-advertisement`
	
	`ifconfig gpu_eth# (server)`
	
	`ip -6 route <prefix>`
	`ip -6 route <rio-prefix>`
		Check “prefix”, “interface”, “rio-prefix”, “timers”.	Indicate SLAAC was configured correctly, and the server is autconfiguring its IPv6 addresses and installing proper routing information.

Performance Data for EVPN Multi-Tenancy Server Isolation with NVIDIA H100 GPUs

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

- 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: 25% (200G), 75%(600G) & 100%(800G) traffic load sent to all the 4 leaf's
- 16 QPs per port from IXIA + Model to the Leafs as ingress traffic
- Only v4 flows with single DSCP code-point marking

Test Results

Table 1.1: DLRM Model Job Completion Time (JCT) Test Results

Traffic Profile	Tuned parameters in leaf/spine (all other parameters as per defaults above)	JCT [sec]
Only Model	DLB disabled (ECMP LB)	3.23
Only Model	flowset-table-size 2048 inactivity-interval 256 sampling rate 1000000 flowlet rate-weightage 80 flowlet egress-quantization min 10 flowlet egress-quantization min 90	2.49
Only Model	flowset-table-size 2048 inactivity-interval 128 sampling rate 1000000 flowlet rate-weightage 80 flowlet egress-quantization min 10 flowlet egress-quantization min 90	2.47
Model + Ixia (100% traffic)	flowset-table-size 2048 inactivity-interval 128 sampling rate 1000000	2.66
Model + Ixia (100% traffic)	flowset-table-size 2048 inactivity-interval 128 sampling rate 1000000 flowlet rate-weightage 80 flowlet egress-quantization min 10 flowlet egress-quantization min 90	3.16
Model + Ixia (100% traffic)	flowset-table-size 2048 inactivity-interval 256 sampling rate 1000000	2.75

Table 1.2: BERT Model Job Completion Time (JCT) Test Results

Traffic Profile	Tuned parameters in leaf/spine (all other parameters as per defaults above)	JCT [sec]
Only Model	flowset-table-size 2048 inactivity-interval 128 sampling rate 1000000	1.98
Model + Ixia (100% traffic)	flowset-table-size 2048 inactivity-interval 128 sampling rate 1000000	1.86
Model + Ixia (100% traffic)	flowset-table-size 2048 inactivity-interval 128 sampling rate 1000000 flowlet egress-quantization min 10 flowlet egress-quantization min 90	1.97

Table 1.3: NCCL test Average Bus Bandwidth

Traffic Profile	Tuned parameters in leaf/ (all other parameters as per defaults above)	Ave. Bus Bandwidth (GB/s)
Only Model (NCCL All-reduce 4QPs with msg-size 16G)	flowset-table-size 2048 inactivity-timer 128 sampling rate 1000000	390.9
Only Model (NCCL All-reduce 4QPs with msg-size 1G)	flowset-table-size 2048 inactivity-timer 128 sampling rate 1000000	375.1
Only Model (NCCL All-reduce 4QPs with msg-size 8G)	flowset-table-size 2048 inactivity-timer 128 sampling rate 1000000	381.5
Only Model (NCCL All-reduce 4QPs with msg-size 16G)	flowset-table-size 2048 inactivity-timer 128 sampling rate 1000000 reassignment threshold 3 reassignment quality-delta 6	390.8
Model (NCCL All-reduce 4QPs) + ixia traffic 100% load	flowset-table-size 2048 sampling rate 1000000 inactivity-timer 128	389.9
Model (NCCL All-reduce 4QPs) + ixia traffic 100% load	flowset-table-size 2048 sampling rate 1000000 inactivity-timer 128 flowlet rate-weightage 80 flowlet egress-quantization min 10 flowlet egress-quantization min 90	390.58

Model (NCCL All-reduce 4QPs) + ixia traffic 100% load	flowset-table-size 2048 sampling rate 1000000 flowlet egress-quantization min 10 flowlet egress-quantization min 90 reassignment threshold 3 reassignment quality-delta 6	390.6
Only Model (NCCL alltoall 4QPs)	flowset-table-size 2048 sampling rate 1000000 inactivity-timer 128 flowlet egress-quantization min 10 flowlet egress-quantization min 90	42.1
Only Model (NCCL alltoall 4QPs) + ixia 100% traffic load	flowset-table-size 2048 sampling rate 1000000 inactivity-timer 128 flowlet egress-quantization min 10 flowlet egress-quantization min 90	41.15
Only Model (NCCL alltoall 4QPs) + ixia 100% traffic load	DLB disabled (ECMP)	37.75

Table 1.4: LLAMA2 JCT

Traffic Profile	**Tuned parameters in leaf/spine (all other parameters as per defaults above)**	**JCT [min]**
Only Model	flowset-table-size 2048 inactivity-interval 128 sampling rate 1000000 flowlet rate-weightage 80 flowlet egress-quantization min 10 flowlet egress-quantization min 90	8.15
Only Model	flowset-table-size 2048 inactivity-interval 256 sampling rate 1000000 flowlet rate-weightage 80 flowlet egress-quantization min 10 flowlet egress-quantization min 90	9.89

Only Model + ixia 100% traffic load	flowset-table-size 2048 inactivity-interval 128 sampling rate 1000000 flowlet rate-weightage 80 flowlet egress-quantization min 10 flowlet egress-quantization min 90	9.55

Table 1.5: LLAMA3 JCT

Traffic Profile	Tuned parameters in leaf/spine (all other parameters as per defaults above)	JCT [min]
Only Model	flowset-table-size 2048 inactivity-interval 128 sampling rate 1000000 flowlet rate-weightage 80 flowlet egress-quantization min 10 flowlet egress-quantization min 90	26.68
Only Model	flowset-table-size 2048 inactivity-interval 256 sampling rate 1000000 flowlet rate-weightage 80 flowlet egress-quantization min 10 flowlet egress-quantization min 90	26.69
Only Model + ixia 100% traffic load	flowset-table-size 2048 inactivity-interval 128 sampling rate 1000000 flowlet rate-weightage 80 flowlet egress-quantization min 10 flowlet egress-quantization min 90	24.69
Only Model + ixia 100% traffic load	flowset-table-size 2048 inactivity-interval 256 sampling rate 1000000 flowlet rate-weightage 80 flowlet egress-quantization min 10 flowlet egress-quantization min 90	26.7

Performance Data for EVPN Multi-Tenancy Server Isolation with AMD MI-300 GPUs

Table 2.1: LLAMA3 JCT

Traffic Profile	Tuned parameters in leaf/spine (all other parameters as per defaults above)	JCT [min]
Only Model	flowset-table-size 2048 inactivity-interval 128 sampling rate 1000000 flowlet rate-weightage 80 flowlet egress-quantization min 10 flowlet egress-quantization min 90	32
Only Model	flowset-table-size 2048 inactivity-interval 256 sampling rate 1000000 flowlet rate-weightage 80 flowlet egress-quantization min 10 flowlet egress-quantization min 90	29.73
Only Model	flowset-table-size 2048 inactivity-interval 256 sampling rate 1000000 reassignment threshold 3 reassignment quality-delta 6 flowlet egress-quantization min 10 flowlet egress-quantization min 90	32.1
Only Model	flowset-table-size 2048 inactivity-interval 256 sampling rate 1000000 flowlet rate-weightage 50 flowlet egress-quantization min 10 flowlet egress-quantization min 90	27.59
Only Model + ixia 100% traffic load	flowset-table-size 2048 inactivity-interval 256 sampling rate 1000000 flowlet rate-weightage 80 flowlet egress-quantization min 10 flowlet egress-quantization min 90	29.76

Table 2.2: RCCL tests Average Bandwidth

Traffic Profile	Tuned parameters in leaf/spine (all other parameters as per defaults above)	Bandwidth (GB/s)
Only Model (RCCL All-reduce 32QPs)	flowset-table-size 2048 sampling rate 1000000 inactivity-timer 128 flowlet egress-quantization min 10 flowlet egress-quantization min 90	336.7
Only Model (RCCL All-reduce 32QPs)	flowset-table-size 2048 sampling rate 1000000 inactivity-timer 256 flowlet egress-quantization min 10 flowlet egress-quantization min 90	335.2
Model (RCCL All-reduce 32QPs) + ixia traffic 100% load	flowset-table-size 2048 sampling rate 1000000 inactivity-timer 128 flowlet egress-quantization min 10 flowlet egress-quantization min 90 rate-weightage 80	338.3
Model (RCCL All-reduce 32QPs) + ixia traffic 100% load	Default DLB parameters	330.3
Model (RCCL All-reduce 32QPs) + ixia traffic 100% load	DLB disabled (ECMP)	335.8

Performance Data for EVPN Multi-Tenancy GPU Isolation with NVIDIA H100 GPUs with IPv6 SLAAC

Test Environment

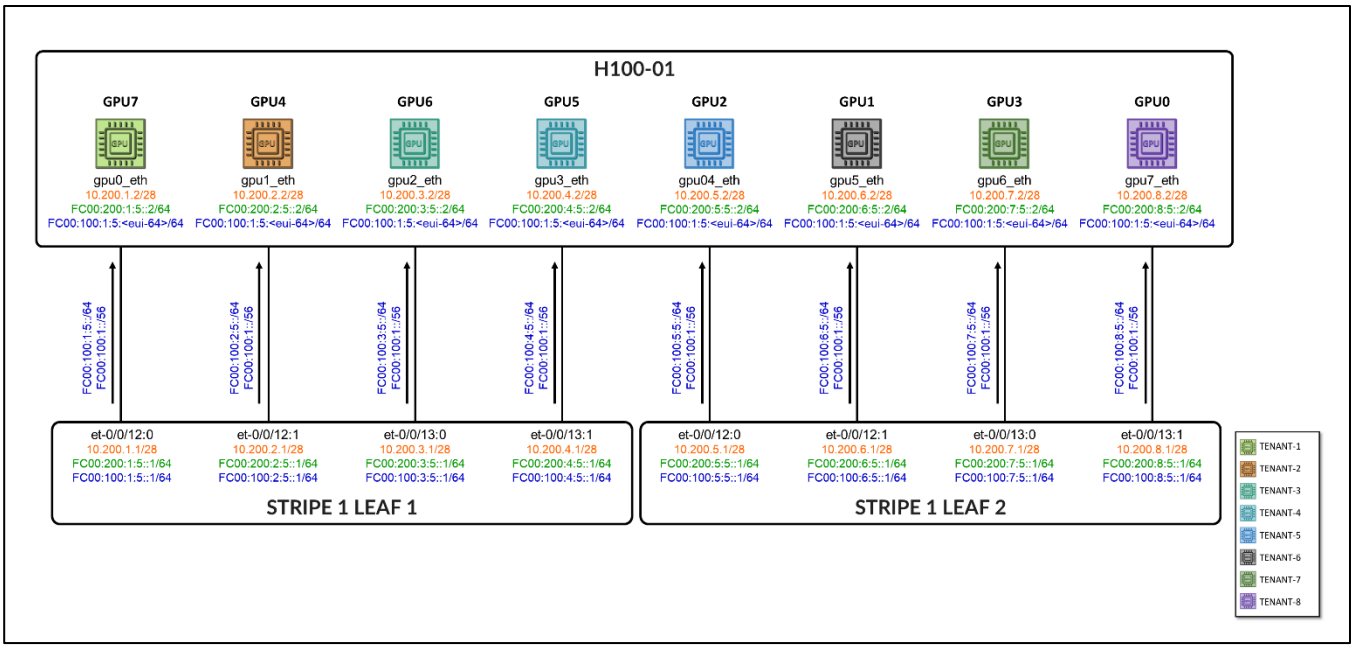
Buffer management values (default):

- 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
- flowlet egress-quantization min 20

- flowlet egress-quantization max 50
- flowlet egress-quantization rate-weightage 50
- flowlet reassignment disabled



NOTE: All Other servers are connected the same way.

Table 3.1: NCCL Bandwidth (Tested with 4 spines for 3.2T fabric Bandwidth)

Traffic Profile	Tuned parameters in leaf/spine (all other parameters as per defaults above)	Bandwidth x Tenant (GB/s)
Only Model (NCCL All-reduce 8QPs)	DLB disabled	48.5
Only Model (NCCL All-reduce 8QPs)	flowset-table-size 2048 inactivity-timer 128 sampling rate 1000000	48.53
Model (NCCL All-reduce 8QPs)	flowset-table-size 2048 sampling rate 1000000 inactivity-timer 128	48.52
Model (NCCL All-reduce 8QPs) + ixia traffic 100% load	flowset-table-size 2048 sampling rate 1000000	48.55

Traffic Profile	Tuned parameters in leaf/spine (all other parameters as per defaults above)	Bandwidth x Tenant (GB/s)
	inactivity-timer 128	

Reference Documents:

- https://www.juniper.net/documentation/us/en/software/junos/ai-ml-evo/index.html
- https://www.juniper.net/documentation/us/en/software/nce/congestion-control-ai-ml/congestion-control-ai-ml.pdf
- https://www.juniper.net/documentation/us/en/software/nce/ai-clusters-data-center-design/ai-clusters-data-center-design.pdf
- https://www.juniper.net/content/dam/www/assets/white-papers/us/en/networking-the-ai-data-center.pdf
- https://www.juniper.net/documentation/us/en/software/nce/nce-225-bgp-unnumbered/nce-225-bgp-unnumbered.pdf

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