NCCL Test Report for AIDC Scenarios

Asterfusion Data Technologies Co., Ltd.

Content

1 Summary

To meet the growing bandwidth demands of large-scale model inference and training, Asterfusion conducted in-depth NCCL performance validation tailored for AIDC scenarios, based on the CX864E-N switch. The tests were performed using a mainstream Spine-Leaf network topology and leveraged the Keysight (IXIA) AresONE-S 400GE testing platform to build a highly realistic simulation environment with authentic AI workloads and RDMA communication characteristics, fully reproducing distributed GPU communication scenarios.

Test Highlights

1. Spine-Leaf end-to-end latency as low as 2 μs

The Asterfusion AI switch demonstrates high-performance, low-latency characteristics under the RoCEv2 + DCQCN architecture. In testing, end-to-end latency as low as 2 μs was observed within the Spine-Leaf network, fully meeting the UEC requirement of 2–10 μs, and proving suitable for large-scale AI training and inference clusters.

2. INT-Driven Adaptive Routing (IAR) bandwidth utilization reaches 97%

Within a typical Spine-Leaf architecture, the Asterfusion CX864E-N switch was tested by comparing three types of collective communication traffic scheduling modes in a Spine-Leaf network: ECMP 5-tuple hash[1]，ECMP RDMA QP hash, and INT-Driven Adaptive Routing. Under the INT-Driven Adaptive Routing mode, bandwidth utilization reached 97%, fully meeting the 85% target specified by UEC.

[1] With E-ECMP, link utilization varies largely: 40-90% of max bandwidth

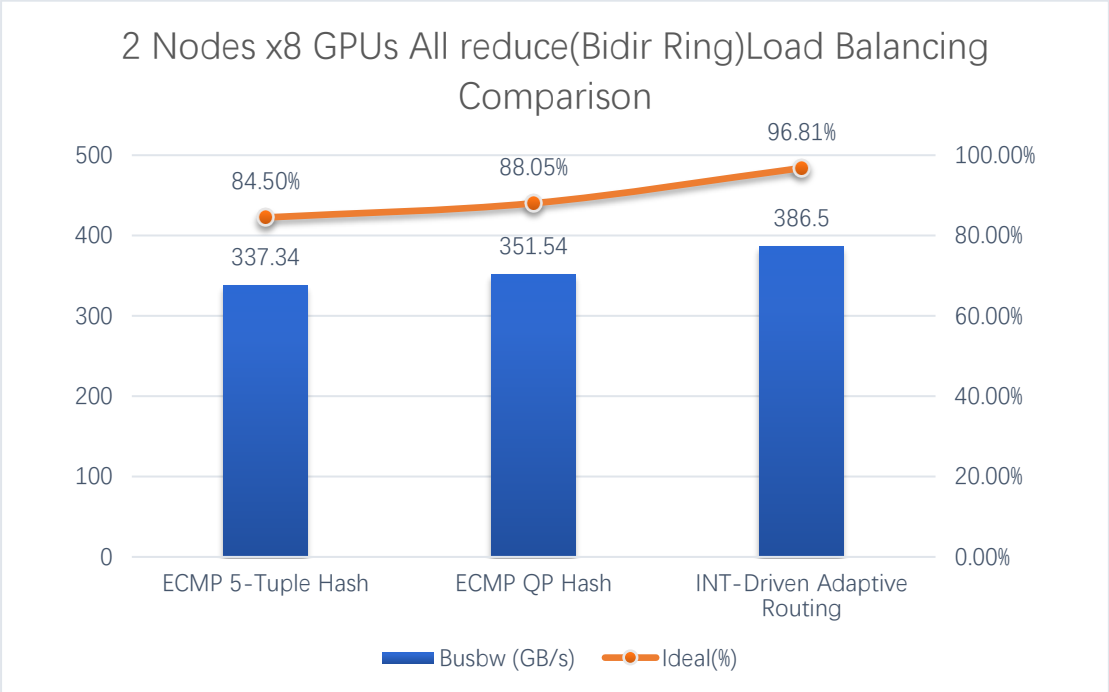


Figure 1-1 Load Balancing Result Comparison

3. Tail Latency P95 FCT（Flow Completion Time） reduced by 11.13%

A comprehensive Tail Latency test was also conducted on the switch. The following are the P95 FCT tail latency results of the switch, comparing Spine-Leaf network testing and single-switch benchmark testing. In scenarios with multi-port concurrency and burst traffic, tests were performed using flows of the same size (flow size = 8 GB). It can be seen that after enabling the INT-Driven Adaptive Routing feature, the P95 FCT tail latency values decreased by 11.13% compared with ECMP 5-tuple hashing and by 6.70% compared with ECMP RDMA QP hashing. The tail latency under IAR was basically consistent with the single-switch benchmark, demonstrating excellent congestion management and forwarding capabilities.

The convergence of tail latency directly improved the overall efficiency of the training job cycle. For example, in a typical LLaMA-39B[2] model training task where

[2] For example, with the Llama-39B model, TP- and CP-incurred communication occupies 55% of the execution time;

communication accounts for about 55% of the total time, compared with traditional ECMP 5-tuple hashing, the optimized tail latency performance reduced job completion time (JCT) by approximately 6.12%, showing significant benefits in cluster communication scheduling optimization.

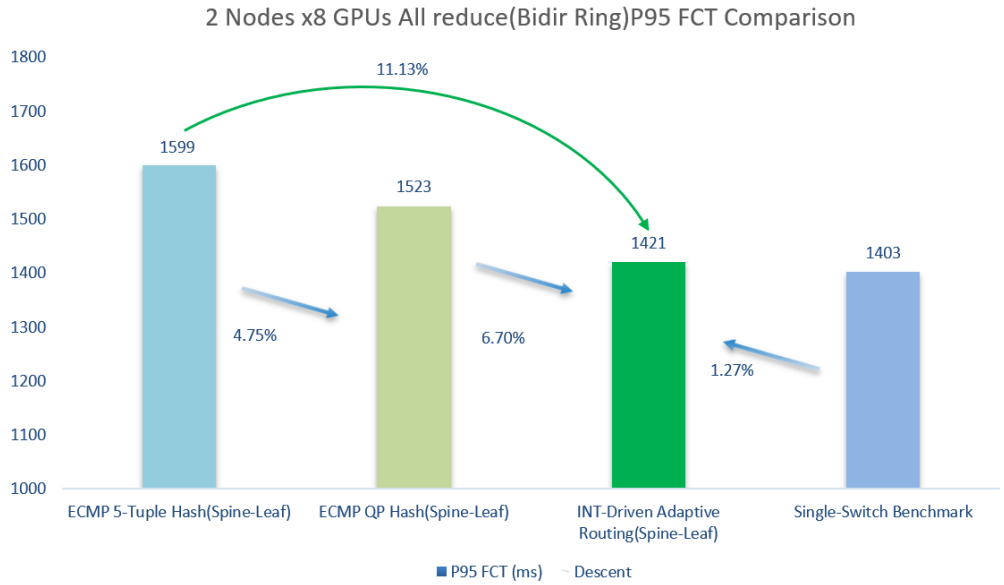


Figure 1-2 P95 FCT Comparison

4. Intelligent Flow Control Optimization (ECN + PFC)

By employing a tunable ECN (Explicit Congestion Notification) algorithm and dynamic queue threshold configuration, fine-grained congestion feedback control is achieved, ensuring low packet loss and high throughput. With optimized PFC (Priority Flow Control) and ECN parameters, queue latency and retransmission rates are significantly reduced, achieving an optimal balance between bandwidth utilization and network stability.

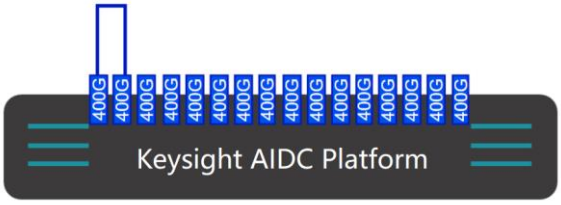
5. Strong Topology Adaptability

In Spine-Leaf tests across various GPU cluster architectures—such as 2 nodes × 8 GPUs, 4 nodes × 4 GPUs, and 8 nodes × 2 GPUs—collective communication algorithms like All-Reduce, All-Gather, and All-to-All consistently achieved around 97% bandwidth utilization. This demonstrates the switch's excellent traffic scheduling capabilities in large-scale AI clusters.

1.1 AIDC Test Topology for AIDC Scenarios

- **Back-to-back Benchmark Test Network Topology**

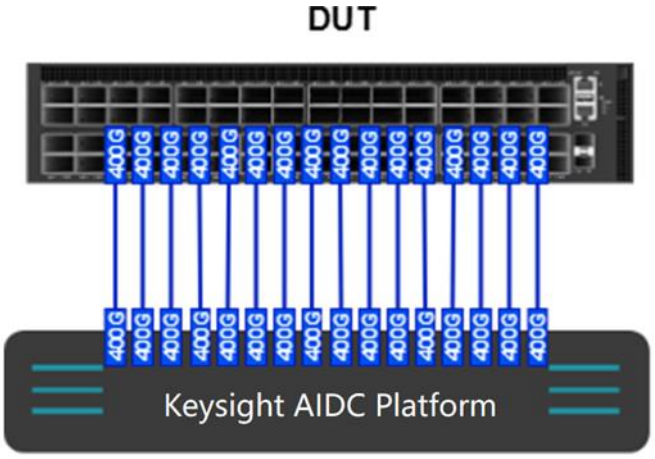
The purpose of this topology is to establish a performance benchmark, serving as a reference for comparison in subsequent multi-node complex scenarios. Due to the limitations of the tester's own mechanism, the back-to-back loopback test can only be conducted on 2 nodes × 1 GPU.



Picture 1-1 Back-to-Back Benchmark Topology

- **Single Switch Benchmark Test Network Topology**

This topology is designed to validate the core performance and flow control behavior of a single switch under AI communication workloads, ensuring that the switch does not become a performance bottleneck in GPU communication.

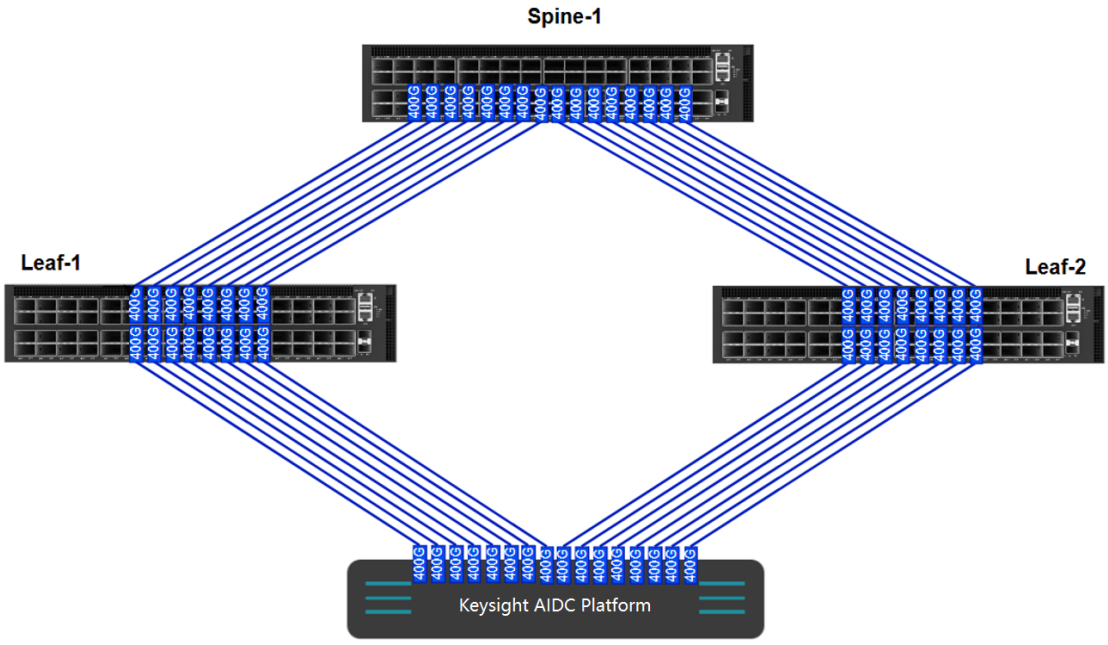


Picture 1-2 Single-Switch Benchmark Topology

- **Spine-Leaf Test Network Topology**

This topology is primarily designed to simulate multi-node communication traffic patterns in an AI Data Center (AIDC) under a Spine-Leaf architecture. It is used to evaluate the stability and performance bottlenecks of the network in complex scenarios such as cross-switch path scheduling, multi-path load balancing,

congestion control, and failover handling.



Picture 1-3 Spine-Leaf Topology

- **Test Platform 1: Asterfusion CX864E-N AI Switch**

This test was conducted using Asterfusion's self-developed high-performance AI networking platform — the CX864E-N ultra-low-latency switch. Based on a high-speed RoCEv2 communication architecture, it is optimized for large-scale GPU interconnect and offers:

- Nanosecond-level latency control, with single-switch latency as low as 560 ns
- Support for large-scale ECMP / INT-Driven Adaptive Routing traffic scheduling strategies (256-way, 4K groups)
- Full support for PFC and ECN, with DCQCN to enable end-to-end flow control
- Line-rate forwarding capabilities for 64 × 800GE, or 128 × 400GE, or 512 × 100GE

Networks built with CX864E-N in Spine-Leaf topologies demonstrate outstanding performance in key metrics such as end-to-end latency, network utilization, and high reliability. This provides a strong technical reference for the design and deployment of next-generation AI data center networks.

- **Test Platform 2: IXIA AresONE-S 400GE**

The test utilized the industry-leading high-performance traffic generation platform — the IXIA AresONE-S 400GE test system, featuring:

- Up to 16 × 400G QSFP-DD ports, fully covering real-world GPU/NIC communication rates
- Accurate emulation of RDMA/RoCEv2 traffic patterns, with fine-grained rank-to-rank traffic simulation
- Protocol support for PFC, ECN, DCQCN, combined with path-level traffic simulation to accurately reflect NCCL bandwidth utilization and congestion control behavior
- Nanosecond-resolution timestamping for precise measurement of one-way delay, round-trip time (RTT), and jitter

As a benchmark platform for AI network validation, AresONE-S effectively supports full-path performance verification and comprehensive network optimization.

2 Test Environment

2.1 Hardware Environment

Name	Model	Description	Quantity	Remark
Switch	Asterfusion CX864E-N	64*800GE OSFP Ethernet optical ports	3	2 as Leaf, 1 as Spine
Tester	IXIA Keysight AresONE-S 400GE	Number of network interfaces: 16*400G	1	/
800G Optical Transceiver	Asterfusion OT-800G-OSFP-2VR4	800G-VR8/2*VR4, OSFP, Dual MPO12/APC, MMF 850nm, 50m/OM4, 13.5W, COB craft	28	20 for Spine-Leaf interconnection, 8 for tester connection
400G Optical Transceiver	Asterfusion OT-400G-QDD-VR4	400G, QSFP56-DD, VR4, MPO-12, 850nm MMF, 50m/OM4, 8W	16	For tester connection
Optical Fiber	MPO-MPO-12S-OM3-APC-030-L	MPO(Female) to MPO(Female), 12 Strands, OM3-300, APC head, Length 3m	36	20 for switch interconnection, 16 for tester connection

Table 2-1 Hardware Table

2.2 Physical Environment

Picture 2-1 Physical Wiring Environment

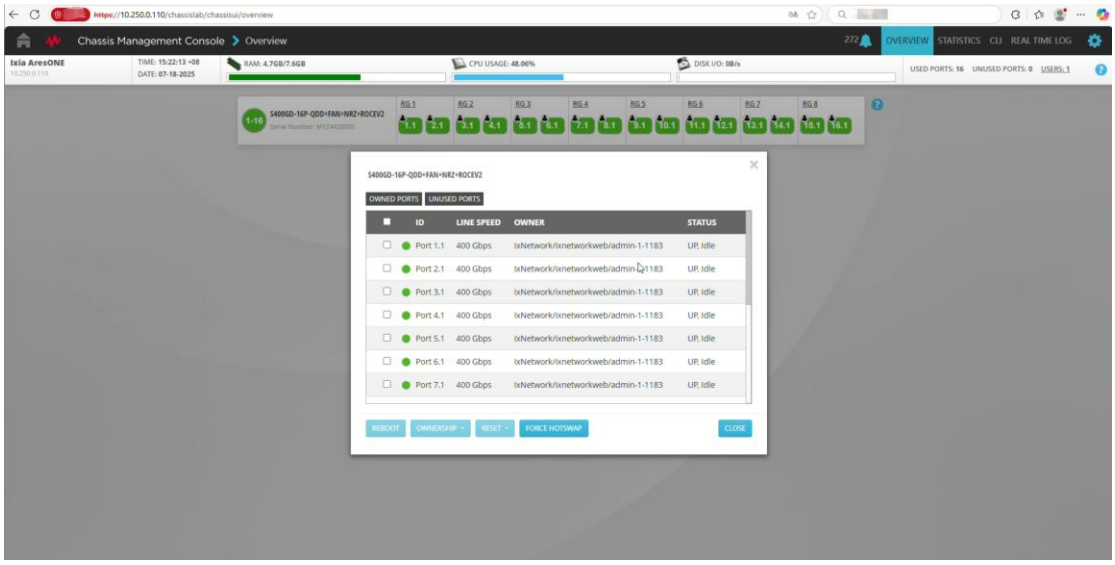
The current view shows the actual wiring diagram of the data center, arranged from top to bottom as the test instrument, Leaf1, Leaf2, and Spine.

2.3 Test Environment Interface

- **Tester Chassis Management Console**

  The chassis management console is the central platform for managing and monitoring the tester chassis (such as the AresONE-S 400GE), including hardware status, port assignments, power supply, fan operations, and other system resources. It is typically accessed via a Web GUI or CLI (SSH).

  The following image displays the physical port information of the test instrument.



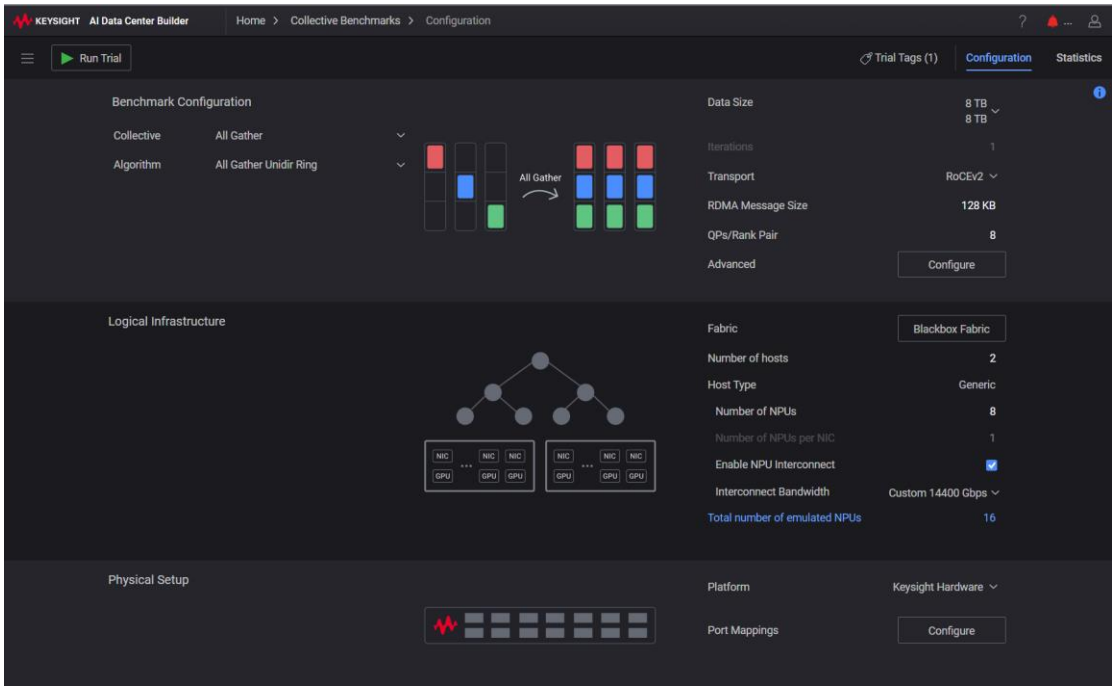
Picture 2-2 Test Platform Chassis Management Console

- **AI Data Center Builder**

AI Data Center Builder is a comprehensive pre-configured testing platform solution launched by Keysight (IXIA) for AI infrastructure testing. It is designed to help users quickly deploy, validate, and optimize AI cluster networks, and is particularly suited for high-performance computing (HPC), GPU/NPU integration, and RoCEv2 communication environments. As an integrated solution, it includes the following components:

- High-density test hardware platforms (e.g., AresONE-S 400GE)
- Traffic generation and control systems (e.g., IxNetwork, IxExplorer)
- Specialized RoCEv2/NCCL test scripts (All Reduce, All to All, etc.)
- Data center topology templates (Leaf-Spine, multi-tenant architectures)
- QoS, PFC/ECN tuning tools
- Visualization reports and path analysis tools

The figure below shows the NCCL configuration page.



Picture 2-3 AI Data Center Builder

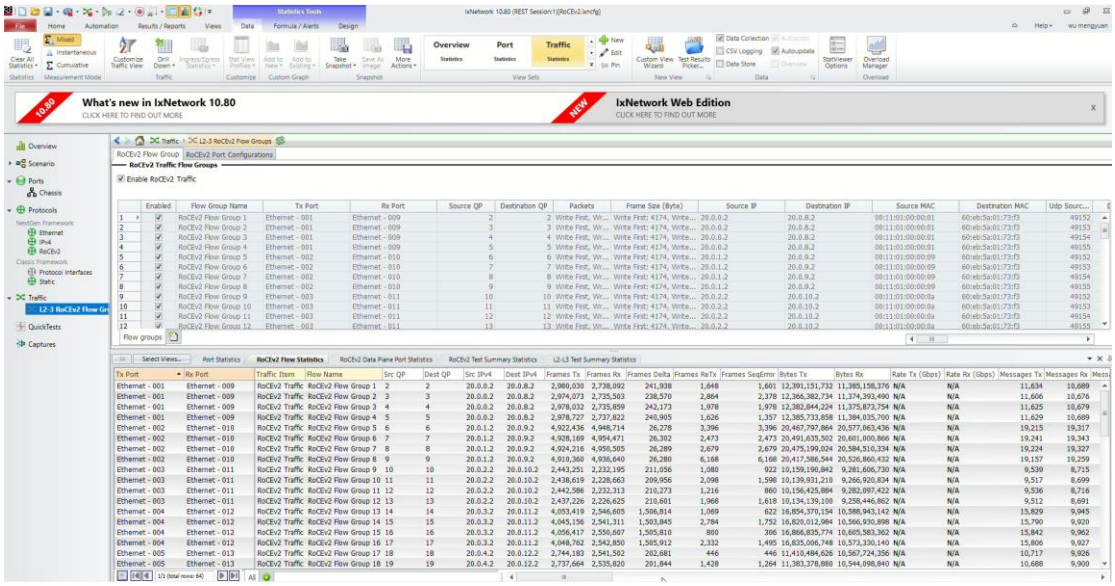
- **IxNetwork**

IxNetwork, as a core platform for protocol and performance testing, supports the creation of complex network topologies and the emulation of protocols such as BGP, ISIS, EVPN, and VXLAN. It can simulate thousands of ports, hosts, and virtual devices, enabling the construction of large-scale AI/NPU network environments.

Typical use cases include:

- Convergence testing of Leaf-Spine networks
- Performance analysis of ECN and PFC mechanisms in collective communication scenarios
- Comparison of bandwidth utilization and congestion control effectiveness

The image below shows real-time RoCEv2 traffic flow information.



Picture 2-4 IxNetwork

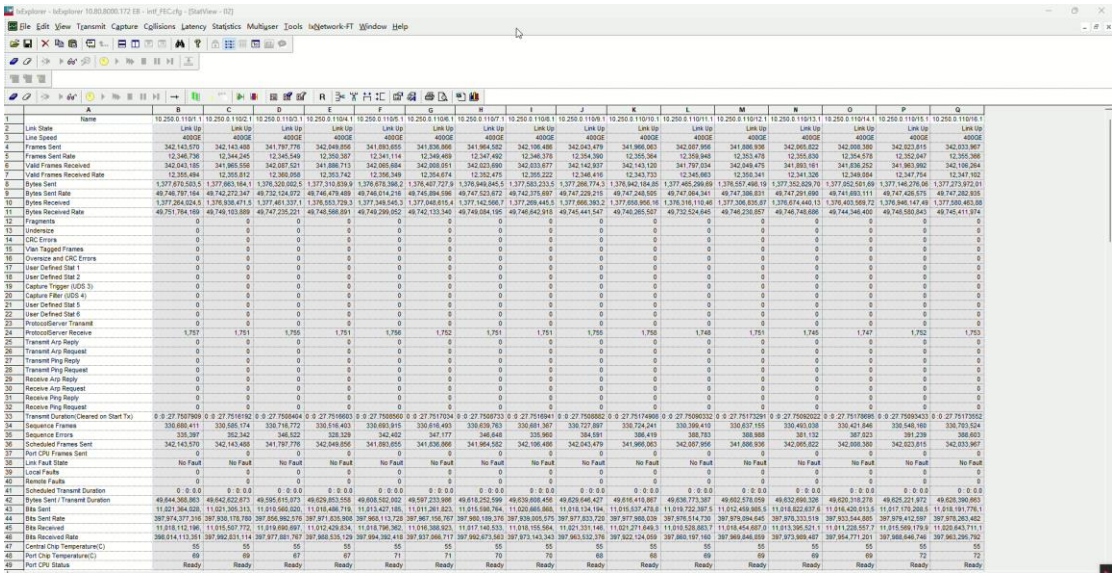
- **IxExplorer**

IxExplorer serves as a port-level control and debugging tool, similar to a network interface control panel. It is used to configure test port attributes, manually send packets, and generate Pause frame traffic. Users can configure port MAC/IP addresses, transmission rates, frame formats, etc., to test link responsiveness and device loopback.

Common use cases include:

- Verifying whether RoCE flow control mechanisms are triggered
- Tuning PFC and ECN thresholds
- Simulating RDMA communication paths and sending tagged traffic

The figure below displays real-time port attribute information from the test instrument.

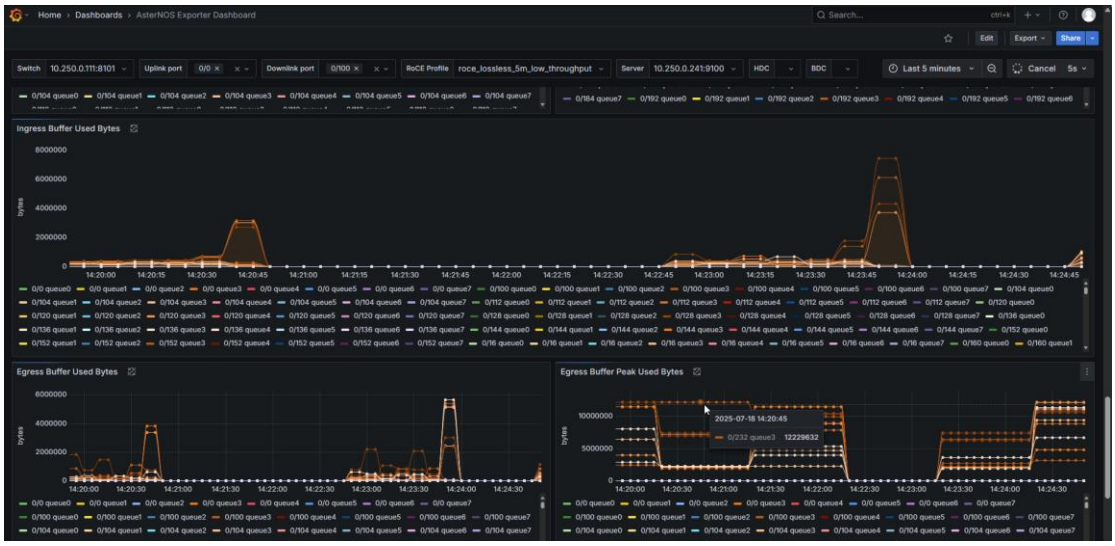


Picture 2-5 IxExplorer

- **Grafana**

Grafana, in combination with the AsterNOS Prometheus Node Exporter, provides real-time visualization of Asterfusion switch monitoring metrics (such as buffer utilization and ECN/PFC trigger events). This enables AI data centers to quickly identify issues like network congestion, packet loss, or uneven path utilization, thereby enhancing the observability and operational efficiency of RDMA networks.

The diagram below shows the real-time buffer metrics of the switch test interface as monitored by Grafana.

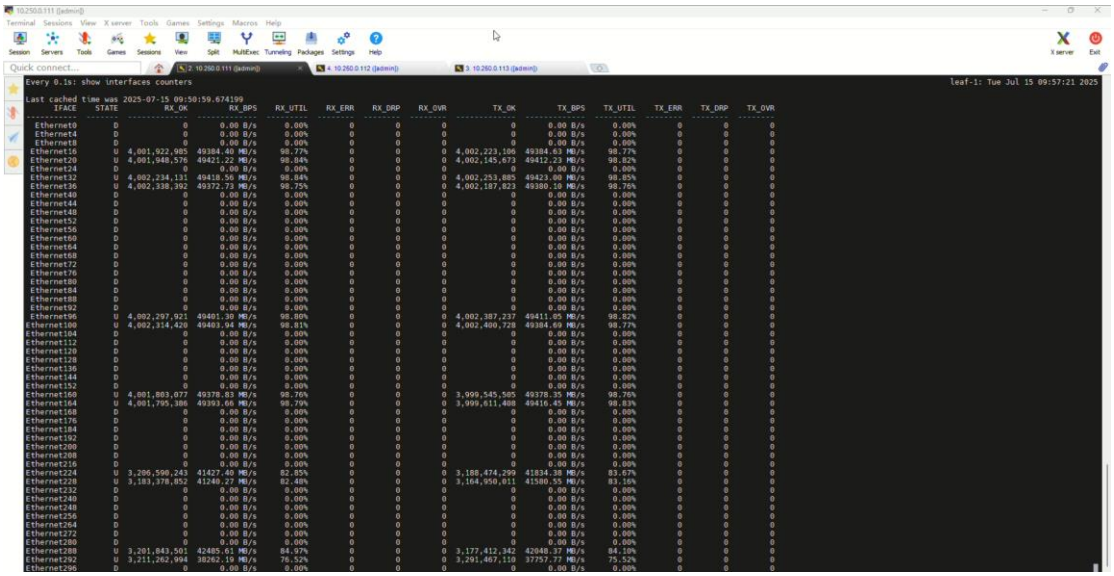


Picture 2-6 Grafana

- **Switch Command Line Interface (CLI)**

The switch CLI is the primary interface for network administrators to perform device configuration, status monitoring, troubleshooting, and performance tuning. The following query commands can be used to monitor real-time packet transmission and reception statistics on interfaces.

The image below shows the real-time interface traffic statistics page on the switch.



Picture 2-7 Switch CLI

3 Test Results Overview

3.1 Back-to-Back Benchmark Test

The following results are from 2 nodes × 1-GPU back-to-back test. The Type in the table below refers to the collective communication primitive, and Algorithm refers to the algorithm used for collective communication. Busbw refers to the bus bandwidth — the bandwidth of data sent and received over the Ethernet interface — used to evaluate the data throughput pressure on the network interface. Ideal refers to the ideal bandwidth utilization, which measures the actual network transmission efficiency. It is calculated as the ratio of bus bandwidth to the theoretical maximum throughput (expressed as a percentage).

It can be observed that the ideal bandwidth utilization in the back-to-back test

exceeds 97%.

Type	Algorithm	Data Size (Byte)	Busbw (GB/s)	Ideal (%)
All-Reduce	Bidir Ring	2G	47.84	97.14
All-Reduce	Bidir Ring	4G	48.12	97.70
All-Reduce	Bidir Ring	8G	48.26	98.00
All-Reduce	Bidir Ring	16G	48.35	98.18

Table 3-1 2 Nodes x 1GPU Back-to-Back Benchmark

Index	Data Size	Collective	Completion Time (ms)	Algbw (GB/s)	Busbw (GB/s)	Ideal (%)	PFC Rx	ECN CE Rx	Frames ReTx	Min FCT (ms)	P50 FCT (ms)
1	2,147,483,648	ALL_REDUCE-1	44.89	47.84	47.84	97.14	0	0	0	11.083	11.085
2	4,294,967,296	ALL_REDUCE-2	89.26	48.12	48.12	97.70	0	0	0	22.167	22.168
3	8,589,934,592	ALL_REDUCE-3	177.99	48.26	48.26	98.00	0	0	0	44.337	44.340
4	17,179,869,184	ALL_REDUCE-4	355.31	48.35	48.35	98.18	0	0	0	88.667	88.672

Picture 3-1 2 Nodes x 1GPU Back-to-Back Benchmark Test Summary

3.2 Single-Switch Benchmark Test

The following are results from a single-switch benchmark test.

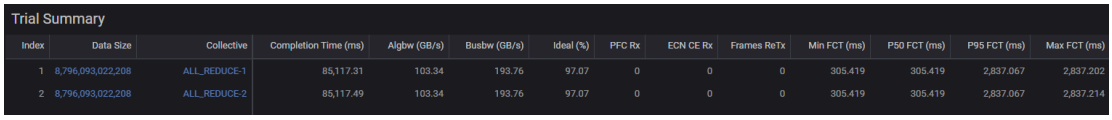
- **2 Nodes × 8 GPUs**

Type	Algorithm	Busbw (GB/s)	Ideal (%)
All-Reduce	Unidir Ring	193.76	97.07
All-Reduce	Bidir Ring	391.75	98.12
All-Reduce	Halving Doubling	88.71	97.72
All-to-All	PXN	91.52	98.16

Table 3-2 2 Nodes x 8 GPUs Single-Switch Benchmark

The original test report image is shown below

All-Reduce Unidir Ring



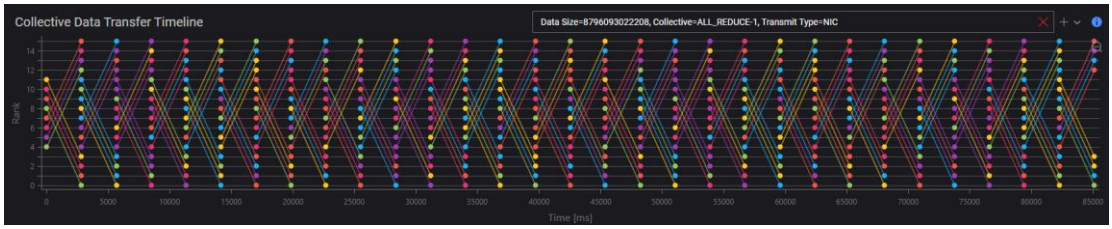
Trial Summary

Index	Data Size	Collective	Completion Time (ms)	Algbw (GB/s)	Busbw (GB/s)	Ideal (%)	PFC Rx	ECN CE Rx	Frames ReTx	Min FCT (ms)	P50 FCT (ms)	P95 FCT (ms)	Max FCT (ms)
1	8,796,093,022,208	ALL_REDUCE-1	85,117.31	103.34	193.76	97.07	0	0	0	305.419	305.419	2,837.067	2,837.202
2	8,796,093,022,208	ALL_REDUCE-2	85,117.49	103.34	193.76	97.07	0	0	0	305.419	305.419	2,837.067	2,837.214

Picture 3-2 2 Nodes x 8 GPUs Single-Switch All-Reduce Unidir Ring Test Summary

The following figure shows the collective communication data transfer timeline. The colored dots and connecting lines represent the time points at which each Rank completes communication during each collective operation cycle. The regular diamond-shaped patterns indicate highly consistent synchronization and communication among nodes, with no noticeable jitters or tail latency anomalies,

demonstrating excellent tail latency stability.



Picture 3-3 2 Nodes x 8 GPUs Single-Switch All-Reduce Unidir Ring Collective Data Transfer Timeline

The upper figure shows the CDF (Cumulative Distribution Function) of Data Chunk Completion Time, which is used to evaluate the distribution of communication latency.

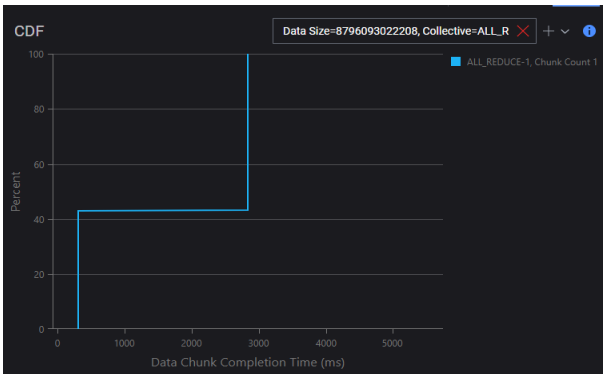
The curve displays two distinct steps:

The first step corresponds to intra-node NVLINK communication, which features extremely low latency and high transmission efficiency.

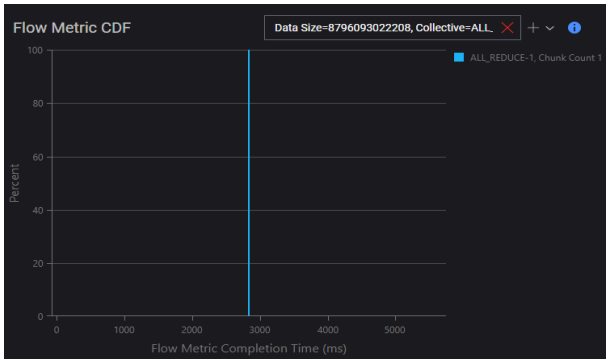
The second step corresponds to inter-node communication via the Spine-Leaf network topology, where latency is higher due to network path traversal, queuing, and scheduling overhead.

The below figure shows the Flow Metric CDF (Cumulative Distribution Function of Flow Metrics), a critical statistical tool used to assess the performance distribution of individual data flows in AI cluster communications — particularly in terms of latency consistency, tail latency, and link load balancing.

By observing the slope and extension at the tail of the CDF curve, one can identify the presence of “long-tail” flows. A steeper curve indicates good latency consistency and high synchronization efficiency across the cluster, whereas a flatter curve suggests larger latency variation between flows, which may slow down collective operations.



Picture 3-4 2 Nodes x 8 GPUs Single-Switch All-Reduce Unidir Ring Data Chunk Completion Time CDF



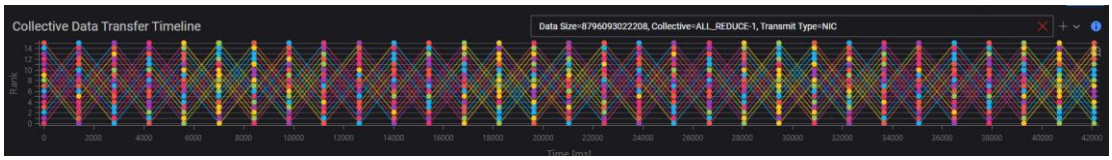
Picture 3-5 2 Nodes x 8 GPUs Single-Switch All-Reduce Unidir Ring Flow Metric CDF

All-Reduce Bidir Ring

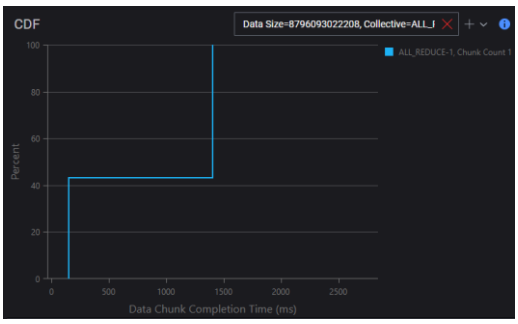
Trial Summary

Index	Data Size	Collective	Completion Time (ms)	Algbw (GB/s)	Busbw (GB/s)	Ideal (%)	PFC Rx	ECN CE Rx	Frames ReTx	Min FCT (ms)	P50 FCT (ms)	P95 FCT (ms)	Max FCT (ms)
1	8,796,093,022,208	ALL_REDUCE-1	42,100.21	208.93	391.75	98.12	0	0	0	152.709	152.709	1,403.152	1,403.235
2	8,796,093,022,208	ALL_REDUCE-2	42,100.39	208.93	391.75	98.12	0	0	0	152.709	152.709	1,403.154	1,403.249

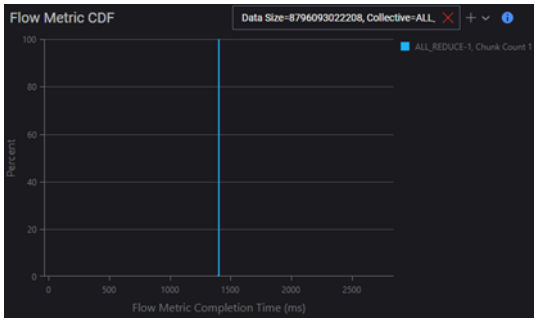
Picture 3-6 2 Nodes x 8 GPUs Single-Switch All-Reduce Bidir Ring Test Summary



Picture 3-7 2 Nodes x 8 GPUs Single-Switch All-Reduce Bidir Ring Collective Data Transfer Timeline



Picture 3-8 2 Nodes x 8 GPUs Single-Switch All-Reduce Bidir Ring Data Chunk Completion Time CDF



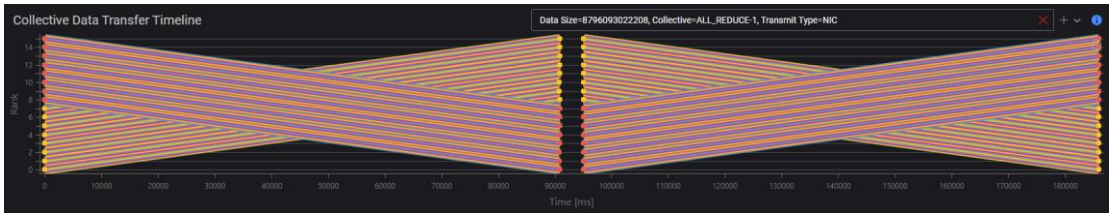
Picture 3-9 2 Nodes x 8 GPUs Single-Switch All-Reduce Bidir Ring Flow Metric CDF

All-Reduce Halving Doubling

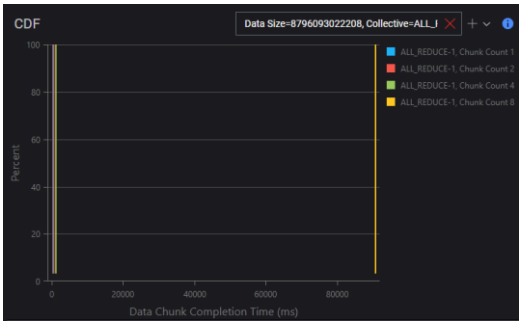
Trial Summary

Index	Data Size	Collective	Completion Time (ms)	Algbw (GB/s)	Busbw (GB/s)	Ideal (%)	PFC Rx	ECN CE Rx	Frames ReTx	Min FCT (ms)	P50 FCT (ms)	P95 FCT (ms)	Max FCT (ms)
1	8,796,093,022,208	ALL_REDUCE-1	185,911.22	47.31	88.71	97.72	0	0	0	305.419	916.259	90,816.363	90,817.230
2	8,796,093,022,208	ALL_REDUCE-2	185,911.22	47.31	88.71	97.72	0	0	0	305.419	916.259	90,816.372	90,817.264

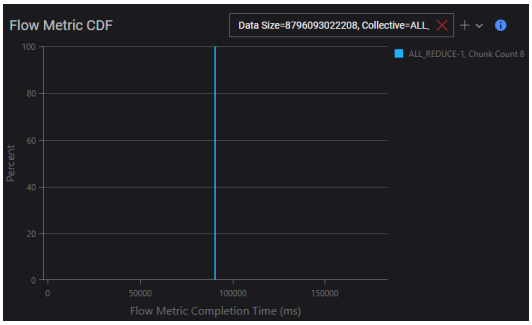
Picture 3-10 2 Nodes x 8 GPUs Single-Switch All-Reduce Halving Doubling Test Summary



Picture 3-11 2 Nodes x 8 GPUs Single-Switch All-Reduce Halving Doubling Collective Data Transfer Timeline



Picture 3-12 2 Nodes x 8 GPUs Single-Switch All-Reduce Halving Doubling Data Chunk Completion Time CDF



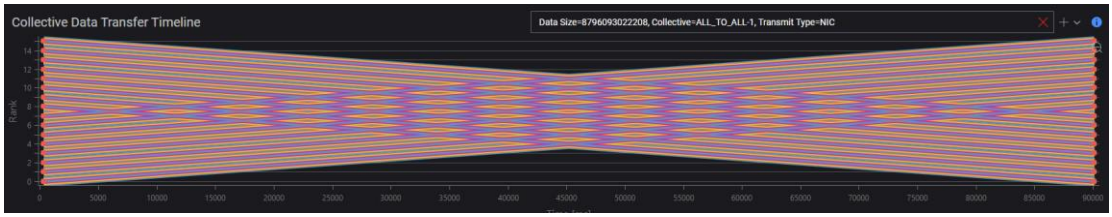
Picture 3-13 2 Nodes x 8 GPUs Single-Switch All-Reduce Halving Doubling Flow Metric CDF

All-to-All PXN

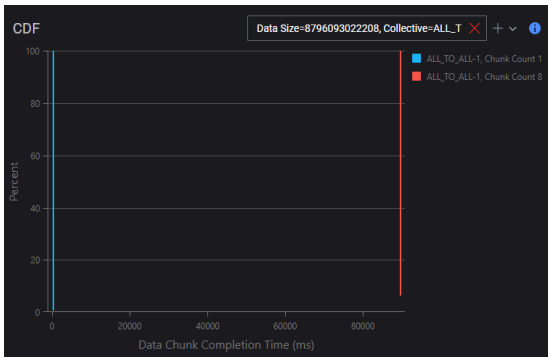
Trial Summary

Index	Data Size	Collective	Completion Time (ms)	Algbw (GB/s)	Busbw (GB/s)	Ideal (%)	PFC Rx	ECN CE Rx	Frames ReTx	Min FCT (ms)	P50 FCT (ms)	P95 FCT (ms)	Max FCT (ms)
1	8,796,093,022,208	ALL_TO_ALL-1	90,105.82	97.62	91.52	98.16	0	0	0	305.419	305.419	89,799.985	89,800.021
2	8,796,093,022,208	ALL_TO_ALL-2	90,105.77	97.62	91.52	98.16	0	0	0	305.419	305.419	89,799.984	89,800.019

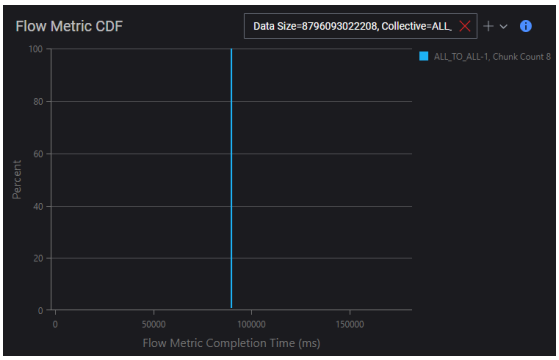
Picture 3-14 2 Nodes x 8 GPUs Single-Switch All-to-All PXN Test Summary



Picture 3-15 2 Nodes x 8 GPUs Single-Switch All-to-All PXN Collective Data Transfer Timeline



Picture 3-16 2 Nodes x 8 GPUs Single-Switch All-to-All PXN Data Chunk Completion Time CDF



Picture 3-17 2 Nodes x 8 GPUs Single-Switch All-to-All PXN Flow Metric CDF

- **4 Nodes × 4 GPUs**

Type	Algorithm	Busbw (GB/s)	Ideal (%)
All-Reduce	Unidir Ring	96.89	97.26
All-Reduce	Bidir Ring	193.68	97.21
All-Reduce	Halving Doubling	60.10	98.01
All-to-All	PXN	60.79	98.01

Table 3-3 4 Nodes x 4 GPUs Single-Switch Benchmark

In the single-switch benchmark test, only one switch is inserted between the test instruments. Both All-Reduce and All-to-All operations achieved over 97% bandwidth utilization. Compared to the back-to-back benchmark test, this indicates that flow

control and path scheduling in the RoCEv2 network have essentially reached an ideal state, aligning with expected results.

3.3 Spine-Leaf Test

We now introduce the full Spine-Leaf test topology to simulate a real-world AI data center environment.

The first scenario presents test results under a 2-node × 8 GPUs configuration using the default ECMP load balancing. Traditional ECMP mechanisms suffer from hash collisions and path stickiness due to 5-tuple hashing, resulting in uneven network load. In real-world networks with multiple concurrent elephant flows and mouse flows, the ideal bandwidth utilization typically reaches only 60% to 75%. The following test results, which simulate multiple concurrent elephant flows, also support this conclusion.

3.3.1 ECMP Load Balancing

- **2 Nodes × 8 GPUs ECMP 5-Tuple Hash**

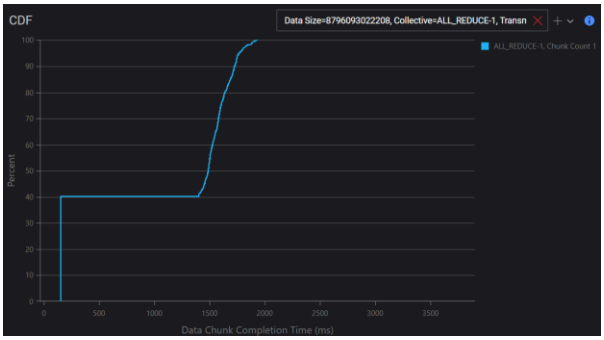
Type	Algorithm	Busbw (GB/s)	Ideal (%)	Remark
All-Reduce	Bidir Ring	337.65	84.57	ECMP 5-Tuple Hash QPs=16

Table 3-4 2 Nodes x 8 GPUs Spine-Leaf ECMP 5-Tuple Hash

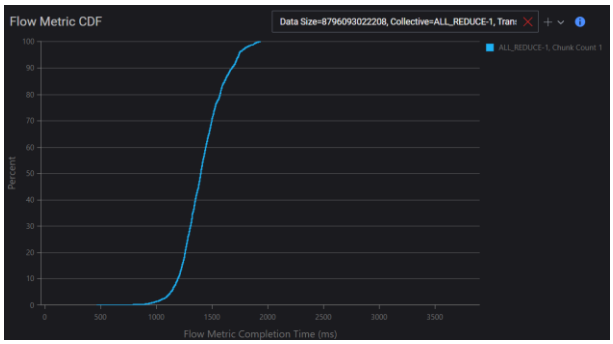
The original test report image is shown below.

Picture 3-18 2 Nodes x 8 GPUs Spine-Leaf ECMP 5-Tuple Hash All-Reduce Bidir Ring Test Summary

Picture 3-19 2 Nodes x 8 GPUs Spine-Leaf ECMP 5-Tuple Hash All-Reduce Bidir Ring Collective Data Transfer Timeline



Picture 3-20 2 Nodes x 8 GPUs Spine-Leaf ECMP 5-Tuple Hash All-Reduce Bidir Ring Data Chunk Completion Time CDF



Picture 3-21 2 Nodes x 8 GPUs Spine-Leaf ECMP 5-Tuple Hash All-Reduce Bidir Ring Flow Metric CDF

- **2 Nodes × 8 GPUs ECMP QP Hash**

Type	Algorithm	Busbw (GB/s)	Ideal (%)	Remark
All-Reduce	Bidir Ring	332.5	83.28	ECMP QP Hash QPs=8
All-Reduce	Bidir Ring	351.54	88.05	ECMP QP Hash QPs=16

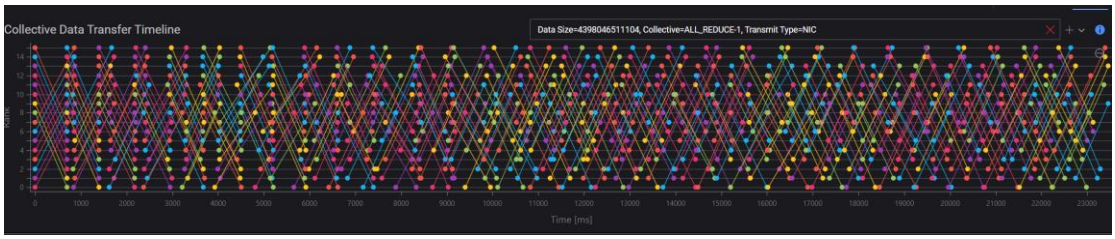
Table 3-5 2 Nodes x 8 GPUs Spine-Leaf ECMP QP Hash

The original test report image is shown below

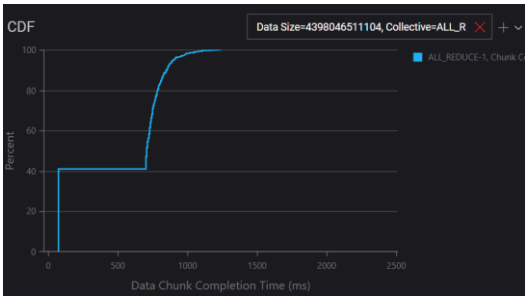
Trial Summary

Index	Data Size	Collective	Completion Time (ms)	Algbw (GB/s)	Busbw (GB/s)	Ideal (%)	PFC Rx	ECN CE Rx	Frames ReTx	Min FCT (ms)	P50 FCT (ms)	P95 FCT (ms)	Max FCT (ms)
1	4,398,046,511,104	ALL_REDUCE-1	23,457.90	187.49	351.54	88.05	1910091	19122222	1830657	152.709	152.709	1523.248	2,238.612
2	4,398,046,511,104	ALL_REDUCE-2	23,472.07	187.37	351.33	88.00	2013530	19094022	1848469	152.709	152.709	1521.003	2,207.833

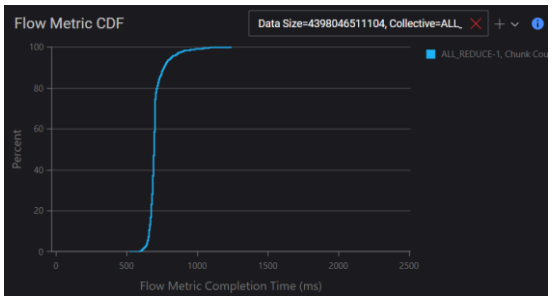
Picture 3-22 2 Nodes x 8 GPUs Spine-Leaf ECMP QP Hash All-Reduce Bidir Ring Test Summary



Picture 3-23 2 Nodes x 8 GPUs Spine-Leaf ECMP QP Hash All-Reduce Bidir Ring Collective Data Transfer Timeline



Picture 3-24 2 Nodes x 8 GPUs Spine-Leaf ECMP QP Hash All-Reduce Bidir Ring Data Chunk Completion Time CDF



Picture 3-252 Nodes x 8 GPUs Spine-Leaf ECMP QP Hash All-Reduce Bidir Ring Flow Metric CDF

3.3.2 INT-Driven Adaptive Routing

With INT-Driven Adaptive Routing enabled in the RoCEv2 network, the NCCL All-Reduce test showed a bandwidth utilization improvement of over 16%, reaching up to 97%, which is close to the performance limit of back-to-back connections.

- **2 Nodes × 8 GPUs INT-Driven Adaptive Routing**

Type	Algorithm	Busbw (GB/s)	Ideal (%)	Remark
All-Reduce	Bidir Ring	386.50	96.81	Spine-Leaf (INT-Driven Adaptive Routing)
All-Reduce	Halving Doubling	87.93	96.84	
All-to-All	PXN	90.62	97.19	

Table 3-6 2 Nodes x 8 GPUs Spine-Leaf INT-Driven Adaptive Routing

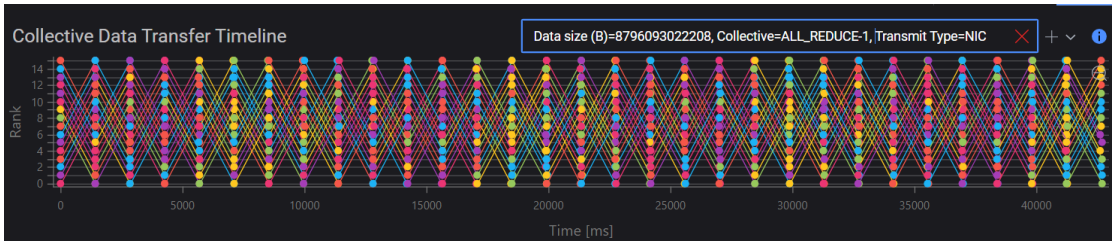
The original test report image is shown below

All-Reduce Bidir Ring

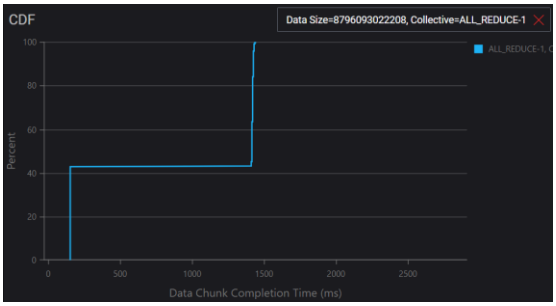
Trial Summary

Index	Data Size	Collective	Completion Time (ms)	Algbw (GB/s)	Busbw (GB/s)	Ideal (%)	PFC Rx	ECN CE Rx	Frames ReTx	Min FCT (ms)	P50 FCT (ms)	P95 FCT (ms)	Max FCT (ms)
1	8,796,093,022,208	ALL_REDUCE-1	42,672.15	206.13	386.50	96.81	0	50156361	2593256	152.709	152.709	1,421.533	1,445.674
2	8,796,093,022,208	ALL_REDUCE-2	42,716.24	205.92	386.10	96.71	0	50463915	2600012	152.709	152.709	1,422.452	1,445.703

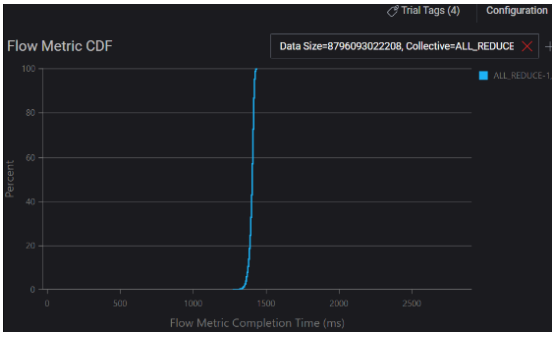
Picture 3-26 2 Nodes x 8 GPUs Spine-Leaf INT-Driven Adaptive Routing All-Reduce Bidir Ring Test Summary



Picture 3-27 2 Nodes x 8 GPUs Spine-Leaf INT-Driven Adaptive Routing All-Reduce Bidir Ring Collective Data Transfer Timeline



Picture 3-28 2 Nodes x 8 GPUs Spine-Leaf INT-Driven Adaptive Routing All-Reduce Bidir Ring Data Chunk Completion Time CDF



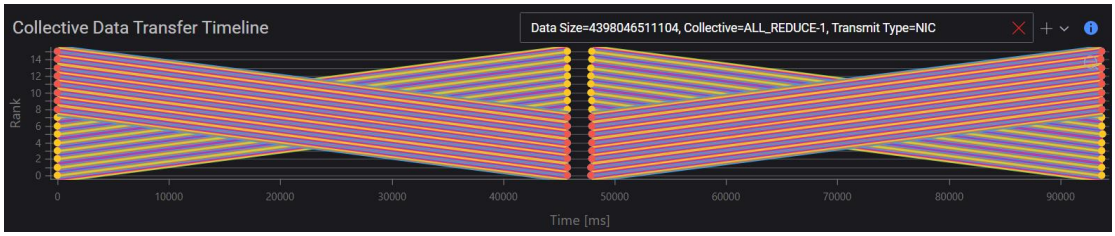
Picture 3-29 2 Nodes x 8 GPUs Spine-Leaf INT-Driven Adaptive Routing All-Reduce Bidir Ring Flow Metric CDF

All-Reduce Halving Doubling

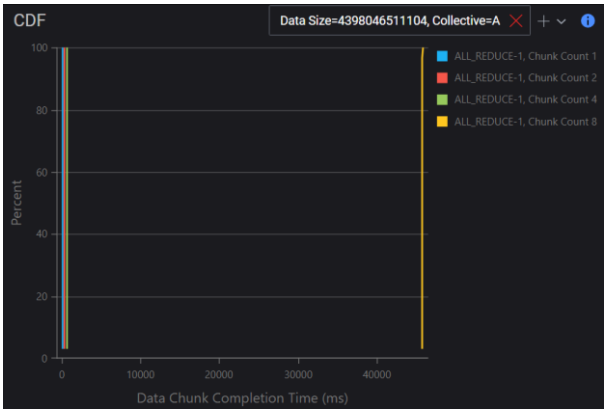
Trial Summary

Index	Data Size	Collective	Completion Time (ms)	Algbw (GB/s)	Busbw (GB/s)	Ideal (%)	PFC Rx	ECN CE Rx	Frames ReTx	Min FCT (ms)	P50 FCT (ms)	P95 FCT (ms)	Max FCT (ms)
1	4,398,046,511,104	ALL_REDUCE-1	93,780.40	46.90	87.93	96.86	0	85509154	7981485	152.709	458.129	45,774.802	45,875.805
2	4,398,046,511,104	ALL_REDUCE-2	93,800.69	46.89	87.91	96.84	206	85702625	7997380	152.709	458.129	45,771.448	45,891.221

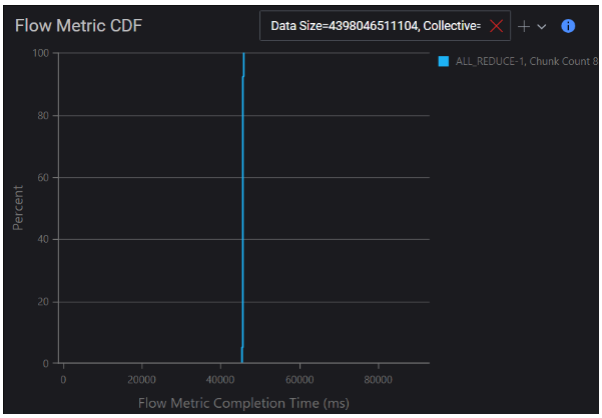
Picture 3-30 2 Nodes x 8 GPUs Spine-Leaf INT-Driven Adaptive Routing All-Reduce Halving Doubling Test Summary



Picture 3-31 2 Nodes x 8 GPUs Spine-Leaf INT-Driven Adaptive Routing All-Reduce Halving Doubling Collective Data Transfer Timeline



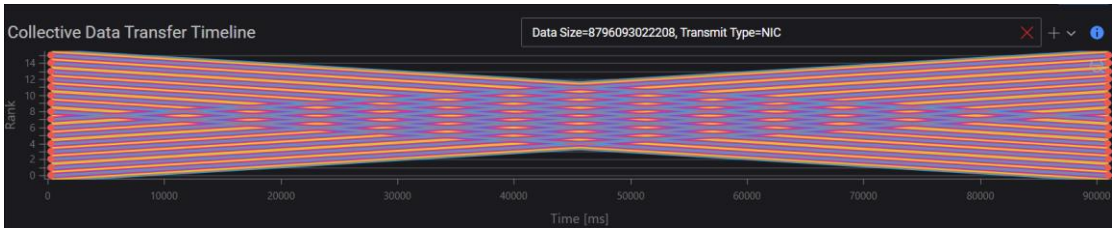
Picture 3-32 2 Nodes x 8 GPUs Spine-Leaf INT-Driven Adaptive Routing All-Reduce Halving Doubling Data Chunk Completion Time CDF



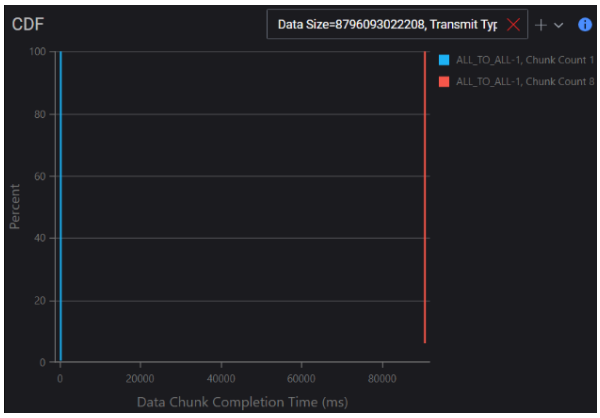
Picture 3-33 2 Nodes x 8 GPUs Spine-Leaf INT-Driven Adaptive Routing All-Reduce Halving Doubling Flow Metric CDF

All-to-All PXN

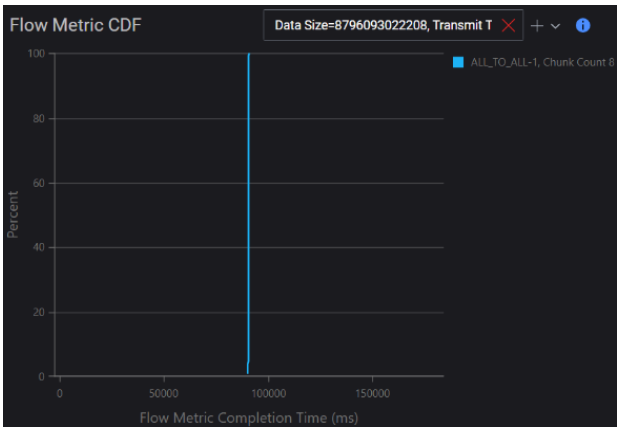
Picture 3-34 2 Nodes x 8 GPUs Spine-Leaf INT-Driven Adaptive Routing All-to-All PXN Test Summary



Picture 3-35 2 Nodes x 8 GPUs Spine-Leaf INT-Driven Adaptive Routing All-to-All PXN Collective Data Transfer Timeline



Picture 3-36 2 Nodes x 8 GPUs Spine-Leaf INT-Driven Adaptive Routing All-to-All PXN Data Chunk Completion Time CDF



Picture 3-37 2 Nodes x 8 GPUs Spine-Leaf INT-Driven Adaptive Routing All-to-All PXN Flow Metric CDF

Test results with INT-Driven Adaptive Routing enabled validated in other topologies.

- **4 Nodes × 4 GPUs INT-Driven Adaptive Routing**

Type	Algorithm	Busbw (GB/s)	Ideal (%)	Remark

All-Reduce	Unidir Ring	96.88	97.25	Spine-Leaf (INT-Driven Adaptive Routing)
All-Reduce	Bidir Ring	193.70	97.22	
All-Reduce	Halving Doubling	59.70	97.36	
All-to-All	PXN	60.74	97.93	

Table 3-7 4 Nodes x 4 GPUs Spine-Leaf INT-Driven Adaptive Routing

- **8 Nodes × 2 GPUs INT-Driven Adaptive Routing**

Type	Algorithm	Busbw (GB/s)	Ideal (%)	Remark
All-Reduce	Unidir Ring	47.88	97.15	Spine-Leaf (INT-Driven Adaptive Routing)
All-Reduce	Bidir Ring	96.85	97.59	
All-Reduce	Halving Doubling	51.27	97.26	
All-to-All	PXN	50.73	95.62	

Table 3-8 8 Nodes x 2 GPUs Spine-Leaf INT-Driven Adaptive Routing

The excellent performance of a Spine-Leaf network relies heavily on robust traffic scheduling and flow control mechanisms. RoCEv2 ensures basic lossless transmission through PFC, enables software-based rate adjustment via ECN + DCQCN, and achieves path-level load balancing with Flowlet-based Adaptive Routing — together building an open, flexible, and highly resilient flow control system.

3.4 Oversubscription Test

The oversubscription ratio is the proportion between downlink access bandwidth (Leaf-to-Host) and uplink aggregation bandwidth (Leaf-to-Spine). It directly determines the network's bandwidth redundancy and concurrent performance.

Type	Algorithm	Oversubscription	Busbw (GB/s)	Ideal (%)	Actual Network Utilization (%)
All-Reduce	Bidir Ring	1:1	386.50	96.81	96.81
All-Reduce	Bidir Ring	1:0.75	292.66	73.31	97.70
All-Reduce	Bidir Ring	1:0.5	193.39	48.44	96.88

Table 3-9 2 Nodes ×8 GPUs Spine-Leaf Oversubscription Test Result

Under non-ideal link conditions (1:0.75 oversubscription), performance still reached 73.31%. A 1:0.75 ratio implies uplink congestion, meaning the Spine cannot handle all

traffic if all GPUs communicate simultaneously. Thus, the theoretical limit is around 75% bandwidth utilization. The actual result of 73.31% implies a 97.7% completion relative to the theoretical limit (73.31% ÷ 75%), which is very close to the ideal.

Similarly, under the extreme 1:0.5 scenario, the bandwidth was maintained at 48.44%, meaning 96.88% completion (48.44% ÷ 50%), demonstrating strong fault tolerance and congestion recovery.

4 Conclusion

Unlike previous tests that primarily focused on the standalone AI performance of a single switch, this round of testing simulated a real-world AIDC (AI Data Center) Spine-Leaf network architecture to comprehensively validate multi-node, multi-GPU distributed communication scenarios.

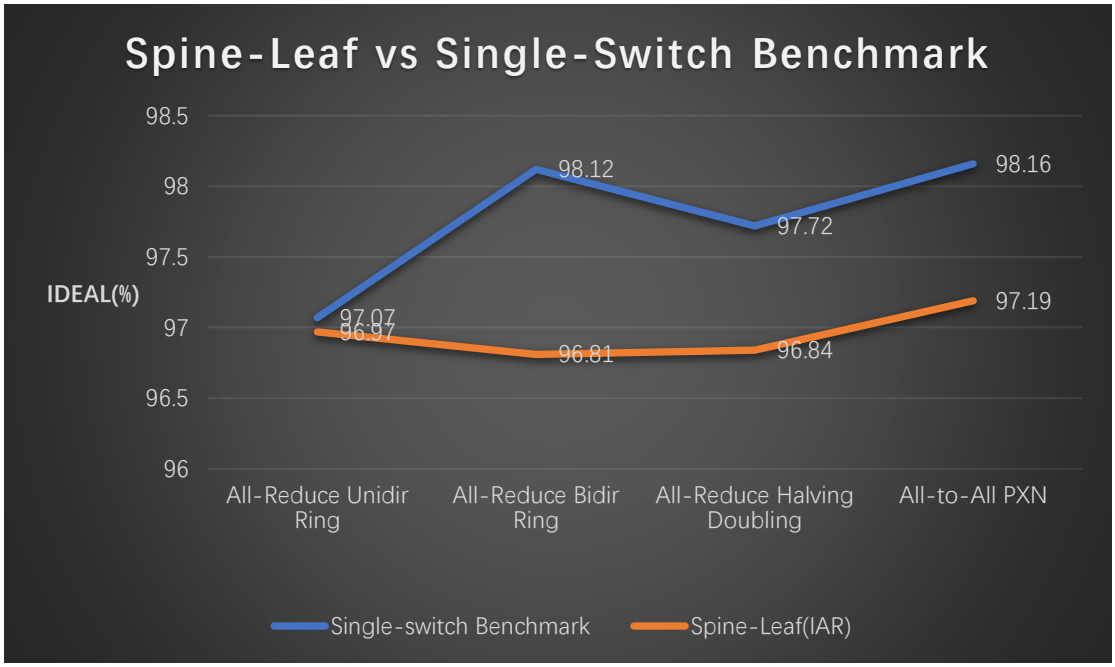


Figure 4-1 2 Node×8 GPUs Spine-Leaf vs Single-Switch Benchmark

- The test covered typical topologies such as 2 nodes × 8 GPUs, 4 nodes × 4 GPUs, and 8 nodes × 2 GPUs, using the IXIA AresONE-400GE traffic generator to simulate fine-grained flow control and accurately reproduce complex NCCL communication patterns like All Reduce, All Gather, and All to All.
- With intelligent scheduling and congestion control mechanisms such as ECMP,

INT-Driven Adaptive Routing, and PFC/ECN enabled, we achieved up to 97% bandwidth utilization for Spine-Leaf communication traffic — significantly outperforming traditional RoCE traffic balancing strategies.

- At the same time, INT-Driven Adaptive Routing significantly improved tail latency performance. In multi-port concurrent and burst traffic scenarios, compared with ECMP 5-Tuple Hash and ECMP QP Hash, the P95 FCT was reduced by 11.13% and 6.70% respectively. The tail latency was close to the single-switch benchmark value, verifying its congestion management and forwarding capabilities.
- The test further verified the device's ultra-low latency performance, with end-to-end latency in the Spine-Leaf network observed as low as 2μs.

This test not only validated the scheduling efficiency and scalability of RoCEv2 networks in large-scale AI clusters, but also highlighted the latency control and traffic consistency advantages of Asterfusion AI switches under extreme communication density — providing a solid foundation for future deployment of large model inference platforms and network architecture planning.