NVIDIA Spectrum-X Ethernet Network Platform Architecture

The First Ethernet Network Designed to Accelerate AI Workloads

White Paper

12652-002 | December 2025

Table of Contents

Improve AI Performance and Efficiency 1
NVIDIA Spectrum-X Ethernet Networking Platform Overview 4
Inside The NVIDIA Spectrum-X Ethernet Networking Platform 4
Key Benefits of Spectrum-X Ethernet 5
NVIDIA Spectrum-X Ethernet Switches 5
NVIDIA Spectrum-X Ethernet SuperNICs 6
NVIDIA End-to-End Physical Layer 8
NVIDIA Acceleration Software 9
NetQ 9
Spectrum Software Development Kit 9
NVIDIA DOCA 10
Comprehensive AI Fabric Security 10
Networking Platform Component Security 10
AI Fabric Network Isolation 11
AI Fabric Data-in-Transit Encryption 11
The Challenges of Deploying AI on Ethernet 12
RDMA and a Lossless Network is Required 12
RoCE Adaptive Routing is Required to Avoid Congestion 13
Entropy Matters 13
NVIDIA RoCE Adaptive Routing Ensures Load Balancing 15
NVIDIA RoCE Congestion Control is Required 16
Key Features of the NVIDIA Spectrum-X Ethernet Networking Platform 16
How NVIDIA RoCE Adaptive Routing Works 17
NVIDIA RoCE Adaptive Routing with NVIDIA Direct Data Placement 17
RoCE Adaptive Routing Summary 20
Performance Isolation Using NVIDIA RoCE Congestion Control 21
RoCE Performance Isolation 23
NVIDIA Full-Stack Optimization 25
Spectrum-X Ethernet Performance Results from Israel-1 26
RDMA Fabric Performance 26

NCCL Collective Performance 27
AI Model Performance 28
Impact of Virtualization on Performance 29
Data Center Scale-Across Performance 30
The Ethernet Designed for AI 31

List of Figures

Figure 1. Differences between control/user access network and E/W AI network 3
Figure 2. NVIDIA Spectrum-X SN5610 Ethernet Switch 6
Figure 3. BlueField-3 Ethernet SuperNIC 7
Figure 4. ConnectX-8 SuperNIC 8
Figure 5. Typical Spectrum-X Ethernet Network Topology 8
Figure 6. NVIDIA NetQ telemetry dashboard 9
Figure 7. RDMA enables BlueField with Direct Access to GPU memory 12
Figure 8. AI workloads reduced to 60 percent of max throughput with static load balancing 13
Figure 9. Why even non-blocking networks can experience long tail latency 14
Figure 10. Flow completion times can vary significantly 15
Figure 11. Incast requires NVIDIA RoCE Congestion Control 16
Figure 12. Flow completion time is reduced with RoCE Adaptive Routing 20
Figure 13. Network congestion creating a victim flow 22
Figure 14. Spectrum-X Ethernet solves congestion with flow metering and congestion telemetry 23
Figure 15. Spectrum-X Ethernet mechanisms involved in delivering noise isolation 24
Figure 16. Spectrum-X Ethernet switches feature a universal shared buffer design 25
Figure 17. Spectrum-X Ethernet achieves higher effective bandwidth and lower latency for RDMA bisection. 27
Figure 18. Spectrum-X Ethernet delivers higher bus bandwidth for NCCL collectives running in an AI factory 28
Figure 19. Spectrum-X Ethernet accelerates training AI workloads in noisy AI factory environments. 29
Figure 20. Spectrum-X Ethernet performs equally well with and without network virtualization. 30
Figure 21. 10km Distance NCCL All-Reduce performance between data centers 31

Improve AI Performance and Efficiency

AI workload demand is growing at an unprecedented rate, while the adoption of generative AI is surging. Every year witnesses the construction of additional AI factories. These facilities, dedicated to the development and operation of artificial intelligence technologies, are increasingly expanding into Cloud Service Providers (CSP) and Enterprise data centers.

AI factories, which are capable of training trillion-parameter foundational models, require three specialized and highly optimized networks:

1. **North/South Fabric:** A dedicated network for user and storage access.
2. **Scale-Up Fabric:** Utilizes **NVIDIA NVLink** as a memory-coherent interconnect, linking multiple GPUs to form a unified, massively powerful compute domain.
3. **East/West Scale-Out Fabric:** Connects thousands of GPUs, enabling demanding AI workloads to scale horizontally.

For deployment in CSP or Enterprise Data Centers, the AI factory must be designed to integrate seamlessly into existing Ethernet service networks and management frameworks, supporting a broader range of users and operational requirements:

- **AI in Software as a Service (SaaS) migration and AI as a Service:** SaaS offerings are increasingly incorporating AI and offering AI as a Service, requiring multi-tenant security and multi-job performance isolation.
- **Cloud-scale Software Defined Networking (SDN):** This provisioning model utilizes Ethernet-based protocols and container orchestration (e.g., Kubernetes). SDN supports scalable provisioning from small clusters to large clusters and enables reprovisioning of compute and network resources into multiple small clusters.
- **Integration and Compatibility:** AI services from CSPs should closely integrate with other cloud services, using Ethernet to manage complete data pipelines in the cloud.
- **Security and Compliance:** CSPs implement robust security measures, including packet brokering and Ethernet-based intrusion detection/protection devices, to safeguard client data and adhere to regional regulations.
- **Open Source:** Supports open network operating systems, like SONiC.

Off-the-shelf Ethernet networks, optimized for shared computing infrastructure, are sub-optimal for AI factories. A few differences between control/user access network and East/west AI fabric are outlined in Figure 1.

Figure 1. Differences between control/user access network and E/W AI network

Control /User Access Network (N-S)		AI Fabric (E-W)
Loosely-Coupled Applications		Tightly-Coupled Processes
TCP (Low Bandwidth Flows)		RDMA (High Bandwidth Flows)
High Jitter Tolerance		Low Jitter Tolerance
Oversubscribed Topologies		Nonblocking Topologies
Heterogeneous Traffic Average Multi-Pathing		Bursty Network Capacity Predictive Performance

When AI factories utilize off-the-shelf (OTS) Ethernet for the AI compute network, they achieve only a portion of the MLPerf performance levels possible with an optimized network. In multi-tenant environments where multiple AI jobs run simultaneously, off-the-shelf (OTS) Ethernet struggles to provide performance isolation, risking the interference of one tenant's jobs with another's. For those deploying AI factories with Ethernet, NVIDIA has developed the Spectrum-X Ethernet Networking Platform, which improves performance, while increasing the predictability and power efficiency of Ethernet-based AI factories.

NVIDIA Spectrum-X Ethernet Networking Platform Overview

The NVIDIA® Spectrum™-X Ethernet Networking Platform is the first Ethernet platform designed specifically to improve the performance and efficiency of Ethernet-based AI factories. Based on open standards, this breakthrough technology delivers markedly better network performance for massive AI workloads, such as LLMs and inferencing, than off-the-shelf (OTS) Ethernet solutions. It also improves power efficiency and ensures consistent, predictable performance in multi-tenant environments. Spectrum-X Ethernet leverages network innovations achieved through the tight coupling of the Spectrum-X Ethernet switches with NVIDIA SuperNICs, like the NVIDIA BlueField®-3 SuperNIC or NVIDIA ConnectX-8 SuperNIC. This integration facilitates the delivery of end-to-end network capabilities purpose-built for AI workloads, reducing runtimes of massive transformer-based generative AI models. As a result, network engineers, data scientists, and cloud service providers can attain faster results and make informed decisions.

Inside The NVIDIA Spectrum-X Ethernet Networking Platform

An effective AI compute network is defined by its ability to support and accelerate AI workloads. Optimization is crucial for every aspect of the network, from the SuperNICs to the switches, cables/optics, networking, and acceleration software. To address these needs, NVIDIA created new innovations that deliver the highest effective bandwidth under load and at scale:

1. NVIDIA RDMA over Converged Ethernet (RoCE) Adaptive Routing on Spectrum-X switches
2. NVIDIA Direct Data Placement (DDP) on Spectrum-X Ethernet SuperNIC
3. NVIDIA RoCE Congestion Control on both Spectrum-X Ethernet switches and SuperNICs
4. NVIDIA AI Acceleration Software
5. End-to-End AI Network Visibility

These innovations form a full-stack solution that is tested, tuned, and benchmarked by NVIDIA to guarantee the highest level of performance. Integrating them into the Spectrum-X Ethernet Networking Platform offers numerous advantages.

Key Benefits of Spectrum-X Ethernet

- **Improved AI Factory Performance**: Spectrum-X Ethernet significantly enhances multi-tenant AI factory network performance.
- **Standard Ethernet Connectivity**: Spectrum-X provides fully standards-based Ethernet and is completely interoperable with Ethernet-based stacks.
- **Increased Power Efficiency**: By improving performance, Spectrum-X Ethernet contributes to a more power-efficient AI environment.
- **Enhanced Multi-Tenant Protection**: Performance isolation in multi-tenant environments ensures that each tenant's workloads perform optimally and consistently, resulting in higher customer satisfaction and improved service quality.
- **Better AI Fabric Visibility:** Visibility into the AI factory's data flows helps identify performance bottlenecks, which is key in a modern, automated fabric-validation solution.
- **Higher AI Scalability**: Spectrum-X Ethernet scales to 128 ports of 400 Gigabit Ethernet (GbE), in a single hop or 8K ports in a two-tier leaf/spine topology, supporting AI factory expansion while maintaining high performance levels.
- **Faster Network Setup**: The automated, end-to-end configuration of advanced networking functionality is fully tuned for AI workloads.
- **Comprehensive AI Fabric Security**: Spectrum-X Ethernet delivers end-to-end protection for AI data centers.
- **Optimized Scale-Across Networking**: Spectrum-XGS Ethernet enables geographically distributed data centers to operate as a single, unified AI factory.

The Spectrum-X Ethernet Networking Platform is a full-stack solution built on the following key components:

NVIDIA Spectrum-X Ethernet Switches

Spectrum-X Ethernet switches are built on a 51.2Tbps ASIC, supporting up to 64 ports of 800GbE or128 ports of 400GbE within a single 2U switch. They set a new standard as the first switch family designed from the ground up for AI workloads. By integrating specialized high-performance, low latency architecture with standard Ethernet connectivity, these switches deliver unparalleled efficiency and reliability, ensuring seamless operation and optimal performance for demanding AI applications.

Spectrum-X Ethernet switches offer RoCE Extensions for AI with unique enhancements:

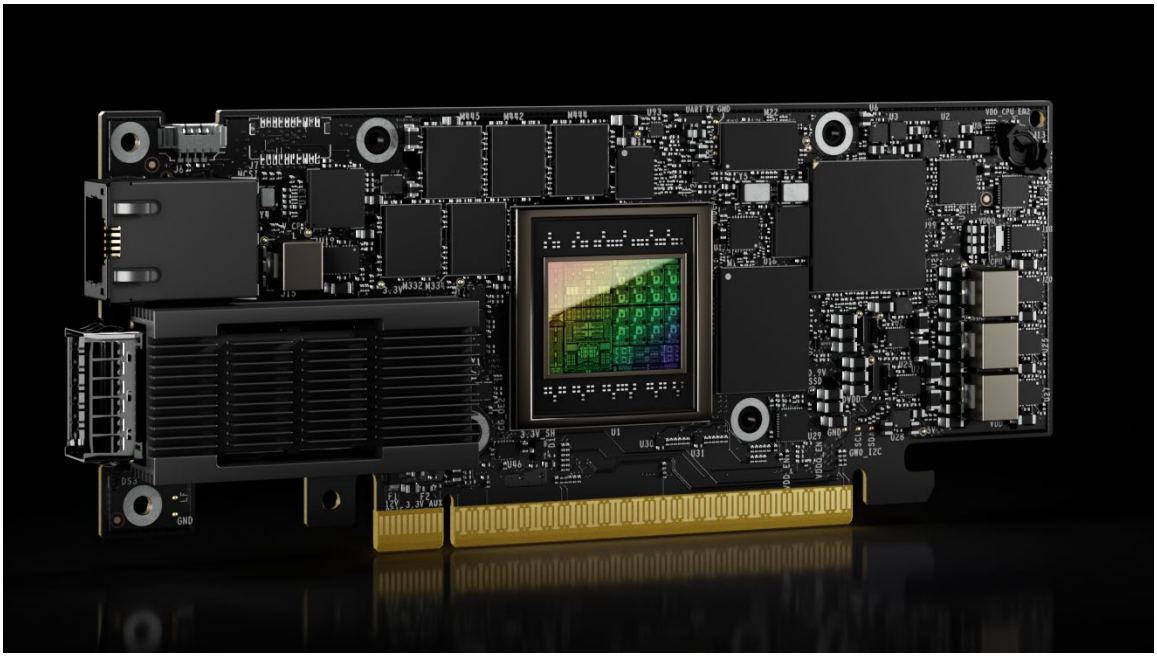
- RoCE Adaptive Routing
- RoCE Performance Isolation
- Highest effective bandwidth on standard Ethernet at scale
- Low latency with low jitter and short tail latency

Figure 2. NVIDIA Spectrum-X SN5610 Ethernet Switch

NVIDIA Spectrum-X Ethernet SuperNICs

The NVIDIA® Spectrum-X Ethernet SuperNIC is the 3rd-generation data center infrastructure-on-a-chip that enables organizations to build software-defined, hardware-accelerated IT infrastructures from cloud to core data center to edge. The BlueField®-3 SuperNIC provides up to 400Gb/s of RoCE network connectivity between GPU servers, optimizing peak AI workload efficiency and delivering deterministic and isolated performance between jobs and tenants.

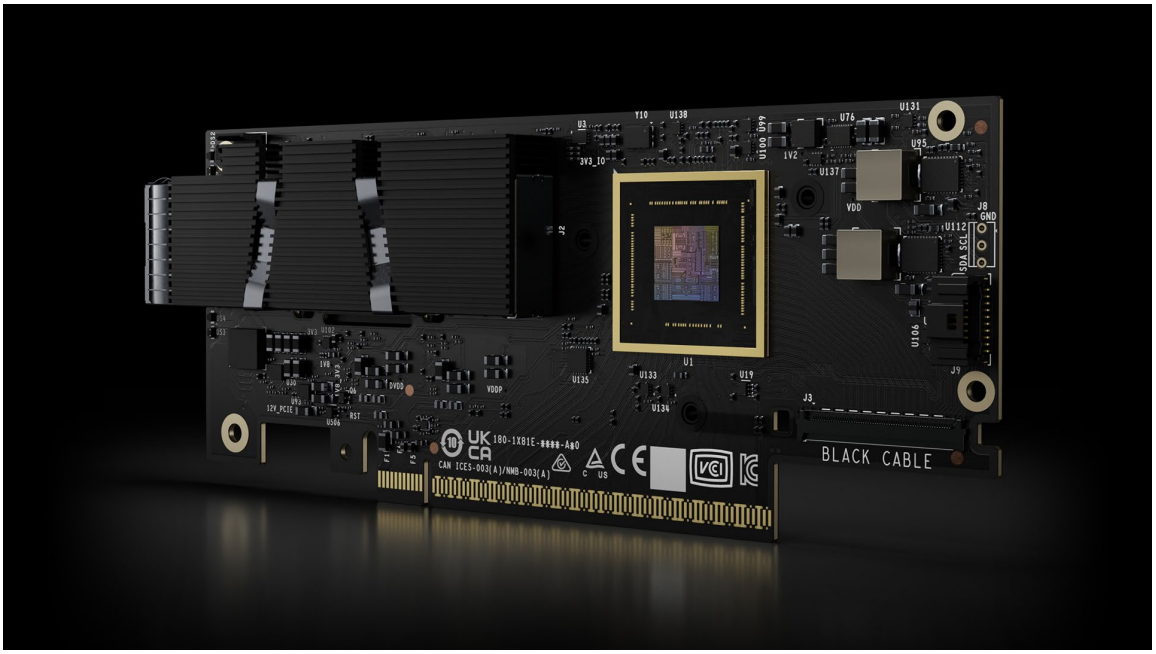
Figure 3. BlueField-3 Ethernet SuperNIC



BlueField-3 has been purpose-built for AI acceleration, featuring an integrated all-to-all engine for AI, NVIDIA GPUDirect®, and NVIDIA® Magnum IO GPUDirect® Storage (GDS) acceleration technology. Additionally, BlueField-3 SuperNICs natively run in network interface card (NIC) mode, taking advantage of local memory to accelerate large AI factories. At scale, these AI factories require a large number of Queue Pairs that can be addressed locally, rather than using system memory. Lastly, it includes NVIDIA Direct Data Placement (DDP) Technology which augments RoCE adaptive routing.

The NVIDIA ConnectX-8 SuperNIC provides up to 800Gb/s of RoCE network connectivity between GPU servers, optimizing peak AI workload efficiency and delivering deterministic and isolated performance between jobs and tenants. ConnectX-8 has been purpose-built for AI acceleration, featuring an integrated all-to-all engine for AI, NVIDIA GPUDirect, and NVIDIA Magnum IO GPUDirect Storage (GDS) acceleration technology. Lastly, it includes the same NVIDIA Direct Data Placement (DDP) Technology and programmable congestion control used in the BlueField-3 SuperNIC.

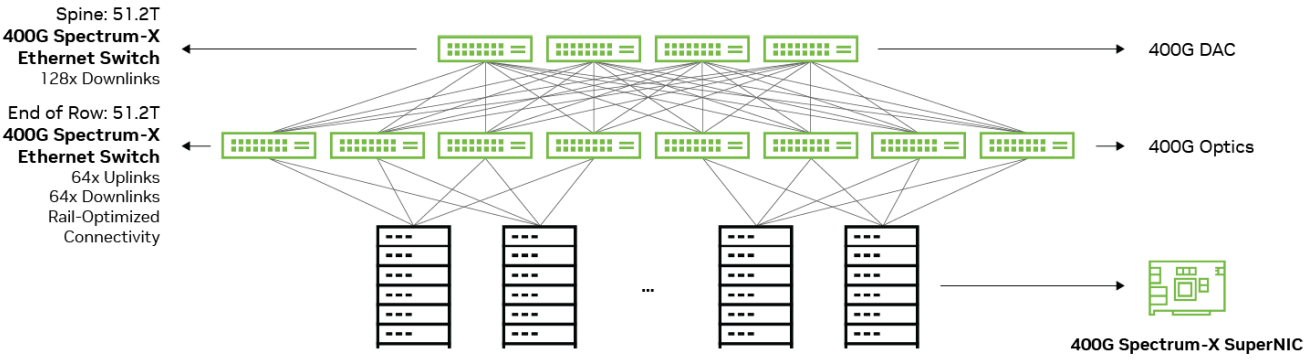
Figure 4. ConnectX-8 SuperNIC



NVIDIA End-to-End Physical Layer

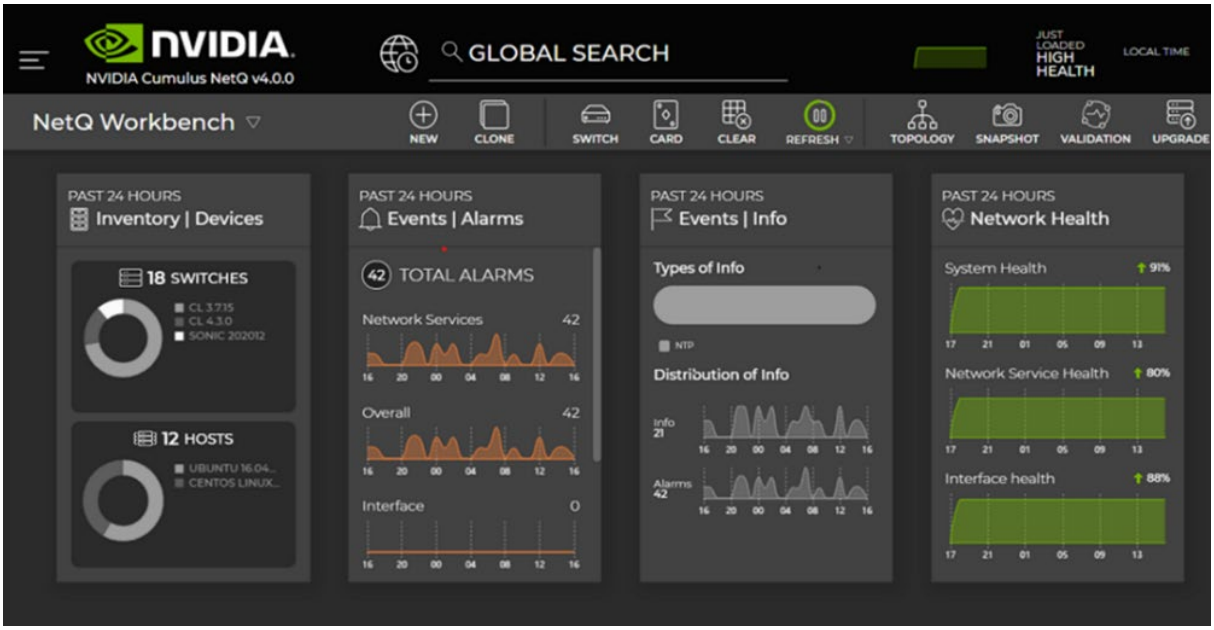
Spectrum-X Ethernet is the only networking platform providing a seamless connection from switch to SuperNIC to GPU. This end-to-end integration ensures exceptional signal integrity and the lowest BER (Bit Error Rate), greatly reducing the power consumption of the AI factory. Leveraging the same advanced Serdes technology found in NVIDIA GPUs, Spectrum-X Ethernet switch and SuperNIC, Spectrum-X Ethernet ensures optimal power efficiency and performance without compromise.

Figure 5. Typical Spectrum-X Ethernet Network Topology



NVIDIA Acceleration Software

Figure 6. NVIDIA NetQ telemetry dashboard



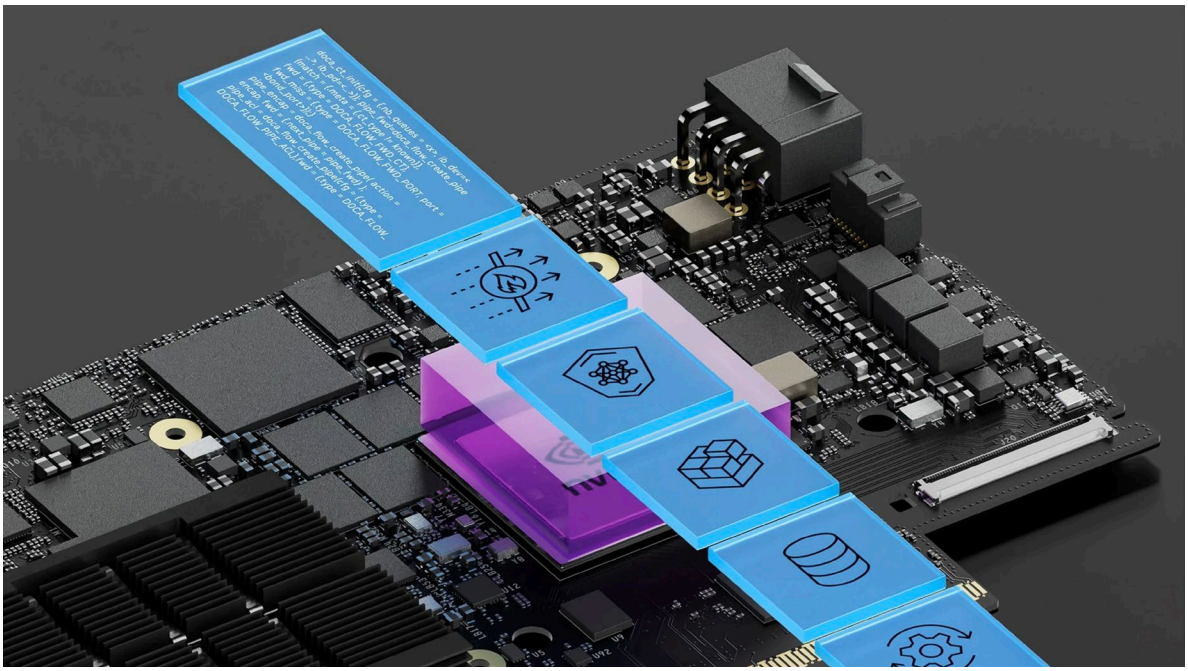
NetQ

NVIDIA NetQ™ is a highly scalable network operations toolset for real-time AI network visibility, troubleshooting, and validation. NetQ leverages 'NVIDIA What Just Happened' switch telemetry data and NVIDIA® DOCA™ telemetry to deliver actionable insights about the health of the Spectrum-X Ethernet switch and SuperNIC, integrating the network into an organization's MLOps ecosystem.

Additionally, NetQ flow telemetry maps flow paths and behavior across switch ports and RoCE queues for analyzing specific application flows. NetQ samples packets, analyzes, and reports per switch latency (maximum, minimum, and average) and buffer occupancy details along the path of the flow. The NetQ GUI reports all the possible paths, per-path details, and flow behavior. Combining NVIDIA What Just Happened® with flow telemetry helps network operators proactively identify root cause server and application issues.

Spectrum Software Development Kit

The NVIDIA Ethernet Switch Software Development Kit (SDK) provides the flexibility to implement any switching and routing functionality, with sophisticated programmability that does not compromise performance in packet rate, bandwidth, or latency. With the SDK, server and networking OEMs and network operating system (NOS) vendors can take advantage of the advanced networking features of the Ethernet switch family of integrated circuits (ICs) and build flexible, innovative, and cost-optimized switching solutions.



NVIDIA DOCA

NVIDIA® DOCA™ software framework is the key to unlocking the potential of the NVIDIA BlueField® networking platform. With DOCA, developers can program the data center infrastructure of tomorrow by creating software-defined, cloud-native, BlueField-accelerated services with zero-trust protection to address the increasing performance and security demands of modern data centers.

Comprehensive AI Fabric Security

AI fabric security is a critical concern in modern data centers that support complex AI workloads, ensuring operational continuity, regulatory compliance, and organizational trust in digital infrastructure. Integral to NVIDIA's Spectrum-X Ethernet AI platform, Spectrum-X Ethernet security encompasses hardware, firmware, and software mitigations to effectively protect these environments .

Networking Platform Component Security

Spectrum-X's components—including Spectrum-X Ethernet switches, BlueField-3 SuperNICs, ConnectX-8 SuperNICs, and LinkX cables—are engineered with security as a foundational element. BlueField-3 and ConnectX-8 devices feature secure boot processes, hardware roots-of-trust, secure firmware updates, flash encryption, and device attestation to prevent firmware-level attacks and unauthorized changes. They also incorporate dedicated hardware countermeasures against physical attacks such as fault injections. These devices support software-defined networking and programmable pipelines, enabling detailed policy enforcement and accelerating security functions. Key

security protocols like PSP and IPsec are offloaded and accelerated to facilitate AI fabric data encryption in transit.

Spectrum-X Ethernet switches implement a comprehensive security framework spanning hardware, firmware, and software layers, all authenticated by a built-in root of trust extending from ASIC firmware through BIOS to the network operating system. Physical protection countermeasures are also applied. On the network security front, Spectrum-X Ethernet supports VXLAN for securing overlay networks and advanced network virtualization with L3EVPN, enabling secure, scalable segmentation and tenant isolation.

AI Fabric Network Isolation

In multi-tenant environments, network isolation is vital to ensure that each tenant's traffic remains fully separated, safeguarding against unauthorized access, preserving privacy, supporting compliance, and preventing cross-tenant performance or security impacts. Spectrum-X Ethernet supports L3EVPN with VXLAN BGP EVPN fabrics, providing logically isolated routing domains (VRFs) for each tenant. Each VRF functions as a unique virtual network with a distinct virtual network interface (VNI), segregating routing tables, policies, and traffic among tenants. Additionally, NVIDIA's patented use of IEEE 802.1X for tenant provisioning offers a distributed and resilient system that allows seamless and secure network access without incremental per-tenant changes.

AI Fabric Data-in-Transit Encryption

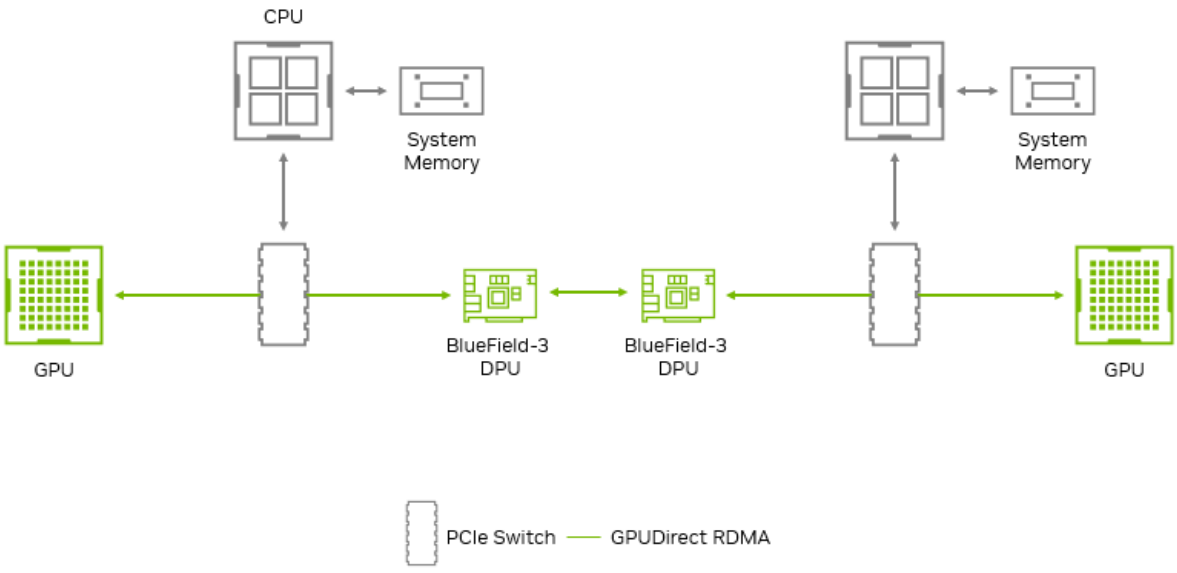
Encrypting data in transit is essential to maintain confidentiality, integrity, and compliance in AI fabric architecture across both single-tenant and multi-tenant setups. BlueField-3 and ConnectX-8 SuperNICs support offloading of security protocols such as IPsec and PSP, which reduces host CPU load and ensures transparent, line-rate encryption of network traffic before it leaves the node. This capability enables comprehensive end-to-end encryption between GPUs within the AI fabric.

The Challenges of Deploying AI on Ethernet

RDMA and a Lossless Network is Required

AI needs GPUs to perform well, and for multiple GPUs to perform well together, they need Remote Direct Memory Access (RDMA)—a networking technology that allows access from the memory of one computer into another computer without involving either computer's CPU.

Figure 7. RDMA enables BlueField with Direct Access to GPU memory

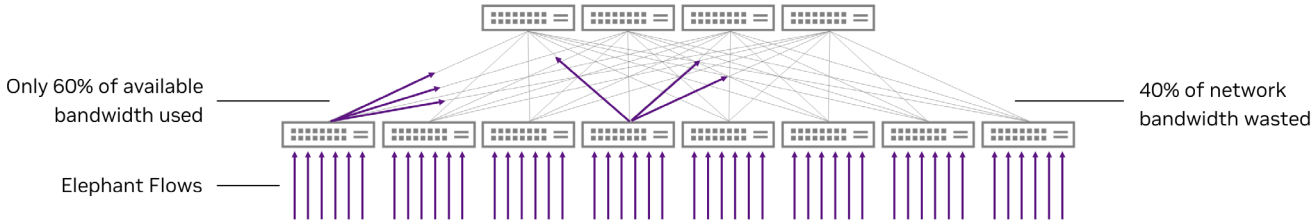


RDMA enables peripheral PCIe devices, such as NVIDIA BlueField, to access GPU memory directly. GPUDirect RDMA, designed specifically for GPU acceleration, provides direct communication between NVIDIA GPUs across remote systems, bypassing system CPUs and eliminating the need for data buffer copies via the system memory. When run over RoCE (RDMA over Converged Ethernet), GPUDirect RDMA achieves optimal performance, when it is deployed on a lossless network, which is critical to make it work reliably.

RoCE Adaptive Routing is Required to Avoid Congestion

A key attribute of scalable IP networks is their ability to distribute massive amounts of traffic and flows across multiple hierarchies of switches.

Figure 8. AI workloads reduced to 60 percent of max throughput with static load balancing



In a perfect world, data flows are completely uncorrelated and are therefore well distributed, smoothly load-balanced across multiple network links. This method relies on modern hash and multipath algorithms, including Equal-Cost MultiPath (ECMP). Modern data centers built on high port count, fixed form factor switches leverage ECMP extensively to build extremely large networks.

However, there are many cases where ECMP flow-based hashing does not work, often including ubiquitous modern workloads like AI and storage.

The issue is the problem of limited entropy and related hash collisions where too many elephant flows are sent on the same path. Entropy refers to the measure of randomness of data within a network packet or network flow. It provides an indication of the amount of information or variability present in the protocol headers.

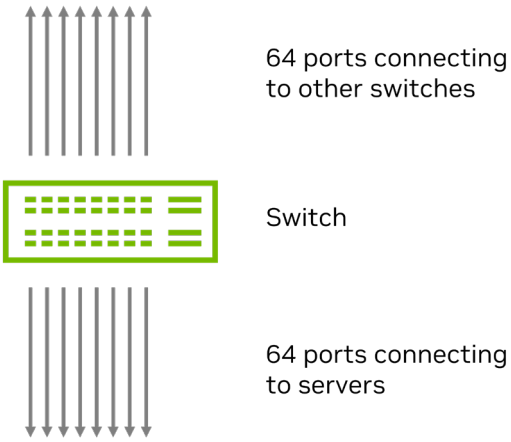
Traditional cloud services generate thousands of flows that are randomly connected to clients around the globe, which gives the cloud service network high entropy. However, AI and storage workloads tend to generate very few, but very large flows. These large AI flows dominate the bandwidth usage per link, significantly reducing the total number of flows and resulting in very low entropy for the network. This low entropy traffic pattern, also known as an elephant flow distribution, is typical with AI and high-performance storage workloads.

Entropy Matters

With traditional IP routing, to achieve load-balancing across all equal cost paths available and to avoid out of order packets, switches use per-flow (usually 3- or 5- tuple) static hashing. This requires high entropy to spread traffic evenly across multiple links without congestion. However, in elephant flow scenarios, when entropy is low while flow size is significant (common in AI networking), multiple flows can be hashed onto the same link, oversubscribing the link. This oversubscription results in congestion, increased latency, packet loss, retransmission, and eventually leads to poor application performance.

For many applications, performance is dictated not only by the average bandwidth of the network but also by the distribution of flow completion time. Long tails or outliers in the completion time distribution will decrease application performance significantly. The following is a common network topology found in AI networks that will most likely experience performance degradation due to long tail latency.

Figure 9. Why even non-blocking networks can experience long tail latency

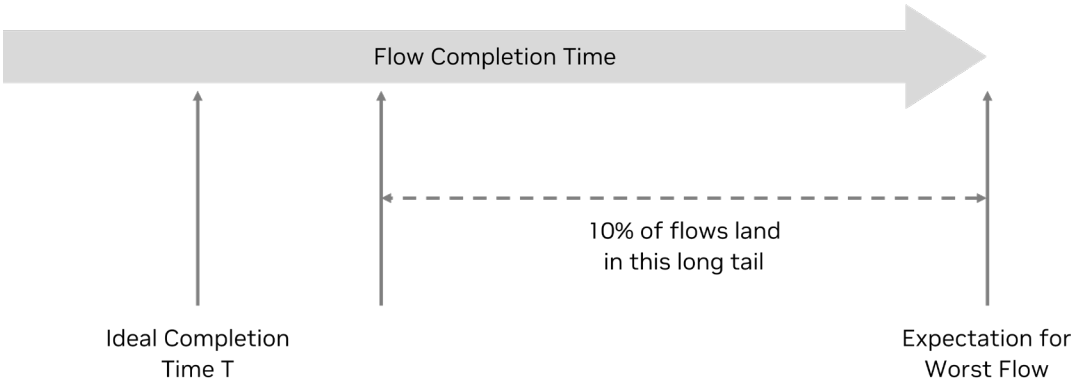


The example in Figure 8 consists of a single Leaf switch with 128 ports of 400GbE.

- 64 ports are 400GbE downstream ports connecting to servers
- 64 ports are 400GbE upstream ports connecting to spine switches
- Each downstream port receives traffic of four flows of equal bandwidth: 100G per flow for a total of 256 flows
- All traffic distribution across uplink ports is handled through ECMP flow-based hashing

As bandwidth levels increase, the likelihood of congestion increases, along with the associated increase in flow completion times. In the worst-case scenario, flows can take up to 2.5X longer to complete compared to the ideal (Figure 9).

Figure 10. Flow completion times can vary significantly



In this case, a few ports are congested while others are unused. The last-flow (worst flow) expected duration is 250 percent of the expected first-flow duration. That is, there is a long tail of flows where the completion time is longer than expected. To avoid congestion with high confidence (98 percent), you must reduce the bandwidth of all flows to under 50 percent.

Many flows suffer from high completion times because static ECMP hashing is not bandwidth-aware so some ports on the switch are highly congested while some others are underutilized. As flows finish transmission and release some port bandwidth, the lagging flows pass through the same congested ports, causing more congestion. This is because forwarding is static after the header has been hashed.

NVIDIA RoCE Adaptive Routing Ensures Load Balancing

RoCE Adaptive Routing is available on Spectrum-X Ethernet switches. With adaptive routing, traffic forwarding to an ECMP group selects the least-congested port for transmission. Congestion is evaluated based on egress queue loads, ensuring that an ECMP group is well-balanced regardless of the entropy size. An application that issues multiple requests to several servers receives the data with minimal time variation.

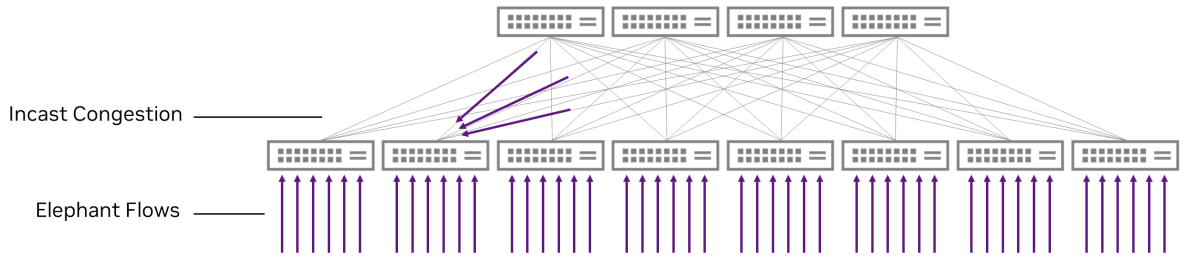
For every packet forwarded to an ECMP group, the switch selects the port with the minimal load over its egress queue. The queues that are evaluated are those that match the packet's traffic class. As different packets of the same flow travel through different paths of the network, they may arrive out of order to their destination. At the RoCE transport layer, the Spectrum-X Ethernet SuperNIC takes care of the out-of-order packets and forwards the data to the application in order. This renders the magic of RoCE Adaptive Routing invisible to the application benefiting from it.

On the sender side, the SuperNIC can dynamically mark traffic for eligibility to packet reordering, thus ensuring that inter-packet ordering can be enforced when required. The switch adaptive routing classifier can classify only these marked RoCE packets to be subjected to its unique forwarding.

NVIDIA RoCE Congestion Control is Required

When multiple applications are running simultaneously on AI factories, they experience subpar performance and inconsistent run times due to congestion at the network level. This can be triggered either by the network traffic from the application itself or by the background network traffic from other applications. The main cause of this type of congestion is referred to as many-to-one, or incast congestion, characterized by multiple data transmitters and a single data recipient.

Figure 11. Incast requires NVIDIA RoCE Congestion Control



Unfortunately, RoCE Adaptive Routing will not solve incast congestion since there is no way to route around it. Spectrum-X Ethernet can solve incast congestion using NVIDIA RoCE Congestion Control.

Key Features of the NVIDIA Spectrum-X Ethernet Networking Platform

AI workloads are characterized by a small number of elephant flows responsible for the large data movement between GPUs, where the tail latency highly impacts the overall application performance. Catering to such traffic patterns with traditional network routing mechanisms can lead to inconsistent and underutilized GPU performance for AI workloads. Spectrum-X Ethernet RoCE Adaptive Routing is a fine-grained load balancing technology. It dynamically reroutes RDMA data to avoid congestion and provide optimal load balancing to achieve the highest effective data bandwidth.

How NVIDIA RoCE Adaptive Routing Works

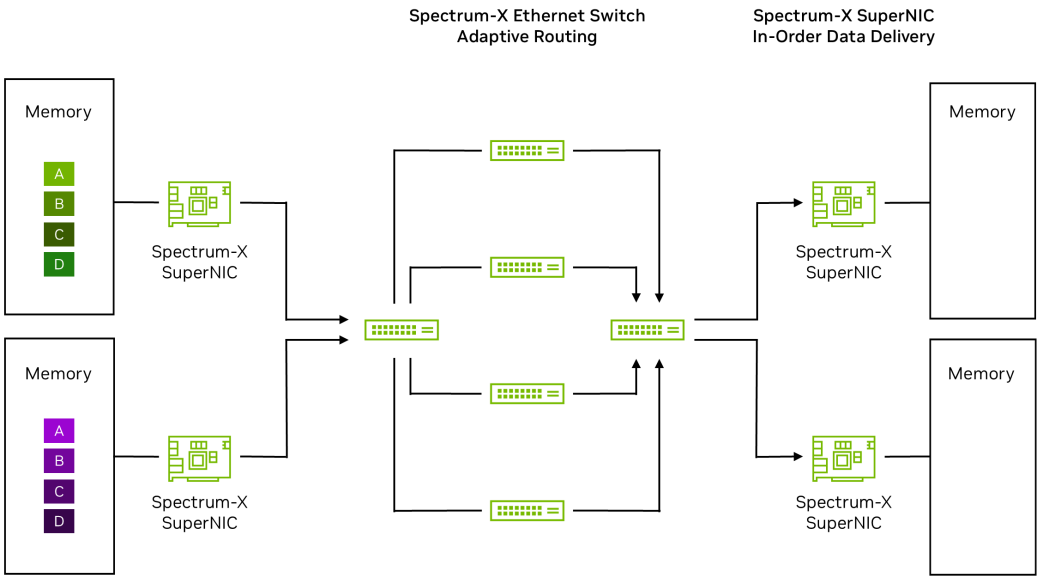
RoCE Adaptive Routing is an end-to-end capability enabled by Spectrum-X Ethernet switches and SuperNICs. Spectrum-X Ethernet switches are responsible for selecting the least congested port for data transmission on a per-packet basis. As different packets of the same flow travel through different paths of the network, they may arrive out-of-order to their destination. The Spectrum-X Ethernet SuperNIC transforms any out-of-order data at the RoCE transport layer, transparently delivering in-order data to the application.

Spectrum-X Ethernet switches evaluate congestion based on egress queue loads, ensuring all ports are well-balanced. Spectrum-X Ethernet switches employ packet spraying, selecting the port for every network packet with the minimal load over its egress queue. Spectrum-X Ethernet switches also receive status notifications from neighboring switches, which can influence the forwarding decision. The queues evaluated match with the appropriate traffic class. As a result, Spectrum-X Ethernet enables up to 95 percent effective bandwidth across the hyperscale system at load, and at scale.

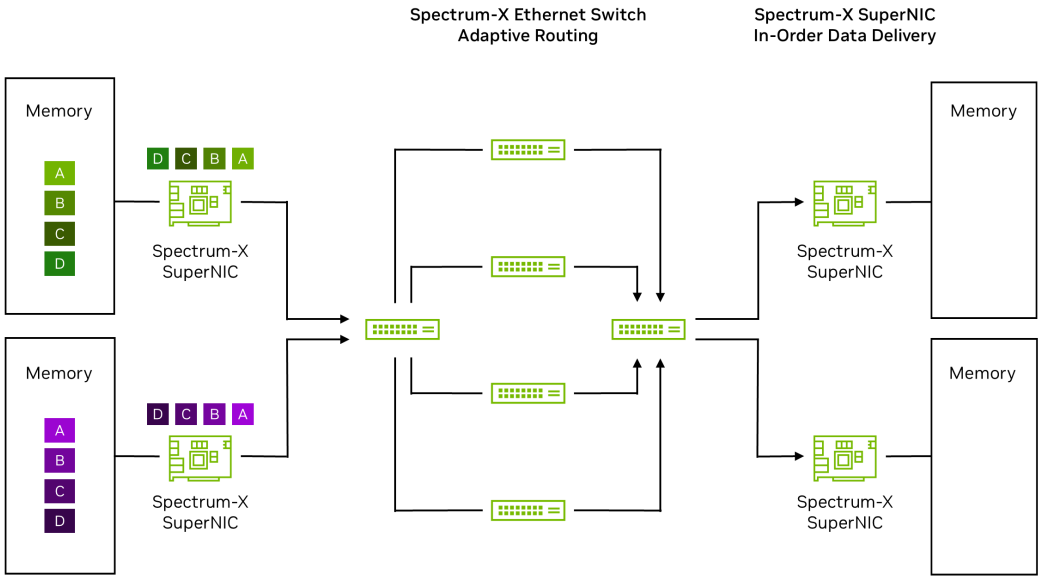
NVIDIA RoCE Adaptive Routing with NVIDIA Direct Data Placement

The following packet-level walkthrough shows how AI flows move between GPUs across a Spectrum-X Ethernet network. It describes how RoCE Adaptive Routing on the Spectrum-X Ethernet switch works in tandem with the NVIDIA Direct Data Placement (DDP) Technology on the Spectrum-X Ethernet SuperNIC:

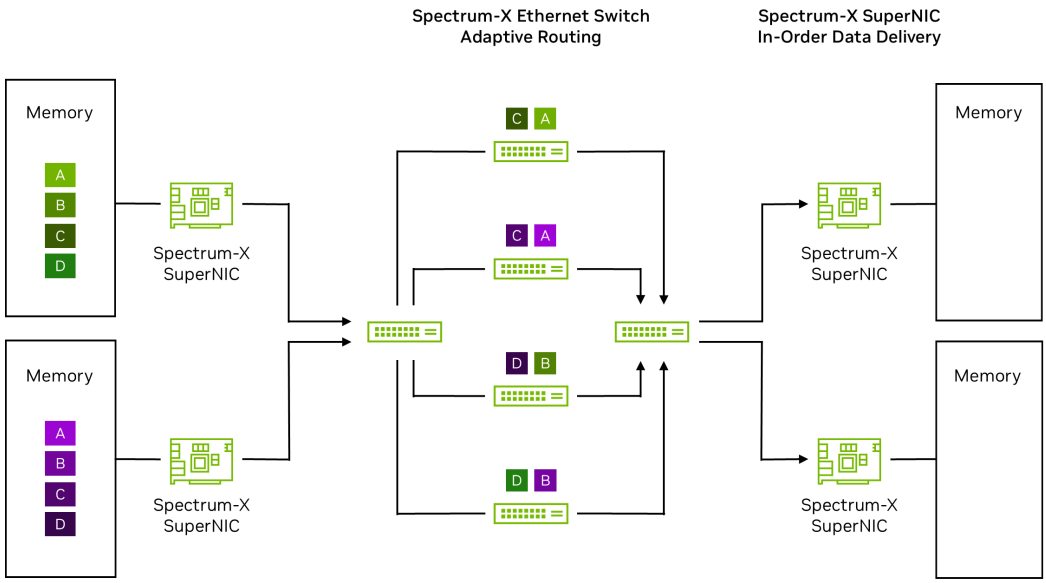
Step 1: Data is sent from server or GPU memory on the left side of the diagram, destined for servers on the right side of the diagram.



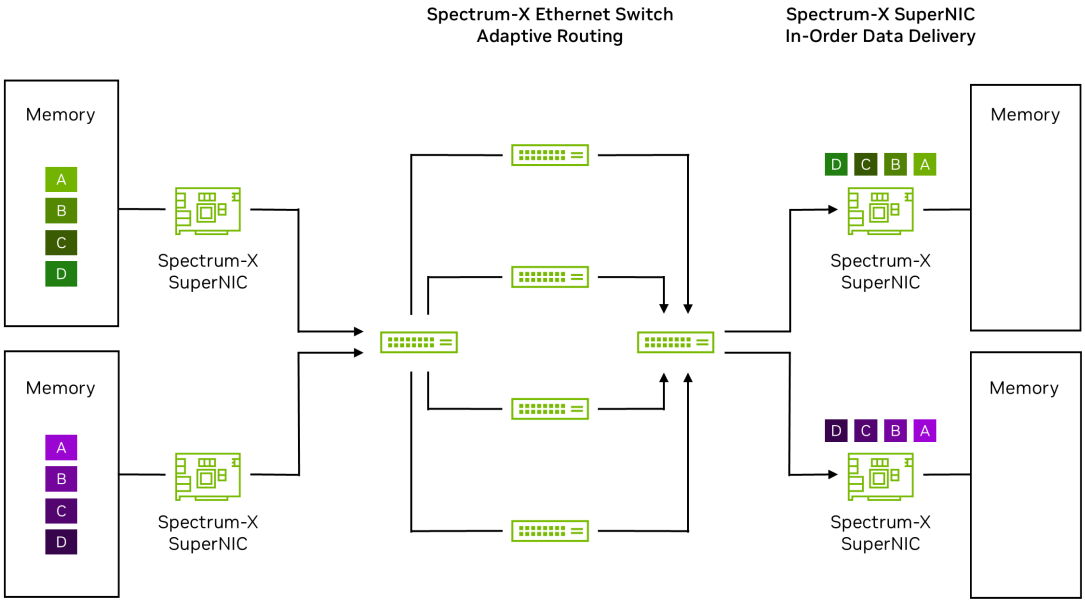
Step 2: The Spectrum-X Ethernet SuperNICs place the data into packets and send those packets to the first Spectrum-X Ethernet leaf switch, while marking these packets as safe for RoCE Adaptive Routing.



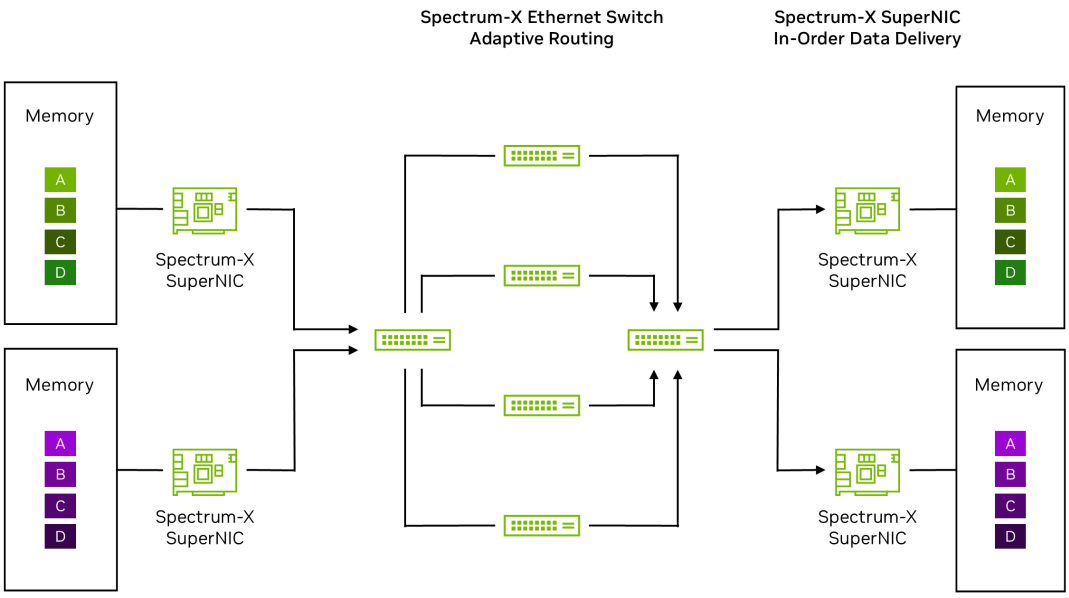
Step 3: The Spectrum-X Ethernet leaf switch on the left uses RoCE Adaptive Routing to load balance the packets from the green and purple data flows across the 4 spine switches, sending packets from each flow to multiple spines. This improves effective bandwidth from 60 percent on off-the-shelf (OTS) Ethernet to 95 percent with Spectrum-X Ethernet.



Step 4: Some of the packets may arrive out of order when they reach the Spectrum-X Ethernet SuperNICs on the right side.



Step 5: The Spectrum-X Ethernet SuperNICs on the right-side use NVIDIA Direct Data Placement (DDP) Technology to place the data in the correct order in the host/GPU memory.

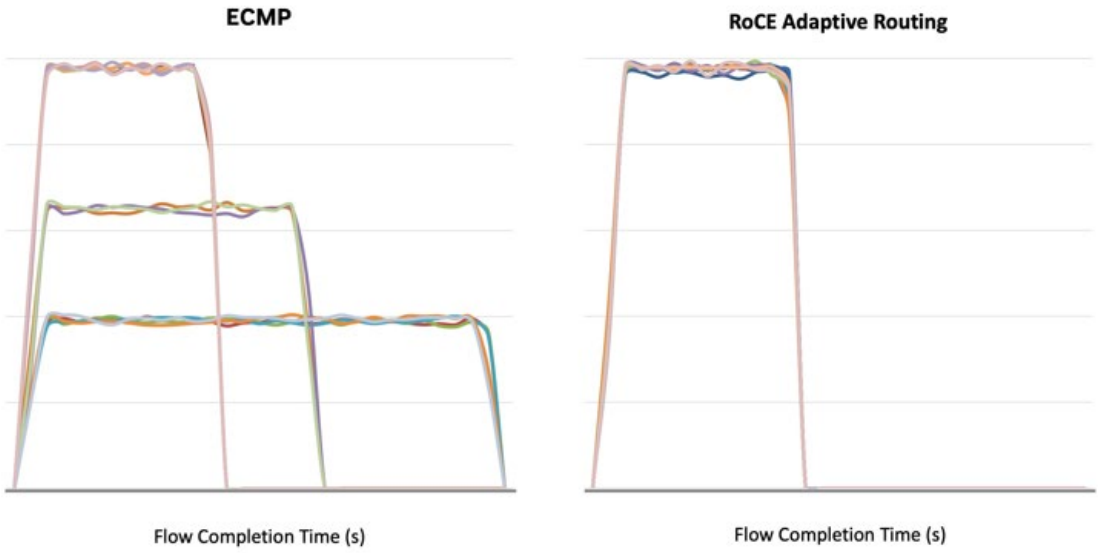


RoCE Adaptive Routing Results

To verify the effect of RoCE Adaptive Routing, we started by testing simple RDMA write test applications. In these tests, we divided the hosts into pairs, and each pair sent each other large RDMA write flows for a long period of time.

Figure 11 shows that the static, hash-based forwarding suffered from collisions on the uplink ports, increased flow completion times, reduced bandwidth, and reduced fairness between the flows. All problems were solved after moving to adaptive routing.

Figure 12. Flow completion time is reduced with RoCE Adaptive Routing



In the ECMP graph, certain flows had the same bandwidth and completion times, while others collided, leading to longer completion times and lower bandwidth. Specifically, in the ECMP scenario, some flows achieved an optimal completion time T of 13 seconds, while the slowest flows took 31 seconds, roughly 2.5X times T longer than the optimal time. In the RoCE Adaptive Routing graph, all flows were completed at approximately the same time, with similar peak bandwidth.

RoCE Adaptive Routing Summary

In many cases, ECMP flow-based hashing leads to high congestion and variable flow completion time, reducing application performance.

Spectrum-X Ethernet RoCE Adaptive Routing solves this issue by increasing the network's good throughput ("goodput"), minimizes the variability of the flow completion time and, as a result, boosts the application performance. By combining RoCE Adaptive Routing with NVIDIA Direct Data Placement (DDP) on Spectrum-X Ethernet SuperNICs, the technology remains transparent to applications. This integration ensures that the NVIDIA Spectrum-X Ethernet platform delivers the accelerated Ethernet performance needed for maximum data center operations.

Performance Isolation Using NVIDIA RoCE Congestion Control

Applications running concurrently on AI factories may experience degraded performance and unpredictable run times due to network-level congestion. This can be caused by the application's own network traffic or by background network traffic from other applications. The primary reason for this congestion is known as many-to-one congestion, involving multiple data senders and a single data receiver.

Such congestion cannot be solved with RoCE Adaptive Routing and instead requires data-flow metering per endpoint. Spectrum-X Ethernet RoCE Congestion Control is an end-to-end technology where the Spectrum-X Ethernet switch provides real-time network telemetry information on real-time congestion data. This telemetry information is processed by Spectrum-X Ethernet SuperNICs, which manage and control the data sender's data injection rate, to maximize network sharing efficiency. Without this congestion control, many-to-one scenarios will cause network back-pressure, congestion spreading, or even packet drops, dramatically degrading network and application performance.

In managing congestion, Spectrum-X Ethernet SuperNICs execute the congestion control algorithm, handling millions of congestion control events per second with microsecond reaction latency and applying fine-grained rate decisions. Spectrum-X Ethernet switch in-band telemetry holds both queuing information for accurate congestion estimation, as well as port utilization indication for fast recovery. NVIDIA's congestion control approach significantly improves congestion discovery and reaction times by allowing telemetry data to bypass queueing delays in congested flows, while still providing accurate and concurrent telemetry.

The following example shows how a network can experience many-to-one congestion, and how Spectrum-X Ethernet uses flow metering and in-band telemetry for RoCE Congestion Control.

Figure 13. Network congestion creating a victim flow

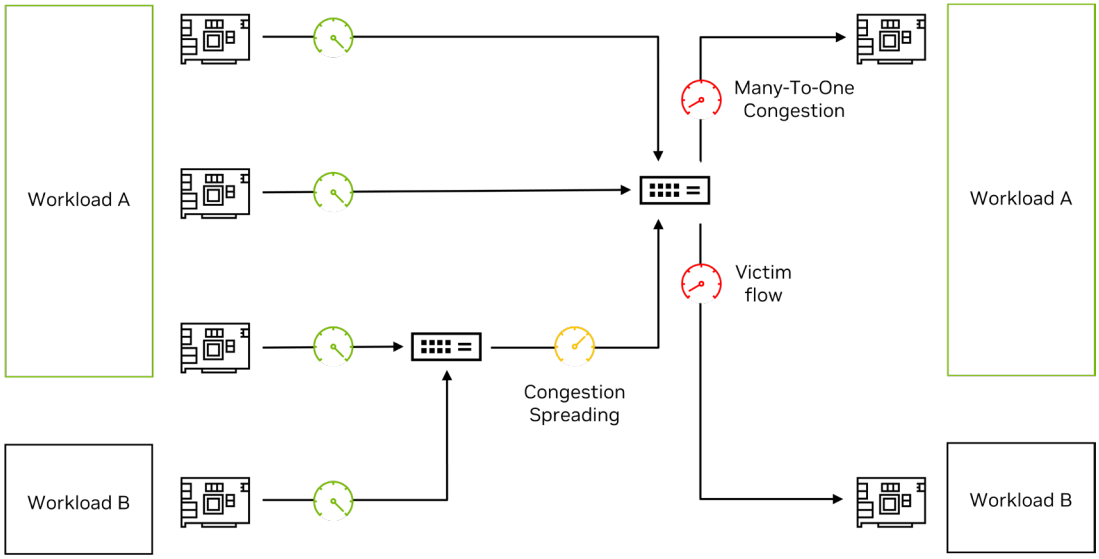


Figure 13 illustrates network congestion creating a victim flow. Four Source SuperNICs transmit data to two Destination SuperNICs. Sources 1, 2, and 3 transmit to Destination 1, with each receiving ⅓ of the available link bandwidth. Source 4 transmits to Destination 2 through the same leaf switch as Source 3, and therefore should receive ⅔ of the available link bandwidth.

Without congestion control, Sources 1, 2, and 3 create 3-to-1 congestion as they all send data to Destination 1. This congestion creates back pressure, spreading from Destination 1 to the leaf switch connected to Source 1 and 2. Source 4 becomes a victim flow, with its throughput to Destination 2 reduced to 33 percent of available bandwidth (50 percent of expected performance). This adversely affects AI application performance, which relies on both average and worst-case performance.

Figure 14. Spectrum-X Ethernet solves congestion with flow metering and congestion telemetry

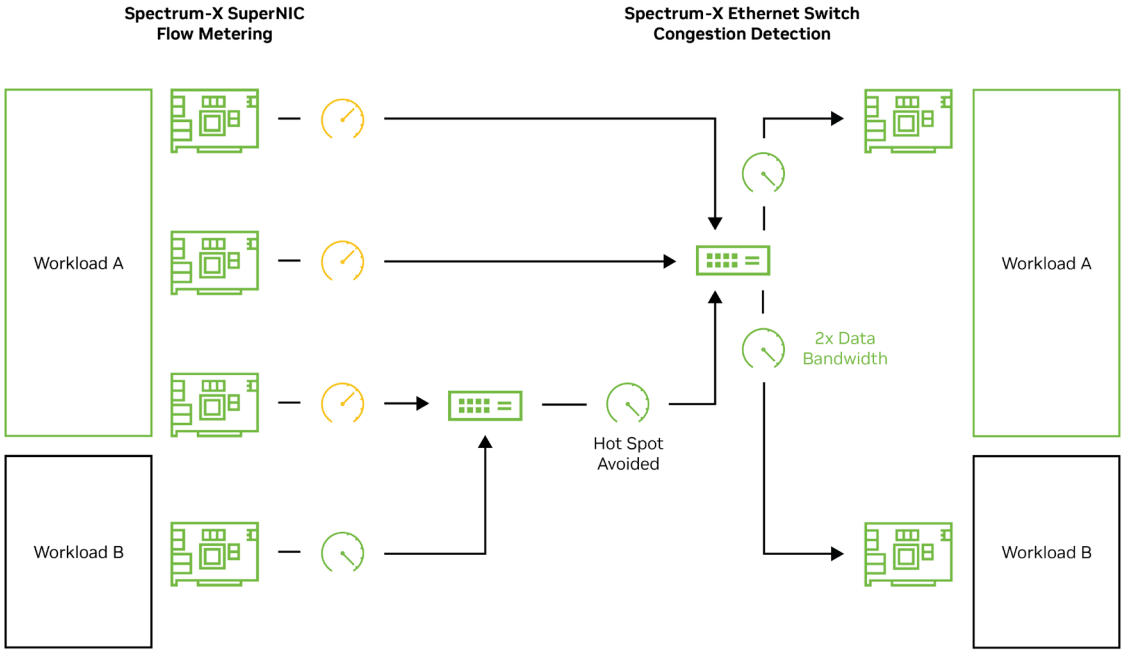


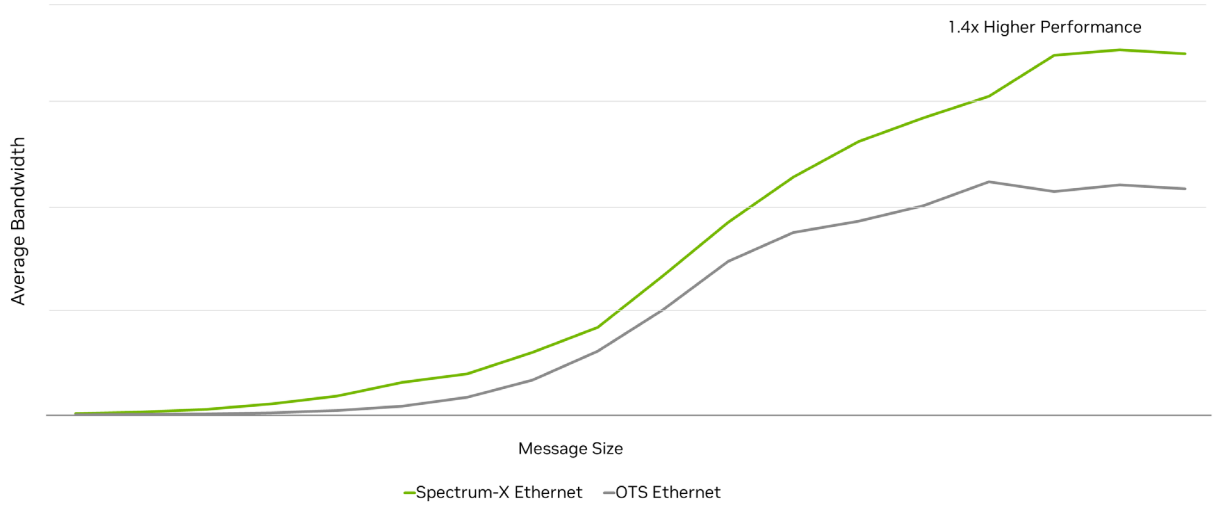
Figure 14 shows how Spectrum-X Ethernet solves the congestion problem in Figure 13. This diagram has the same setup: Four Source SuperNICs send data to two Destination SuperNICs. In this case, congestion at the leaf level is avoided by flow metering at Source 1, 2, and 3. This eliminates the back pressure that impacts Source 4. Source 4 receives its full ⅔ effective bandwidth as expected. In addition, the Spectrum-X Ethernet switch uses the in-band telemetry generated via What Just Happened to dynamically re-allocate flow paths and queue behavior.

RoCE Performance Isolation

AI factories need to support a large number of users (tenants) and parallel applications or workflows. These users and applications generate traffic, competing for the infrastructure's shared resources, such as the network, and therefore may impact performance.

In addition, enhancing the AI network to increase performance with the NVIDIA Collective Communication Library (NCCL) requires coordination and synchronization across all workloads running simultaneously in the AI factory. For AI applications, the traditional benefits of cloud computing, such as elasticity and high availability, are more limited. Consequently, performance degradation becomes a more significant and global concern.

Figure 15. Spectrum-X Ethernet mechanisms involved in delivering noise isolation

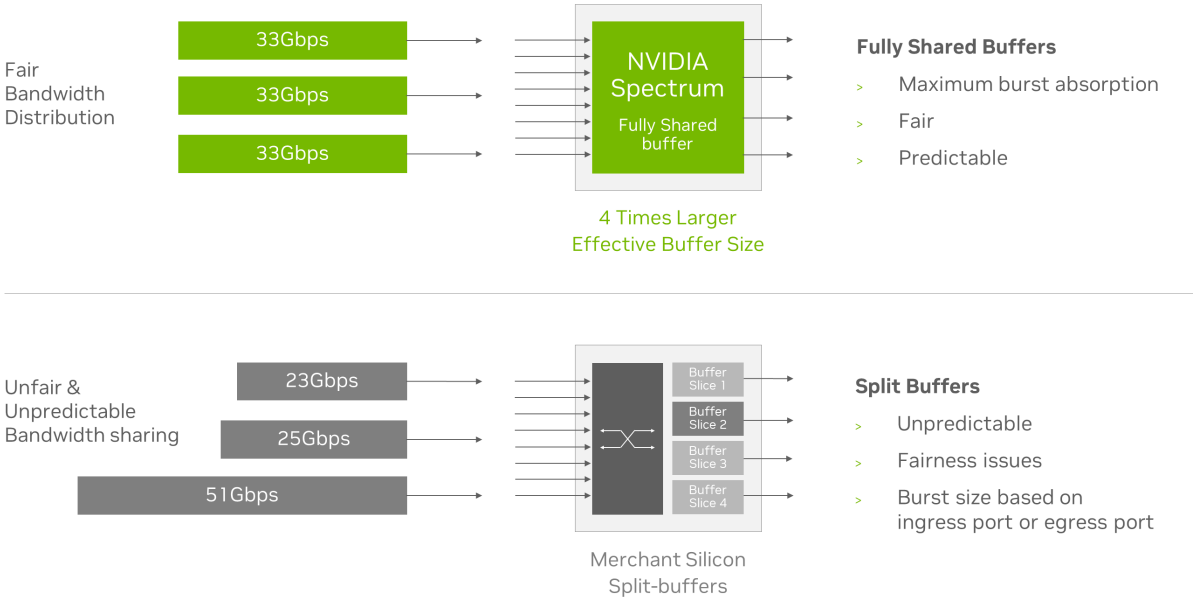


The Spectrum-X Ethernet platform includes mechanisms, which together deliver performance isolation, ensuring that one workload does not adversely impact the performance of another. These Quality-of-Service mechanisms prevent any workload from generating network congestion that could impede the data movement of another workload.

RoCE Adaptive Routing contributes to this isolation by finely balancing data paths, which avoids collision of flows traversing the leaf and spine. RoCE Congestion Control further supports isolation through metering and telemetry to prevent the formation of victim flows due to many-to-one host traffic.

Moreover, Spectrum-X Ethernet switches improve performance isolation through their universal shared buffer design. This design delivers bandwidth fairness across different-sized flows, workload protection from “noisy neighbor” flows, and larger microburst absorption when multiple flows converge on the same destination port.

Figure 16. Spectrum-X Ethernet switches feature a universal shared buffer design



NVIDIA Full-Stack Optimization

An AI factory is a precision-engineered machine whose performance can be negatively impacted by seemingly minor network events, such as a link or device failure, that might be overlooked on a cloud or Enterprise control network. A single sluggish system can cause a ripple effect, slowing down the entire AI factory. Moreover, one jabbering SuperNIC can disrupt adjacent devices if the network lacks a vigilant monitoring system capable of detecting and mitigating such behavior.

NVIDIA follows a multi-stage process to test, certify, and tune the Spectrum-X Ethernet components as part of a full-stack AI solution:

1. **AI Performance Benchmark Testing:** NVIDIA tests the overall AI performance of a typical cluster. It then publishes the performance results so others can know of the expected performance level of an AI cluster built with the Spectrum-X Ethernet network.
2. **Component Testing:** NVIDIA first tests each component (switch, SuperNIC, GPU, and AI libraries) individually to ensure they both function as expected and meet performance benchmarks.
3. **Integration Testing:** After verifying the functionality of individual components, NVIDIA integrates them to create a cohesive AI solution. The integrated system undergoes a series of tests to ensure compatibility, interoperability, and seamless communication between the components.
4. **Performance Tuning:** Once the components are integrated and functioning as a unit, NVIDIA focuses on optimizing the performance of the full-stack solution. This

involves adjusting parameters, identifying bottlenecks, and fine-tuning configurations to maximize the effective bandwidth of the network.

5. **Overall System Performance**: During this stage, NVIDIA also specifically tunes buffers and congestion threshold points to cater to AI workloads like GPT, BERT, and RetinaNet, ensuring optimal performance for these popular deep learning models.
6. **Library and Software Optimization:** NVIDIA optimizes AI libraries, such as NCCL, to ensure efficient communication between the GPUs and other components. This step is crucial in minimizing latency and maximizing throughput in large-scale deep learning applications.
7. **Certification:** After testing and tuning the full-stack AI solution, NVIDIA performs a series of certifications to ensure the system performs reliably and securely. This process includes stress testing, security testing, and validating compatibility with popular AI frameworks and applications.
8. **Real-World Testing:** Finally, NVIDIA deploys the full-stack AI solution in real-world scenarios to evaluate its performance under various conditions and workloads. This step helps identify any unforeseen issues and ensures the solution is ready for widespread adoption by customers.

By following this comprehensive process, NVIDIA ensures the robustness, reliability, and high performance of our full-stack AI solutions, delivering a seamless experience for their customers, especially for widely used AI workloads like GPT, Llama, and DeepSeek..

Spectrum-X Ethernet Performance Results from Israel-1

NVIDIA's full stack optimizations result in significant performance benefits for AI workloads. To fully tune and optimize Spectrum-X Ethernet recommended configurations, NVIDIA uses Israel-1, a generative AI supercomputer that is the largest system in Israel. Consisting of 4 scalable units of HGX-H100 systems featuring B3140 SuperNICs connected by SN5610 Ethernet switches in 3 tiers, Israel-1 was designed as a blueprint for large-scale AI factories that can be replicated by NVIDIA customers and partners to achieve identical performance for their own workloads.

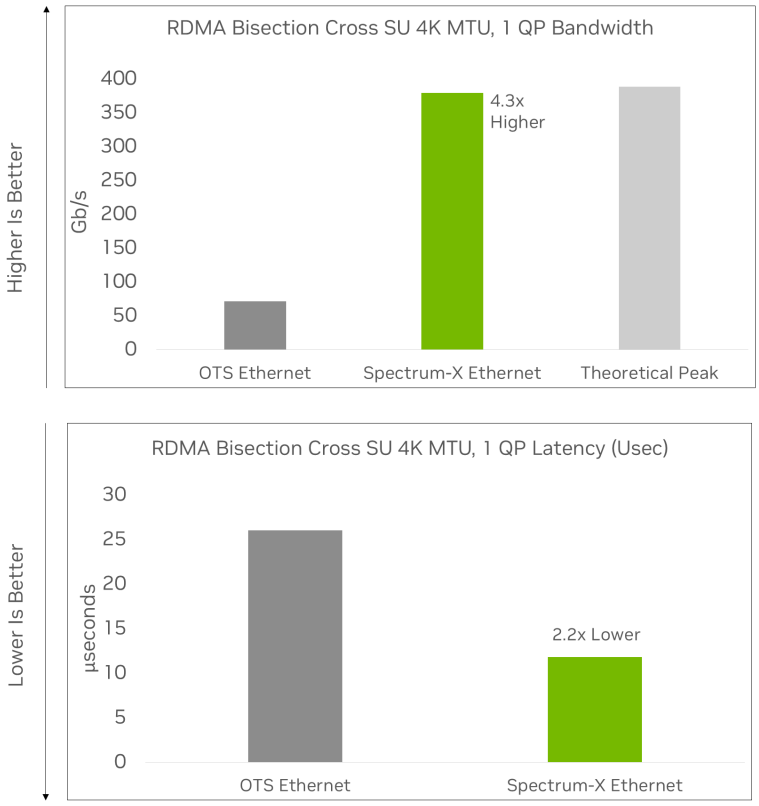
The following benchmarks compare the performance of a fully optimized Spectrum-X Ethernet network to off-the-shelf (OTS) Ethernet networks optimized for traditional data centers. This off-the-shelf (OTS) Ethernet network leverages RoCE and standard Ethernet flow control techniques, but not the RoCE Extensions and NCCL-aware network optimizations introduced with Spectrum-X Ethernet.

RDMA Fabric Performance

AI workloads and their network traffic patterns can differ significantly from model to model, but all models require an efficient, reliable, and highly performant RoCE network.

The first benchmark result shows the effective bandwidth and latency of RDMA bisection. Spectrum-X Ethernet delivered significantly better performance (higher effective bandwidth and lower latency) compared to off-the-shelf (OTS) Ethernet.

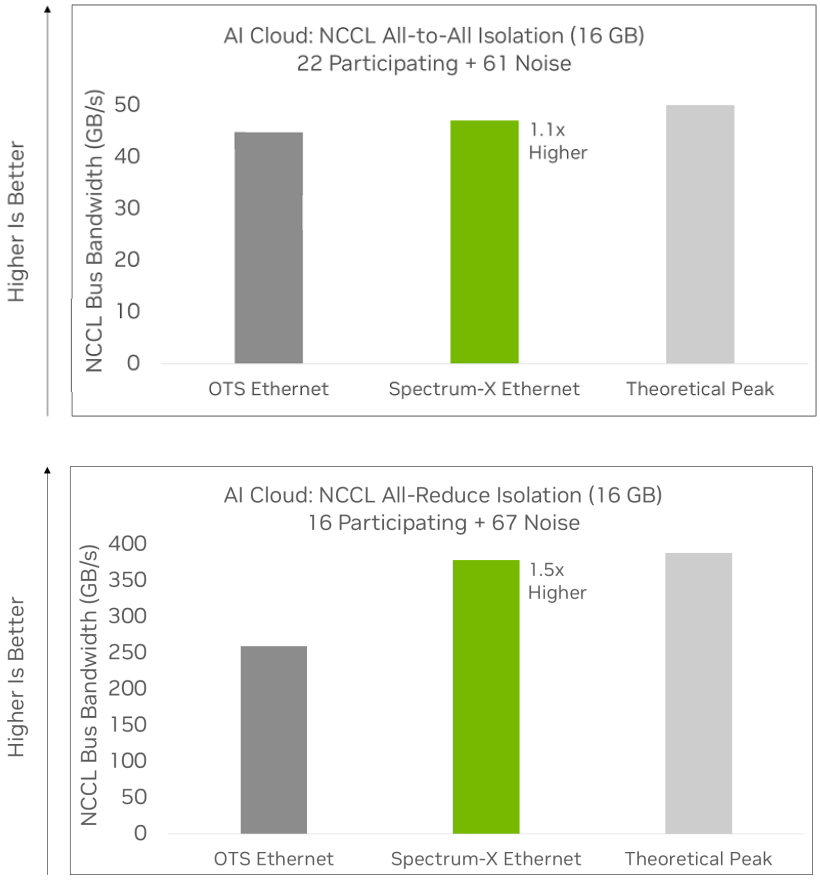
Figure 17. Spectrum-X Ethernet achieves higher effective bandwidth and lower latency for RDMA bisection.



NCCL Collective Performance

NCCL collectives such as all-reduce and all-to-all are essential to distributed AI performance and are the atomic operations of AI training. Spectrum-X Ethernet is designed as a NCCL-aware fabric, which ensures proper tuning for these operations. The following benchmark results show NCCL operation performance in an AI factory scenario where multiple AI jobs are running at the same time and creating noise. Spectrum-X Ethernet delivers higher NCCL bus bandwidth for both all-reduce and all-to-all while providing noise isolation, leading to consistent performance almost as high as performance measured in a “non-noisy” scenario (that is, where no other jobs are running simultaneously in the AI factory).

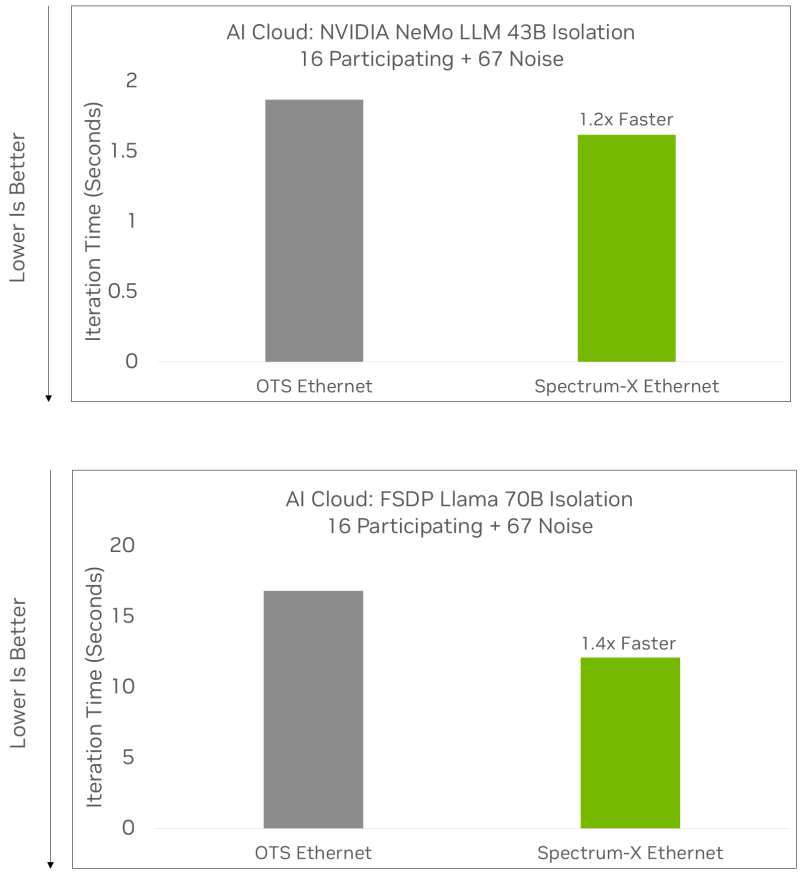
Figure 18. **Spectrum-X Ethernet delivers higher bus bandwidth for NCCL collectives running in an AI factory**



AI Model Performance

The most important objective of network optimization is to deliver the highest application-level performance. For AI workloads like LLM training, this means reducing iteration step times, which leads to shorter job completion times and faster time-to-value for the AI model. NVIDIA benchmarked AI factory training iteration times for several common large language models, including GPT-4, FSDP Llama 70B, and NeMo LLM 43B. In all cases, Spectrum-X Ethernet led to faster training times and showed consistent, predictable performance unaffected by noise.

Figure 19. **Spectrum-X Ethernet accelerates training AI workloads in noisy AI factory environments.**



Impact of Virtualization on Performance

As multi-tenant architectures, AI factories typically require network virtualization through methods such as EVPN VXLAN. These virtualization technologies are focused around optimizing for extensibility and scalability, but not performance. Without careful implementation, virtualization can introduce substantial performance penalties.

NVIDIA compared Spectrum-X Ethernet performance with and without virtualization via EVPN VXLAN. Across all tests, performance was almost identical in the single-tenant and the multi-tenant environments and showed minimal degradation, demonstrating Spectrum-X Ethernet as the ideal Ethernet fabric for multi-tenant AI factories.

Figure 20. Spectrum-X Ethernet performs equally well with and without network virtualization.

Data Center Scale-Across Performance

Scaling AI Factories within a single building or campus eventually runs into hard limits in power availability, cooling capacity, and physical footprint. To extend capacity, enterprises and cloud providers have looked to link multiple data centers together, but this approach has traditionally been constrained by networking challenges such as long-tail latency, jitter, and unpredictable performance, that can stall large-scale training.

Scale-across networking overcomes these barriers by enabling multiple data centers that span hundreds or even thousands of kilometers to operate as a single, unified compute fabric. Instead of functioning as isolated sites, distributed facilities become one massive AI factory, where large-scale training and inference seamlessly span locations without being slowed by latency, congestion, or inconsistent performance. The result is faster job completion, higher infrastructure utilization, and the ability to train the largest models by pooling resources across geographic boundaries.

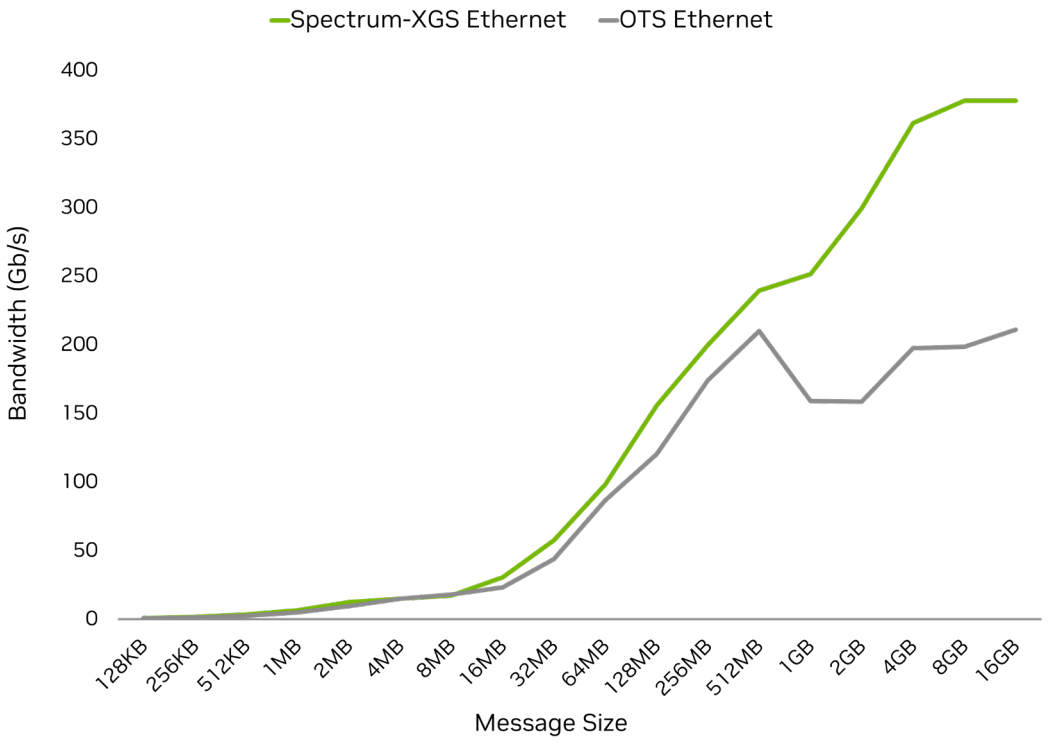
Spectrum-XGS Ethernet is a scale-across technology, seamlessly integrated into the Spectrum-X platform to address the power constraints of modern data centers. Built on industry-standard Ethernet protocol, it introduces:

- Distance-Adaptive Congestion Control – algorithms that factor in site-to-site distances to optimize flow control and minimize tail latency.

- Topology-Aware Load Balancing – traffic distribution across multi-site fabrics to maximize throughput and eliminate hotspots.
- Deterministic Performance – precise latency management with real-time telemetry for visibility, predictability, and control.

NVIDIA Spectrum-XGS Ethernet accelerates multi-site NCCL collectives by a factor of 1.9 compared to off-the-shelf Ethernet, allowing geographically distributed data centers to truly function as one.

Figure 21. 10km Distance NCCL All-Reduce performance between data centers



The Ethernet Designed for AI

The Spectrum-X Ethernet Networking Platform is purpose-built for demanding AI applications, offering a range of benefits over off-the-shelf (OTS) Ethernet. With higher performance, reduced power consumption, lower total cost of ownership, seamless full-stack software-hardware integration, and massive scalability, Spectrum-X Ethernet emerges as the ultimate Ethernet platform for both present and future AI workloads.

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation (“NVIDIA”) makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer (“Terms of Sale”). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer’s own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, BlueField, GPUDirect, MagnumIO GPUDirect Storage, and Spectrum are trademarks and/or registered trademarks of NVIDIA Corporation and/or its affiliates in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated. All other trademarks are property of their respective owners.

Copyright

© 2025 NVIDIA Corporation & Affiliates. All rights reserved.

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com