NVIDIA Spectrum-X Network Platform Architecture

The First Ethernet Network Designed to Accelerate AI Workloads

White Paper

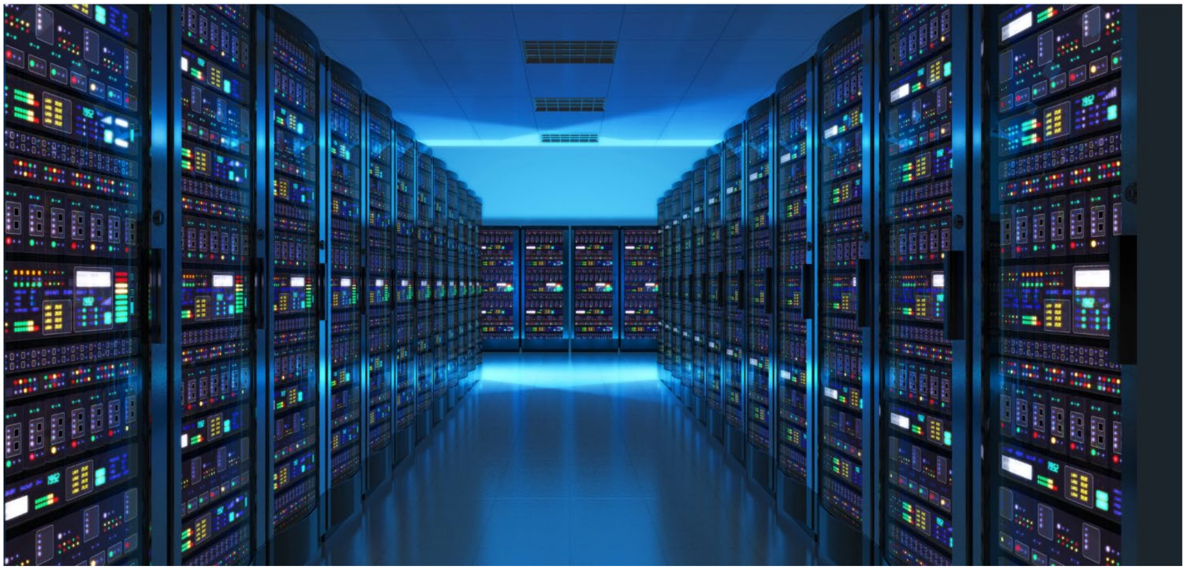
| May 2023

Table of Contents

List of Figures

Improve AI Performance and Efficiency

Cloud 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, which are expanding into Cloud Service Providers (CSP) and cloud-scale data centers.



While AI Factories are built for the highest levels of performance, capable of training trillion-parameter foundational models, they typically operate with just one or a few tenants. AI Factories require a highly optimized network that combines NVIDIA® NVLink™ high-speed GPU interconnect with NVIDIA Quantum InfiniBand for the highest levels of performance. Contrast this with clouds that are being built for AI (AI Cloud), which have several unique requirements, including the need to integrate into the Ethernet service network and management framework.

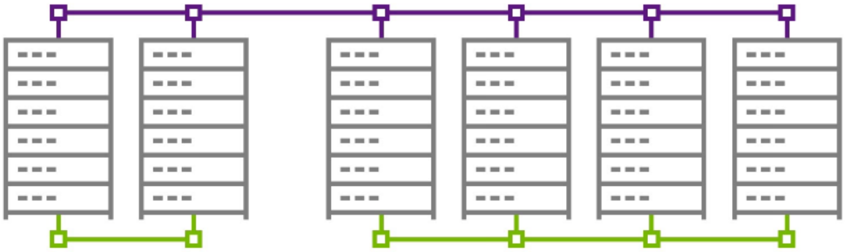
Key requirements specific to AI Clouds include:

- **AI in Software as a Service (SaaS) migration and AI as a Service:** These SaaS offerings are being adapted to include AI, along with AI as a Service, necessitating 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 enables seamless scaling from small clusters to large clusters and allows for the reprovisioning of compute and network resources into multiple small clusters.

- **Integration and Compatibility:** AI services from CSPs should be closely integrated with other cloud services provided by the same company. These services running on Ethernet are necessary for managing 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:** Provides support for open network operating systems, like SONiC.
- **Global Reach:** CSPs globally distribute data centers synchronized with PTP. This allows for AI model deployment closer to the end user while adhering to cross-border data regulations regarding sensitive data.

CSP networks have been optimized for shared cloud computing infrastructure environments and are sub-optimal for addressing needs of AI clouds. A few differences between control/user access network and AI fabric are outlined in Figure 1.

Figure 1. Differences between control/user access network and 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 clouds use traditional IP/Ethernet (hereafter referred to as Ethernet) as their 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, traditional Ethernet is unable to provide performance isolation, which is necessary to protect one tenant's AI jobs from negatively impacting others.

For those deploying AI clouds with Ethernet, NVIDIA created the Spectrum-X Networking Platform which improves performance, while increasing the predictability and power efficiency of Ethernet-based AI clouds.

NVIDIA Spectrum-X Networking Platform Overview

The NVIDIA® Spectrum™-X Networking Platform is the first Ethernet platform designed specifically to improve the performance and efficiency of Ethernet-based AI clouds. This breakthrough technology achieves 1.7X improved overall AI performance for massive AI workloads such as LLM, along with 1.7X better power efficiency and consistent, predictable performance in multi-tenant environments. Spectrum-X is built on network innovations based on the tight coupling of the Spectrum-4 Ethernet switch with the NVIDIA BlueField®-3 data processing unit (DPU). The delivery of end-to-end network capabilities purpose-built for AI workloads reduces run times of massive transformer-based generative AI models, and allows network engineers, data scientists, and cloud service providers to attain faster results and make informed decisions.

Inside The NVIDIA Spectrum-X 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 DPUs to the switches, cables/optics, networking, and acceleration software, to attain this primary goal. NVIDIA created a number on new innovations to achieve the highest effective bandwidth under load and at scale:

1. NVIDIA RoCE Adaptive Routing on Spectrum-4
2. NVIDIA Direct Data Placement (DDP) on BlueField-3
3. NVIDIA RoCE Congestion Control on both Spectrum-4 and BlueField-3
4. NVIDIA AI Acceleration Software
5. End-to-End AI Network Visibility

These innovations work together as part of a full-stack solution that is tested, tuned, and benchmarked by NVIDIA to guarantee the highest level of performance. When implemented as a full stack, there are number of benefits to the Spectrum-X Networking Platform.

Key Benefits of Spectrum-X

- **Improved AI Cloud Performance**: Spectrum-X enhances AI cloud performance by 1.7X.
- **Standard Ethernet Connectivity**: Spectrum-X is fully standards-based Ethernet and is completely interoperable with Ethernet-based stacks.
- **Increased Power Efficiency**: By improving performance, Spectrum-X 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 flows running across the AI cloud makes it possible to identify performance bottlenecks and is a key part of a modern, automated fabric-validation solution.
- **Higher AI Scalability**: Scales to 128X 400G ports in one hop or 8K ports in a two-tier leaf/spine topology, supporting the expansion of the AI cloud while maintaining high levels of performance.
- **Faster Network Setup**: The automated, end-to-end configuration of advanced networking functionality is fully tuned for AI workloads.

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

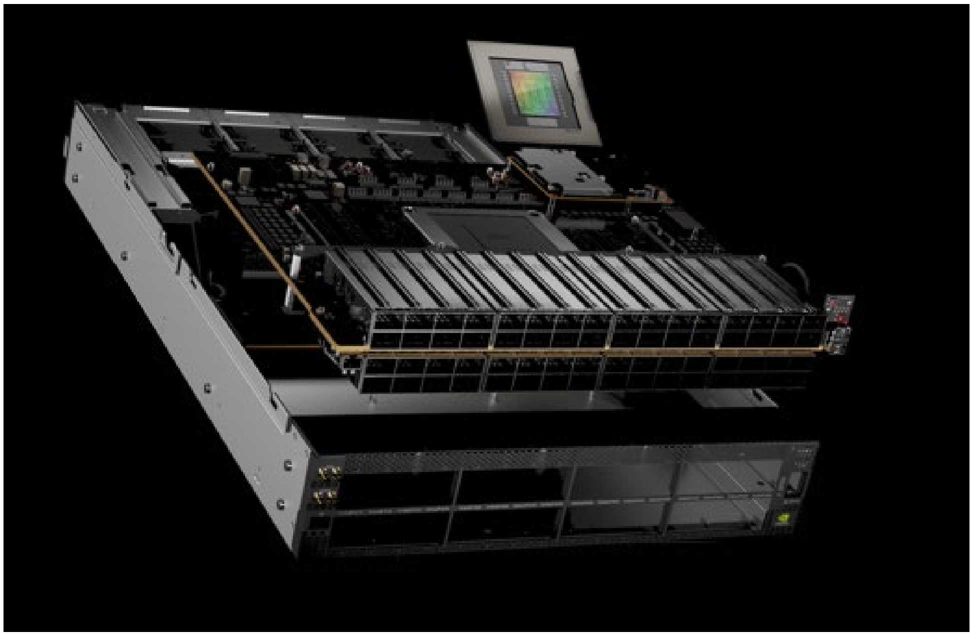
NVIDIA Spectrum-4 Ethernet Switches

Spectrum-4 switches are built on a 51.2Tbps ASIC, supporting up to 128 ports of 400 Gigabit Ethernet (GbE) in a single 2U switch. Spectrum-4 is the first switch designed from the ground up for AI workloads, combining specialized high-performance architecture and lowest latency with standard Ethernet connectivity.

Spectrum-4 offers 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

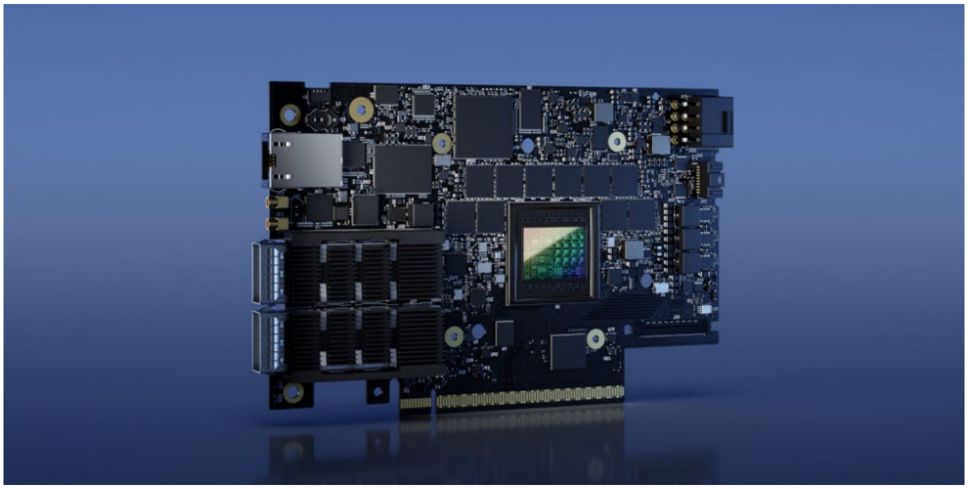
Figure 2. NVIDIA Spectrum-4 400 Gigabit Ethernet Switch



NVIDIA BlueField-3 DPU

The NVIDIA® BlueField®-3 DPU 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. With 400Gb/s Ethernet network connectivity, BlueField-3 DPU offloads, accelerates, and isolates software-defined networking, storage, security, and management functions in ways that profoundly improve data center performance, efficiency, and security. BlueField-3 support for multi-tenancy and security is a key requirement for handling North-South as well as East-West traffic in Cloud AI data centers powered by Spectrum-X.

Figure 3. NVIDIA BlueField-3 400Gb/s Ethernet DPU

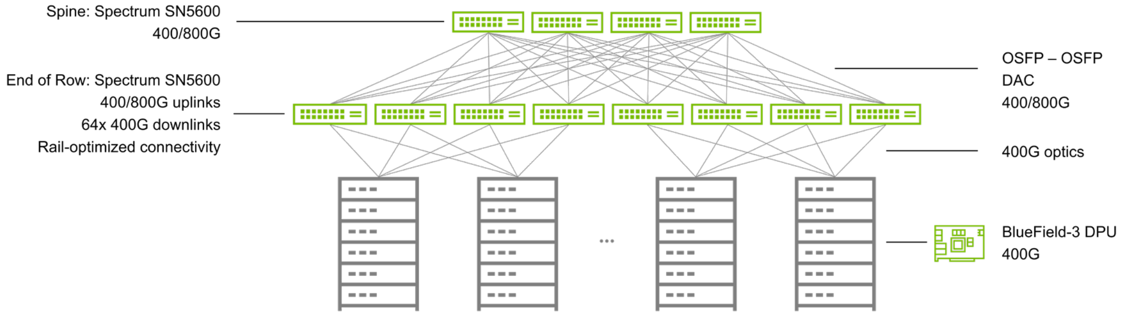


BlueField-3 has been purpose-built for AI acceleration, featuring an integrated all-to-all engine for AI, NVIDIA GPUDirect®, and NVIDIA® Magnum IO GPUDirect® Storage (GDS) acceleration technology. Additionally, it features a special network interface mode (NIC) mode that takes advantage of local memory to accelerate large AI clouds. These clouds consist of an extensive number of Queue Pairs that can be addressed locally instead of using system memory. Lastly, it includes NVIDIA Direct Data Placement (DDP) Technology which augments RoCE Adaptive Routing.

NVIDIA End-to-End Physical Layer (PHY)

Spectrum-X is the only Ethernet networking platform built on the same end-to-end 100G Serializer/Deserializer (SerDes) channels from switch to DPU to GPU. NVIDIA SerDes ensures exceptional signal integrity and the lowest BER (Bit Error Rate), greatly reducing the power consumption of the AI Cloud. This robust SerDes technology, leveraged by NVIDIA Hopper GPUs, as well as the Spectrum-4, BlueField-3, and Quantum InfiniBand portfolios, delivers power efficiency and performance without compromise.

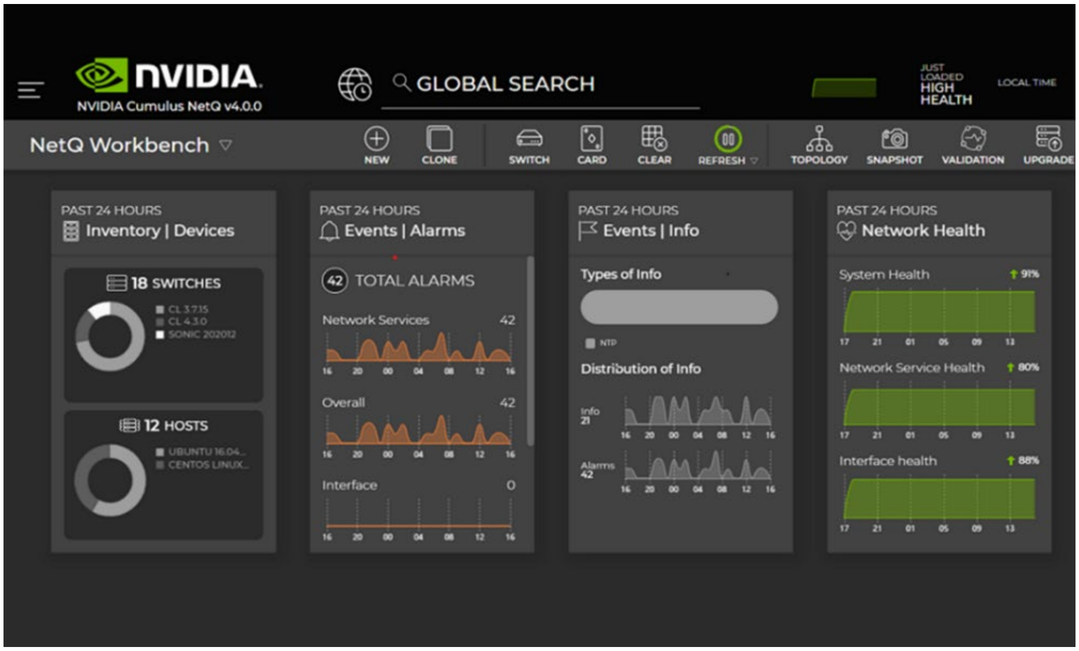
Figure 4. Typical Spectrum-X Network Topology



SerDes technology plays a vital role in modern data transmission, by enabling the conversion of parallel data to serial data and vice versa. Standardizing the employment of SerDes technology across all network devices and components in a network or system offers numerous advantages:

- **Cost and Power Efficiency:** The NVIDIA SerDes used in Spectrum-X is optimized for power efficiency and removes the need for gearboxes in the network, which are used to bridge differences between lane speeds. The usage of gearboxes not only adds complexity to the data path but also additional cost and power consumption. Eliminating the need for these gearboxes reduces both the initial investment and the operational costs associated with power usage and cooling.
- **System Design Efficiency**: Unifying Best-in-Class SerDes technology across data center infrastructure leads to better signal integrity, requiring fewer system components and simplifying system design. Unifying the same SerDes technology also makes operations simpler and improves uptime.

Figure 5. 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 switch and DPU, 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 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

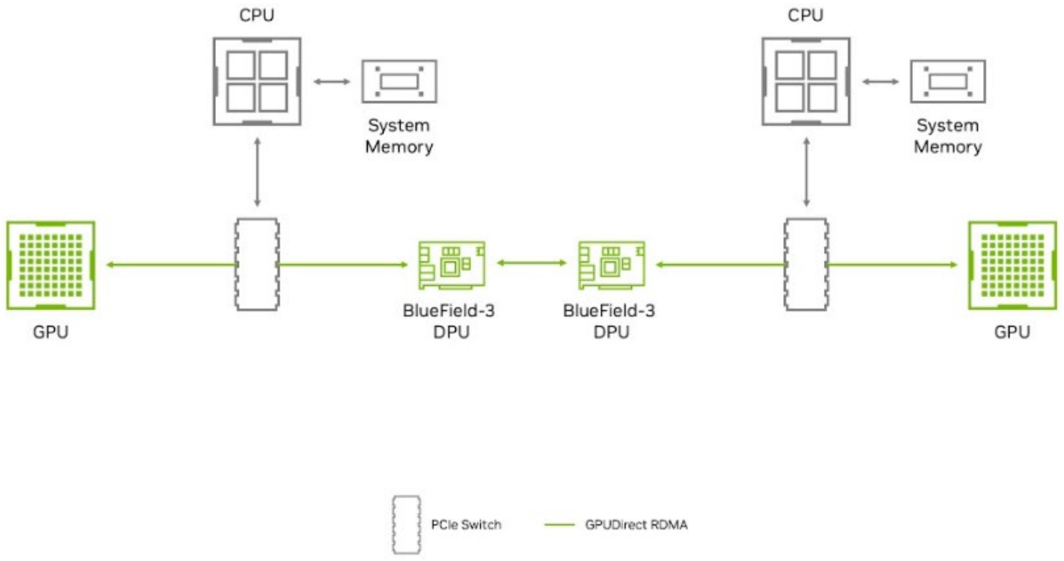
NVIDIA® DOCA™ is the key to unlocking the potential of the NVIDIA BlueField® DPU to offload, accelerate, and isolate data center workloads. With DOCA, developers can program the data center infrastructure of tomorrow by creating software-defined, cloud-native, DPU-accelerated services with zero-trust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

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 RDMA. Remote Direct Memory Access (RDMA) is a networking technology that allows access from the memory of one computer into another computer without involving either computer's CPU.

Figure 6. RDMA enables BlueField with direct access to GPU memory

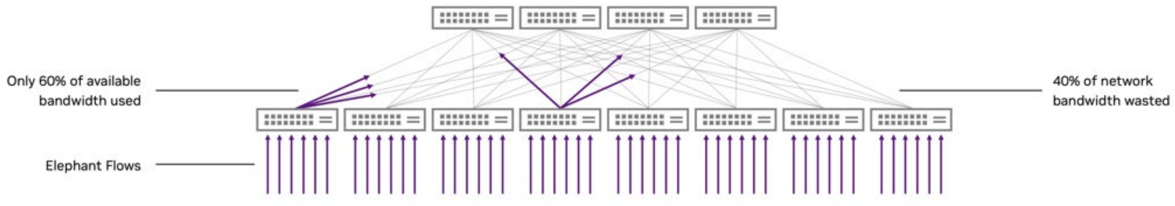


Remote direct memory access (RDMA) enables peripheral PCIe devices, such as NVIDIA BlueField, direct access to GPU memory. Designed specifically for the needs of GPU acceleration, GPUDirect RDMA provides direct communication between NVIDIA GPUs in remote systems. This eliminates the system CPUs and the required buffer copies of data via the system memory. GPUDirect RDMA, when run over RoCE (RDMA over Converged Ethernet), provides peak benefit, 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 7. AI workloads reduced to 60 percent of max throughput with static load balancing


Only 60% of available bandwidth used
40% of network bandwidth wasted
Elephant Flows

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 a network flow. It provides an indication of the amount of information or variability present in the protocol headers.

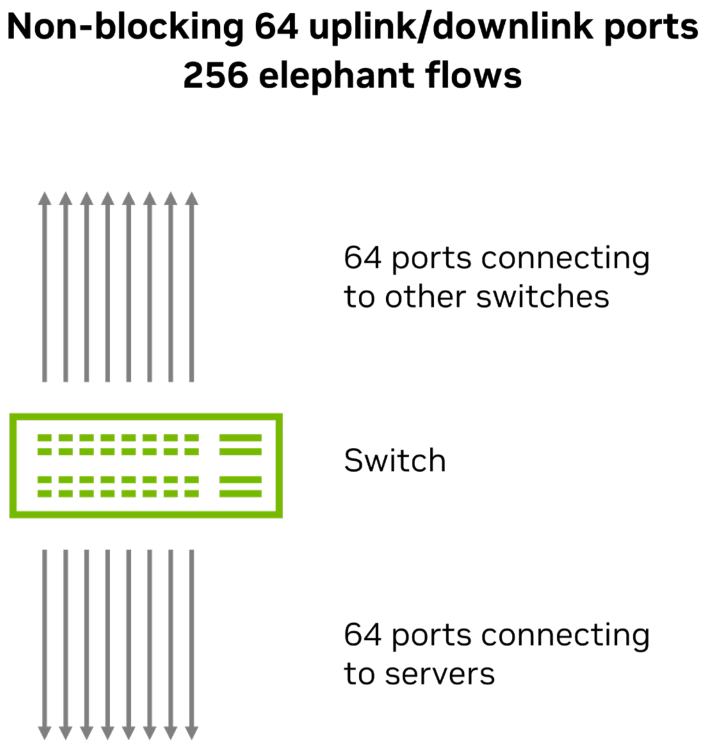
Traditional cloud services generate thousands of flows that are randomly connected to clients around the globe, which gives the cloud service network high entropy. However, AI and storage workloads tend to generate very few, but very large flows. These large AI flows dominate the bandwidth usage per 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 8. Why even non-blocking networks can experience long tail latency

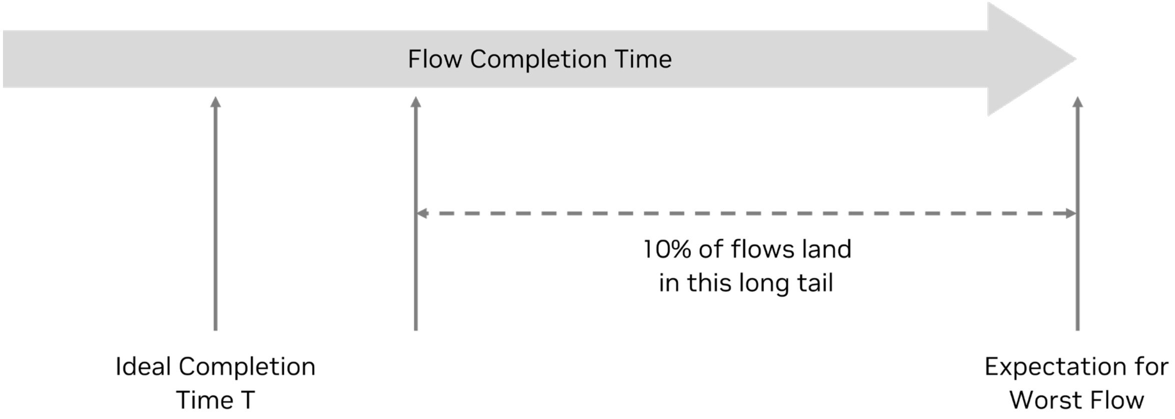

Non-blocking 64 uplink/downlink ports
256 elephant flows
64 ports connecting to other switches
Switch
64 ports connecting to servers

This example consists of a single Leaf switch with 128 ports of 400G.

- 64 ports are 400G downstream ports connecting to servers
- 64 ports are 400G 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 9. 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

With Spectrum-X, RoCE Adaptive Routing is available on Spectrum-4 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.

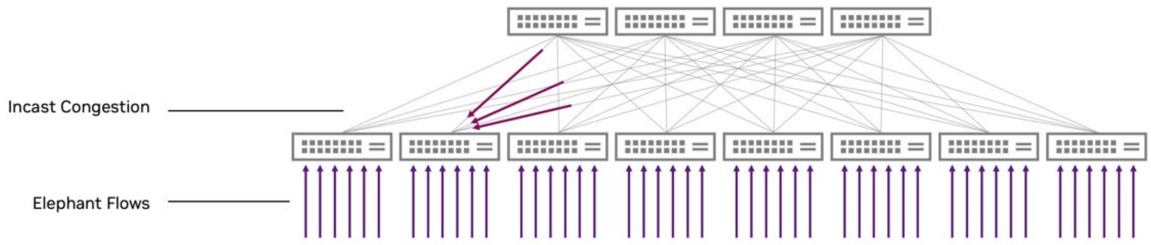
For every packet forwarded to an ECMP group, the switch selects the port with the minimal load over its egress queue. The queues that are evaluated are those that match the packet’s traffic class. As different packets of the same flow travel through different paths of the network, they may arrive out of order to their destination. At the RoCE transport layer, the BlueField-3 DPU 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, BlueField-3 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 Cloud systems, they could 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 10. Incast requires NVIDIA RoCE Congestion Control



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

Key Features of the NVIDIA Spectrum-X 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 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

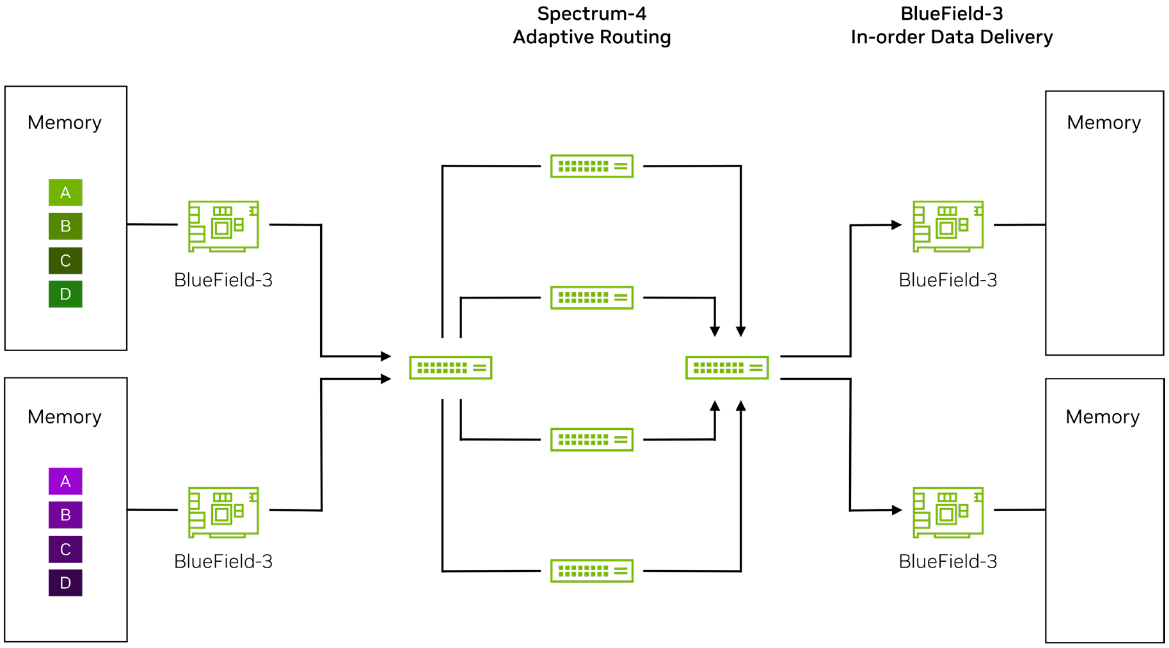
RoCE Adaptive Routing is an end-to-end capability enabled by Spectrum-4 switches and BlueField-3 DPUs. Spectrum-4 switches are responsible for selecting the least-congested port for data transmission on a per-packet basis. As different packets of the same flow travel through different paths of the network, they may arrive out of order to their destination. Bluefield-3 transforms any out-of-order data at the RoCE transport layer, transparently delivering in-order data to the application.

Spectrum-4 evaluates congestion, based on egress queue loads ensuring all ports are well-balanced. For every network packet, the switch selects the port with the minimal load over its egress queue. Spectrum-4 also receives 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 enables up to 95 percent effective bandwidth across the hyperscale system at load, and at scale.

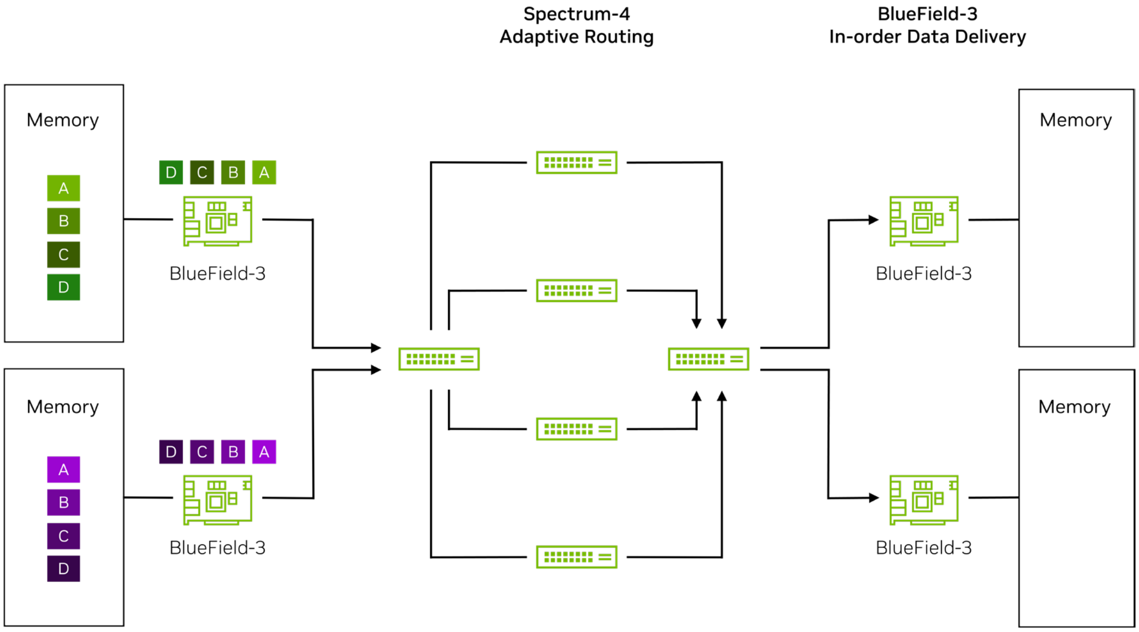
NVIDIA RoCE Adaptive Routing with NVIDIA Direct Data Placement

The following packet-level walkthrough shows how AI flows move between GPUs across a Spectrum-X network. It describes how RoCE Adaptive Routing on the Spectrum-4 switch works in tandem with the NVIDIA Direct Data Placement (DDP) Technology on the BlueField DPU:

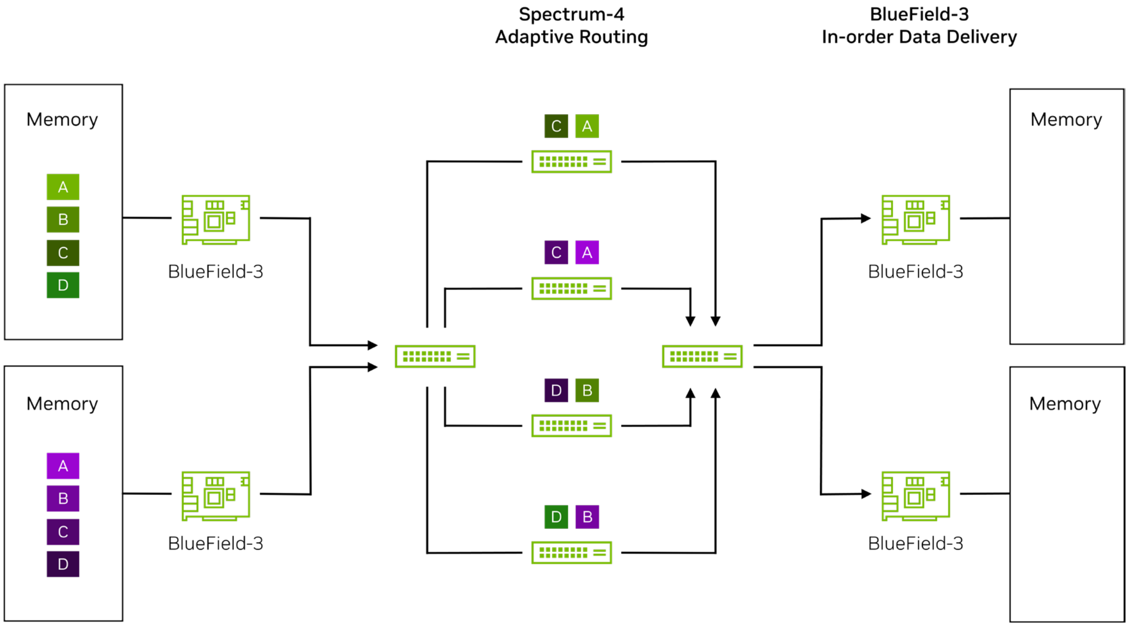
Step 1: Data is sent from server or GPU memory on the left side of the diagram, destined for servers on the right side of the diagram.



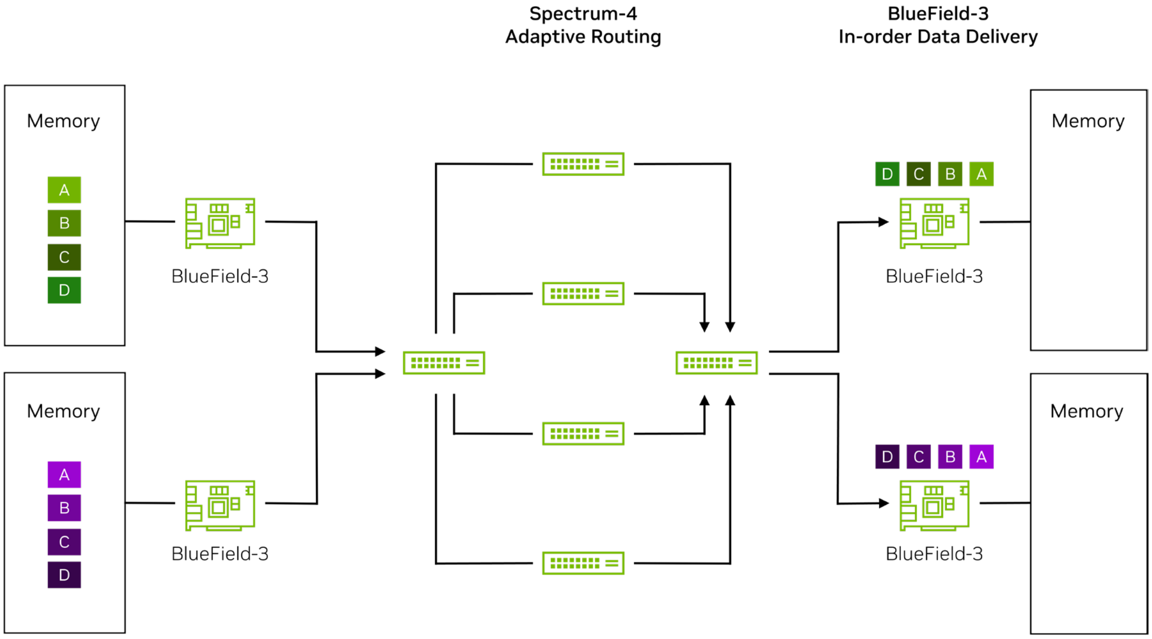
Step 2: The BlueField-3 DPUs place the data into packets and send those packets to the first Spectrum-4 leaf switch, while marking these packets as safe for RoCE Adaptive Routing.



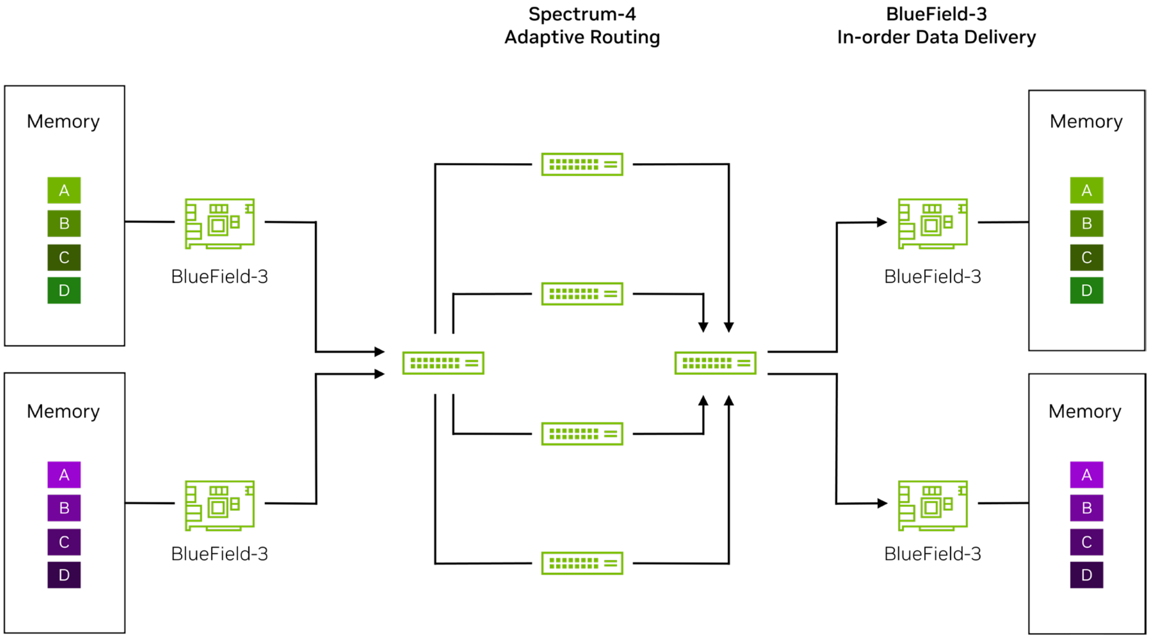
Step 3: The Spectrum-4 leaf switch on the left uses RoCE Adaptive Routing to load balance the packets from the green and purple flows across the 4 spine switches, sending packets from each flow to multiple spines. This improves effective bandwidth from 60 percent on standard Ethernet to 95 percent with Spectrum-X (1.6X).



Step 4: Some of the packets may arrive out of order when they reach the BlueField-3 DPUs on the right side.



Step 5: The BlueField-3 DPUs on the right-side use NVIDIA Direct Data Placement (DDP) Technology to place the data in the correct order in the host/GPU memory.

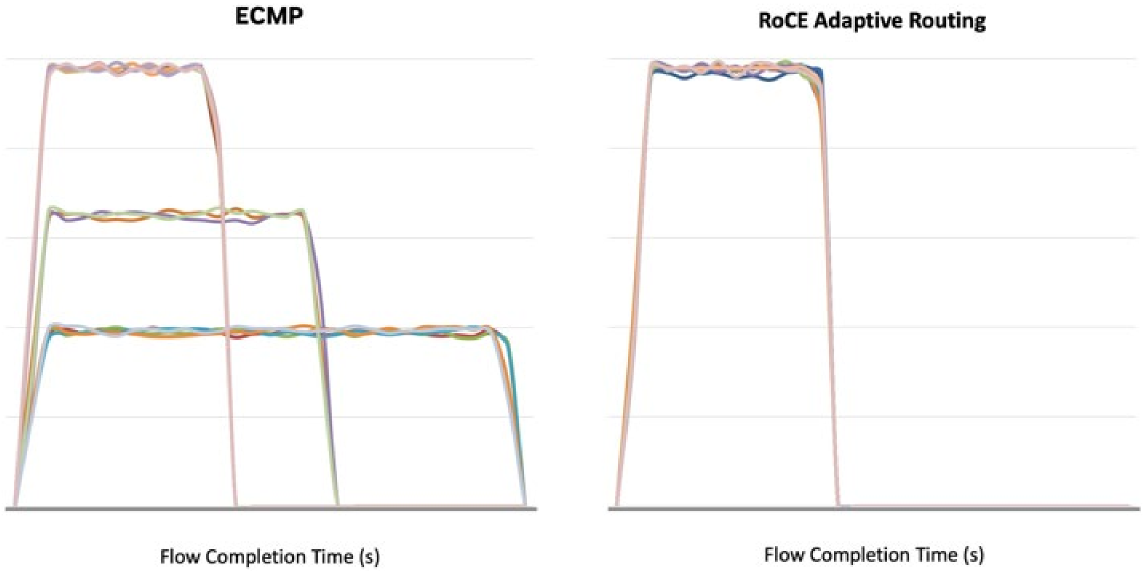


RoCE Adaptive Routing Results

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, causing increased flow completion time, reduced bandwidth, and reduced fairness between the flows. All problems were solved after moving to adaptive routing.

Figure 11. Flow completion time is reduced with RoCE Adaptive Routing



n 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 for AI Workloads

To continue with evaluating adaptive routing in RoCE workloads, we tested the performance gains for common AI benchmarks on a 32-server testbed, built with four NVIDIA Spectrum switches in a two-level, leaf-spine network topology. The benchmark evaluates common collective operations and network traffic patterns found in distributed AI training workloads, such as all-to-all traffic and all-reduce collective operations.

Figure 12. RoCE Adaptive Routing for AI: All-Reduce

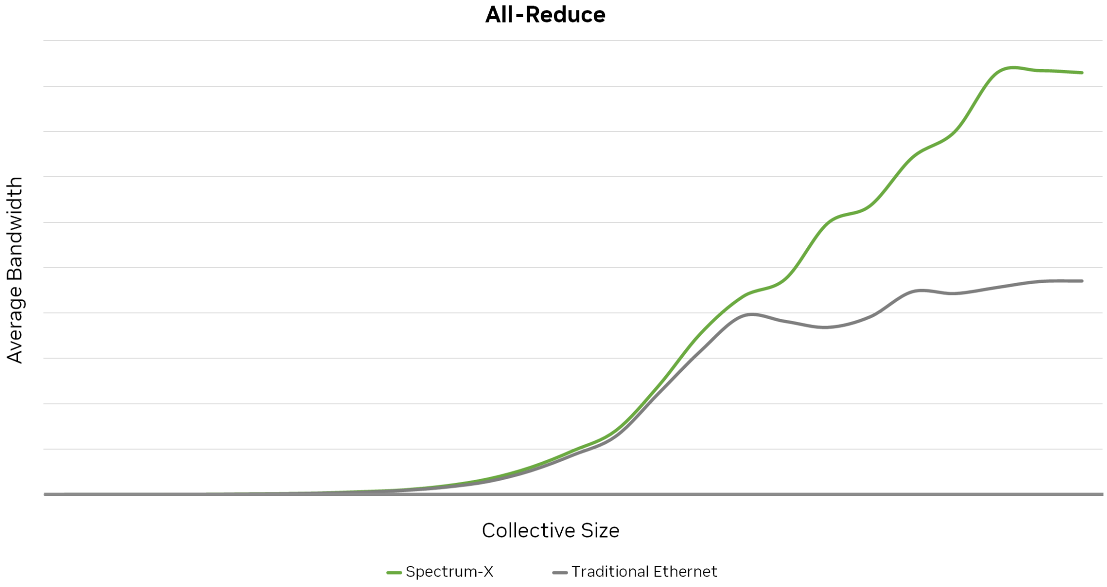
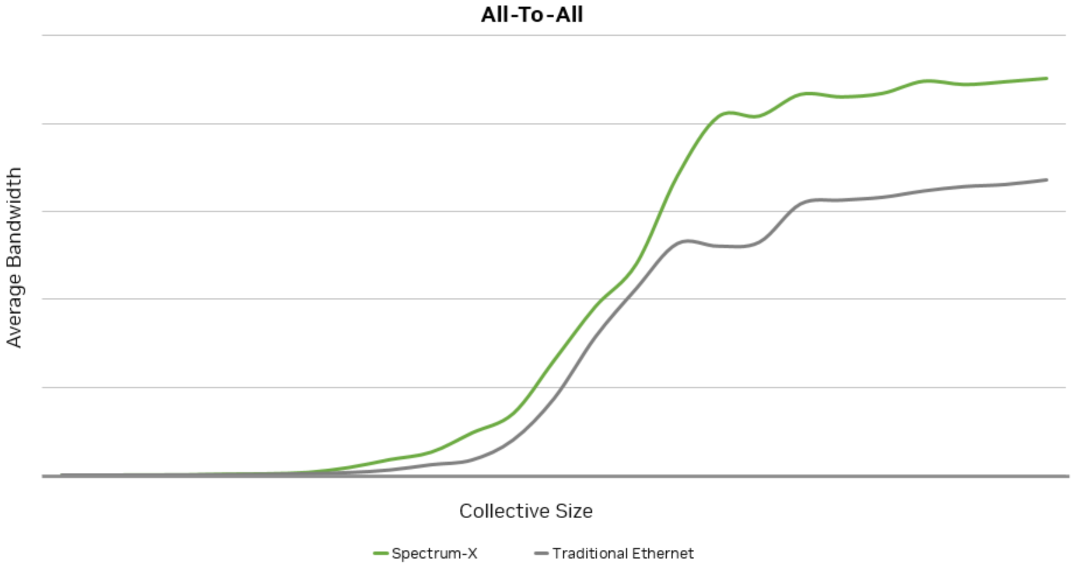


Figure 13. RoCE Adaptive Routing for AI: All-to-All



(Spectrum-X delivers 1.2X Average Bandwidth of Traditional Ethernet)

RoCE Adaptive Routing Summary

In many cases, ECMP flow-based hashing leads to high congestion and variable flow completion time. This reduces applications performance.

Spectrum-X RoCE Adaptive Routing solves this issue. This technology increases the network’s good throughput (“goodput”), minimizes the variability of the flow completion time and, as a result, boosts the application performance.

Combining RoCE Adaptive Routing with NVIDIA Direct Data Placement (DDP) on BlueField-3 DPUs, the technology is transparent to the applications. This ensures that the NVIDIA Spectrum Ethernet platform delivers the accelerated Ethernet needed for maximum data center performance.

Performance Isolation Using NVIDIA RoCE Congestion Control

Applications running concurrently on AI Cloud systems may suffer from degraded performance and reproducible run-times due to network level congestion. This can be caused by the application’s network traffic itself, or background network traffic from other applications. The primary reason for this congestion is known as many-to-one congestion, where there are multiple data senders and a single data receiver.

Such congestion cannot be solved using RoCE Adaptive Routing and actually requires data-flow metering per endpoint. Spectrum-X RoCE Congestion Control is an end-to-end technology where Spectrum-4 switches provide network telemetry information representing real time congestion data. This telemetry information is processed by the Bluefield-3 DPUs which manage and control the data sender’s data injection rate, resulting in maximum efficiency of network sharing. Without congestion control, many-to-one scenarios will cause network back-pressure and congestion spreading or even packet-drop which dramatically degrade network and application performance.

In the congestion control process, BlueField-3 DPUs executes the congestion control algorithm handling millions of congestion control events per second in microsecond reaction latency and applies fine-grained rate decisions. Spectrum-4 switches in-band telemetry holds both queuing information for accurate congestion estimation, as well as port utilization indication for fast recovery. NVIDIA's congestion control significantly improves congestion discovery and reaction time by allowing the telemetry data to bypass the congested flow's queueing delay while still providing accurate and concurrent telemetry.

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

Figure 14. Network congestion creating a victim flow

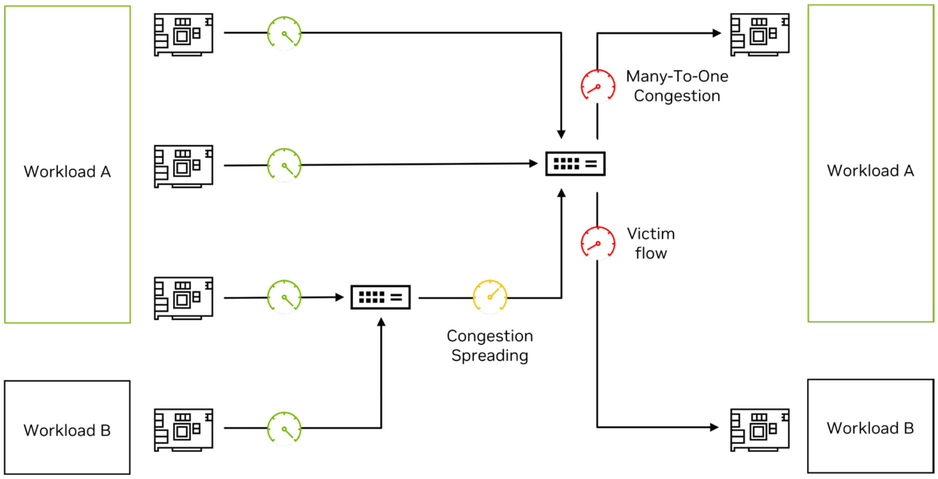


Figure 14 illustrates network congestion creating a victim flow. Four Source DPUs transmit data to two Destination DPUs. Sources 1, 2, and 3 are all transmitting to Destination 1, each receiving ⅓ of the available link bandwidth. Source 4 transmits to Destination 2 through the same leaf switch as Source 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 15. Spectrum-X solves congestion with flow metering and congestion telemetry

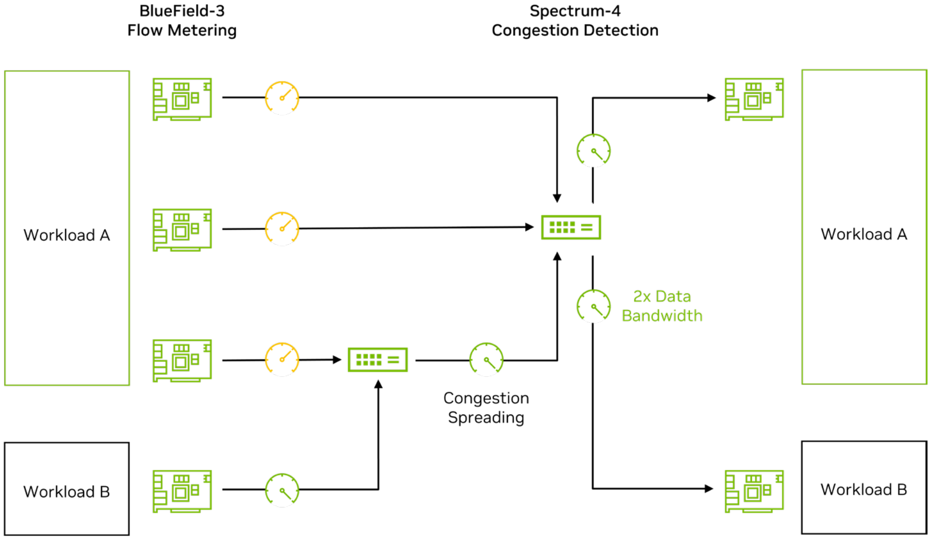


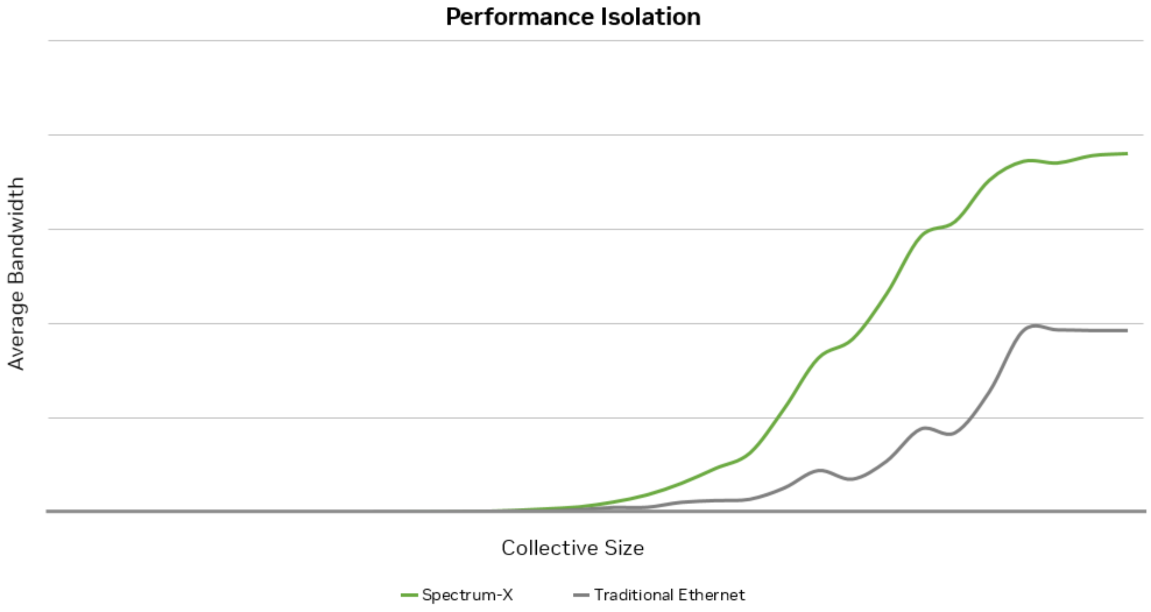
Figure 15 shows how Spectrum-X solves the congestion problem in Figure 13. This diagram has the same setup: Four Source DPUs are sending data to two Destination DPUs. In this case, congestion at the leaf level is avoided by flow metering at Source 1, 2, and 3. This eliminates the back pressure that impacts Source 4. Source 4 receives its full ⅔ effective bandwidth as expected. In addition, Spectrum-4 uses the in-band telemetry generated via What Just Happened to dynamically re-allocate flow paths and queue behavior.

RoCE Performance Isolation

AI Cloud infrastructure needs to support a large number of users (tenants) and parallel applications or workflows. These users and applications compete on the infrastructure's shared resources, such as the network, and therefore may impact performance.

In addition, optimizing the AI network for NVIDIA Collective Communication Library (NCCL) performance requires coordination and synchronization by all workloads running simultaneously on the cloud. The traditional cloud benefits of elasticity and high availability are more limited for AI applications, and performance degradation is a greater and more global concern.

Figure 16. Spectrum-X mechanisms involved in delivering performance isolation



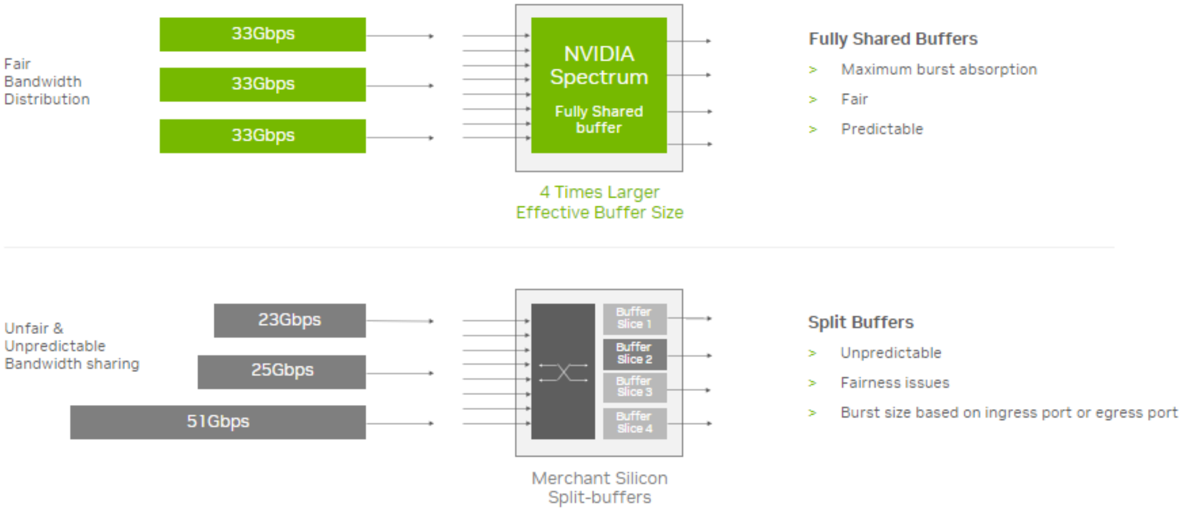
Spectrum-X delivers 2X Average Bandwidth of Traditional Ethernet

The Spectrum-X platform includes mechanisms that, when combined, deliver performance isolation. It ensures that one workload cannot impact the performance of another. These Quality-of-Service mechanisms ensure that any workload cannot create network congestion that will impact data movement of another workload.

With RoCE Adaptive Routing, performance isolation is achieved via fine-grained data path balancing which avoids collision of flows traversing the leaf and spine. With RoCE Congestion Control, performance isolation is achieved via metering and telemetry to prevent victim flows forming from many-to-one host traffic.

Moreover, Spectrum-4 switches deliver performance isolation through a universal shared buffer design. Shared buffers deliver bandwidth fairness for flows of different size, workload protection from “noisy neighbor” flows, and larger microburst absorption for cases where multiple flows have the same destination port.

Figure 17. Spectrum-4 switches feature a universal shared buffer design



NVIDIA Full-Stack Optimization

An AI Cloud is a precision-engineered machine. Its performance can be negatively affected by network events that might be overlooked on a cloud control network, including minor glitches such as a link or device failure. A single sluggish system can cause a ripple effect, slowing down the entire cloud. Moreover, one jabbering DPU 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 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 network.

2. **Component Testing:** NVIDIA first tests each component (switch, DPU, GPU, and AI libraries) individually to ensure they 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
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, BERT, and RetinaNet.

The Ethernet Designed for AI

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

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