DIPL.ING.(FH)KLAUS ROCK

HTTP-QuSS

HTTP - QUANTUM SPEED AND SECURITY

November 25, 2023

AI NODE SERVER NETWORKING

ROCK TECHNOLOGIES
Bonhoefferstr. 37 | 73432 Aalen | Germany | +49-7367-9222-958

Networking for the Era of AI:
The Network Defines the Data Center White Paper

Table of Contents

Introduction 4

AI is a Distributed Computing Problem 5

- Building Networks for AI 5
- Lossless networking and RDMA 5

NVIDIA Spectrum-X: Designed for the Era of Generative AI 6

- Lossless Networking and RDMA 6
- Adaptive Routing, Multipathing, and Packet Spraying 7
- Congestion Control 8
- Performance Isolation and Security 9

NVIDIA Quantum InfiniBand: Inherently Optimized for AI 11

- Collective Computational Power 11
- In-Network Computing 12
- NVIDIA Quantum InfiniBand Adaptive Routing 13
- NVIDIA Quantum InfiniBand Congestion Control 14

Avoiding Common Misconceptions 15

- Continued Development of Emerging AI 15
- Cut-through Switching and End-to-end Link Speed 15
- Switch Radix and AI Scalability 16
- Switch Buffer Architectures 17
- Resilience to Network Link Failures 18
- AI Cloud Management 18

Conclusion 19

List of Figures

Figure 1 Diagram of an RDMA implementation for GPU-to-GPU Communication..... 7
Figure 2 Diagram of packet-granular NVIDIA Spectrum-X Ethernet Adaptive Routing implementation..... 8
Figure 3 Example of NVIDIA Spectrum-X Ethernet Congestion Control with switch and NVIDIA BlueField DPU working in tandem..... 9
Figure 4 Diagram highlighting importance of universal shared packet buffer architecture versus split-buffer implementation..... 10
Figure 5 Diagram showing single GPU and using NCCL to scale across multi-GPU and multi-node configurations..... 12
Figure 6 Visual representation of Scalable Hierarchical Aggregation and Reduction Protocol Architecture (SHARP) on the left, and its performance when used with NCCL on the right..... 13
Figure 7 Diagram of NVIDIA Quantum InfiniBand Congestion Control Mechanism 14
Figure 8 Diagram showcasing an optimized 400Gb/s end-to-end AI Cloud Ethernet topology..... 16
Figure 9 Graphs showing real and projected latency compared to buffer size and buffer occupancy..... 17

Introduction

For decades, traditional cloud data centers were focused on providing various resources to a broad user-base. Advancements in the virtualization of infrastructure components enabled the quick spin-up of systems and applications as needed to meet demands.

These data centers were well suited to support a diverse set of users and business applications and were sufficient enough to support smaller-scale workloads connected via commodity-class Ethernet. While Ethernet incorporated an expansive and comprehensive feature-set, it wasn't performant for scaling beyond a few nodes. It also wasn't suitable for high-performance computing.

Today, we are facing new classes of data centers: AI Clouds and AI Factories, that require both accelerated computing and high-performance networking to support artificial intelligence (AI). As a result, the landscape of today's hyperscale and cloud deployments is changing considerably. With the adoption of GPU-accelerated computing architectures, AI researchers and practitioners can harness the power of distributed accelerated computing that would otherwise be infeasible.

The data center's network is ultimately responsible for ushering in the era of AI advancements and performance, because it acts as the backbone of the data center for distributed AI model training or when harnessing the power of Generative AI.

AI is a Distributed Computing Problem

Traditional data center computing refers to the model where all computational resources, including servers, storage, and networking, are centralized within a physical location or facility. Distributed computing, on the other hand, involves the utilization of multiple interconnected servers or nodes working together to perform a task or execute a process. In this model, the workload is distributed across various machines, connected by a high-speed, low latency network.

Deploying generative AI applications, or training a foundational AI model such as ChatGPT, BERT, or DALL-E, can be computationally intensive, especially for large and complex models. As the volume of data and model size increases, distributed computing is employed to tackle this challenge. It accelerates the training process by distributing the workload across multiple interconnected compute nodes. Particularly, the runtime of a single distributed task is governed by that of the slowest participating node. The network plays a significant role in ensuring the timely arrival of messages to all participating nodes. This makes tail latency—the time of arrival of the last participating message— critical, especially in large-scale data center deployments and in the presence of competing workloads. Additionally, the network's ability to scale and handle an increasing number of nodes is essential for training large AI models and handling vast amounts of data.

Building Networks for AI

When evaluating the network architecture for a data center's adoption of AI, it should be examined as a comprehensive, integrated end-to-end solution with distributed computing at the top of mind. It should also encompass all of the attributes, processes, communication libraries, and techniques needed to achieve data center-scale performance, cost, and value. This paper will cover these and other key areas:

Lossless networking and RDMA

- Adaptive routing and out-of-order packet data
- Congestion control
- Security and performance isolation
- In-Network Computing
- Avoiding common misconceptions

NVIDIA Spectrum-X: Designed for the Era of Generative AI

An AI Cloud is a new data center class that supports Generative AI workloads. The AI Cloud bundles all of the traditional cloud's core capabilities, such as multi-tenancy, security, and support for a variety of workloads, and adds support for larger scale Generative AI applications. Generative AI refers to a category of AI algorithms that generate new outputs based on the data they've been trained on. Unlike traditional AI systems that are designed to recognize patterns and make predictions, Generative AI creates new content in the form of images, text, audio, and more. NVIDIA® Spectrum™-X is a revolutionary solution for building multi-tenant, hyperscale AI clouds with Ethernet for the era of Generative AI.

Lossless Networking and RDMA

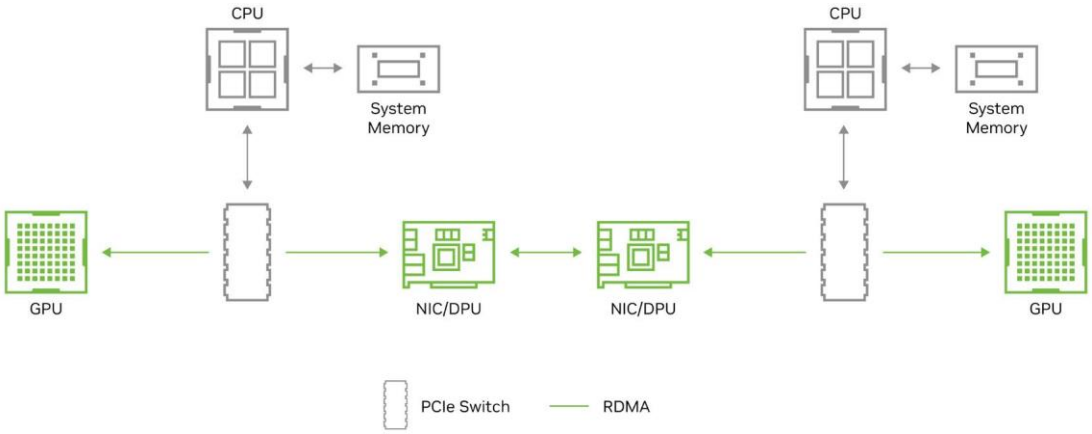
With lossy networking, data is transmitted with the understanding that loss or degradation of information may occur. The network prioritizes data transmission over perfect accuracy. The consequences of implementing a lossy network for AI introduce significant penalties in terms of performance, GPU idle time, power consumption, etc.

With lossless networking, data is transmitted without any loss or corruption. The network guarantees all the data packets reach the destination accurately, and no information is lost during the transmission. While Ethernet was designed as inherently lossy, lossless networking is fundamental with InfiniBand networking, and has been the de facto standard for large-scale deployments. Today, with the adoption of GPU computing and large-scale AI use cases within cloud environments, Ethernet can be a practical solution when it is running RDMA over Converged Ethernet (RoCE) and Priority Flow Control (PFC) coupled with a lossless-network implementation such as Spectrum-X.

Remote Direct Memory Access (RDMA) enables high-speed, low-latency data transfers over a network. It allows data to be directly transferred between the memory of remote systems, GPUs, and storage without involving the CPUs of those systems. With traditional networking, data transfer involves multiple steps: The data is first copied from the source system's memory to the network stack, and then sent over the network. Finally, it's copied into the destination system's

memory after multiple steps on the receive side. RDMA bypasses these intermediate steps, resulting in more efficient data transfers.

Figure 1 Diagram of an RDMA implementation for GPU-to-GPU Communication



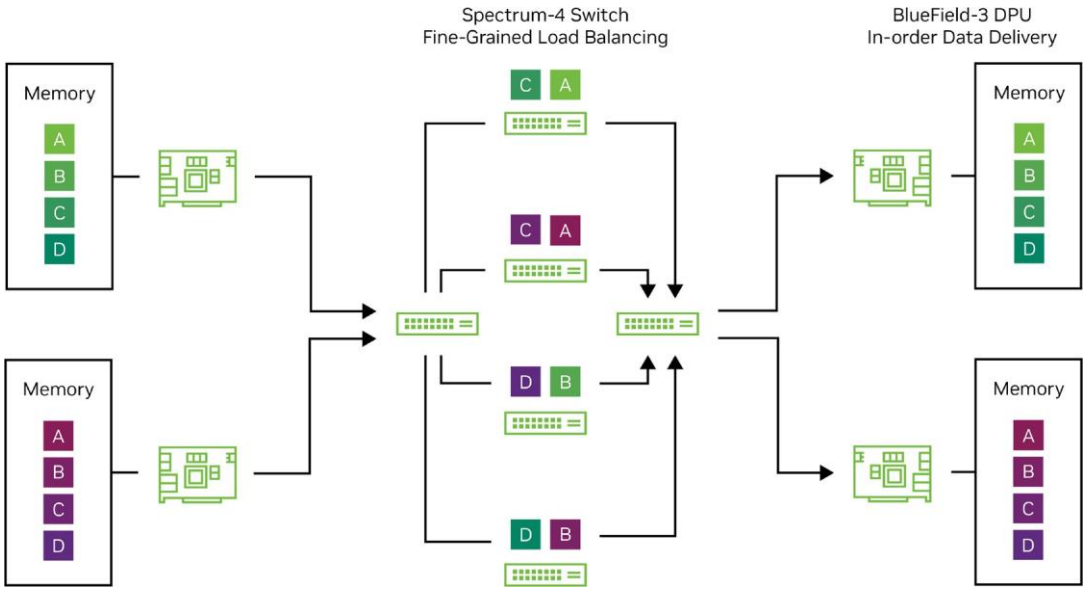
Adaptive Routing, Multipathing, and Packet Spraying

Traditional data center applications tend to generate many small data flows that enable statistical averaging to reflect network traffic. This means that for flow routing, simple, static hash-based algorithms such as Equal Cost Multi-Path (ECMP) implemented by the switch are sufficient for avoiding network traffic issues. In contrast, AI workloads generate a small number of large data flows, known as “elephant flows.” These large data flows use up a significant amount of link bandwidth, and if multiple elephant flows are routed to the same link, congestion and high latency will occur. Using ECMP with AI, the odds are very high that, even with a non-blocking topology, such collisions will happen. Because jobs are dependent on worst-case performance, these collisions will result in model training times that are both higher than desired and very unpredictable.

As such, adaptive routing algorithms are needed to dynamically load balance the data traversing the network. Additionally, the routing needs to be very granular to avoid collisions. If the routing is done flow-by-flow, there's still a strong statistical possibility that congestion will occur. However, when packet spraying (routing packet-by-packet), it's likely that packets will arrive out of order at their destination. For packet-granular adaptive routing, a flexible reordering mechanism must be put in place so that the adaptive routing is invisible to the application. Spectrum-X achieves this by combining the load balancing capabilities of the

Spectrum-4 switch with the Direct Data Placement (DDP) performed by the BlueField-3 DPU to deliver end-to-end adaptive routing.

Figure 2 Diagram of packet-granular NVIDIA Spectrum-X Ethernet Adaptive Routing implementation.



Congestion Control

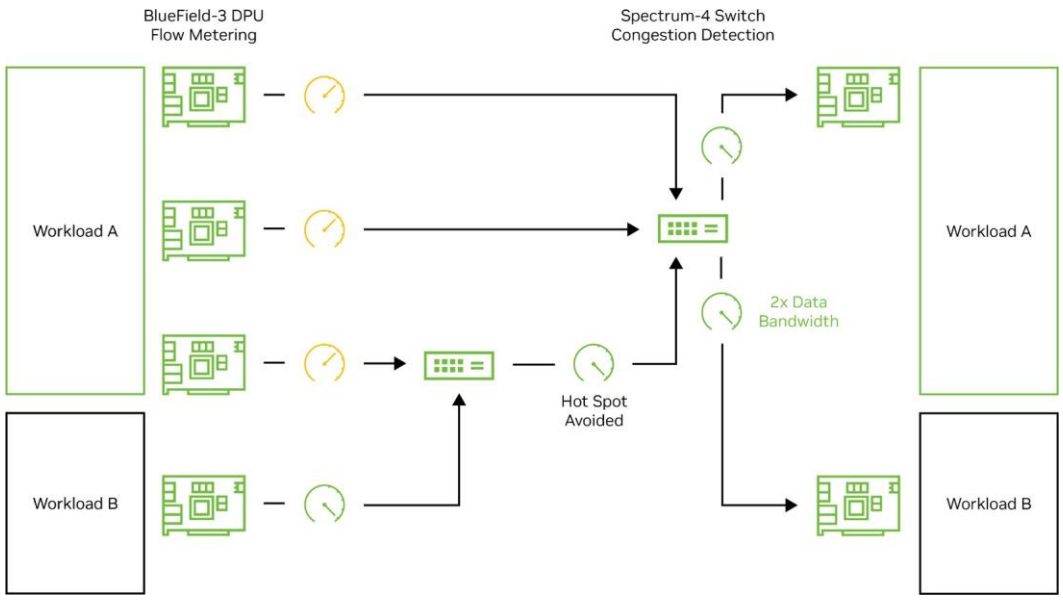
Within multi-tenant AI Cloud environments where different AI jobs run simultaneously, network congestion can arise. This is particularly evident when n senders try to transmit data to one destination (or even n senders with n different destinations transmit data to a switch that has background network traffic from other applications). Such network congestion not only leads to higher latency and decreased effective bandwidth, but the propagation of network “hot spots” and associated back pressures also result in victimization; that is, adjacent tenants may be affected by congestion from another tenant.

The most typical congestion control method, Explicit Congestion Notification (ECN), isn't sufficient when deploying Generative AI over Ethernet. Ultimately, to relieve congestion, the network device transmitting the data (NIC or DPU) must be metered. With ECN, this metering doesn't occur until the switch buffer reaches a particular capacity threshold. The receiver then notifies the sender to meter its throughput until the receiver sees the congestion is cleared up. However, in a bursty traffic situation common to large scale AI

models, this congestion communication's latency could be too high, resulting in over-full buffers and dropped packets. While deep buffer switches can reduce the likelihood of buffers reaching capacity, the added latencies they introduce defeat the intended purpose of congestion control.

There are several ways to implement congestion control, but successful designs ultimately require the switch and NIC/DPU to work in tandem. Spectrum-X leverages in-band, hardware accelerated telemetry data from the Spectrum-4 switch to inform the BlueField-3 DPU to engage in flow metering. Other implementations require deep buffer switches unsuitable for AI and rely on complex, proprietary protocols that deviate from standard Ethernet.

Figure 3 Example of NVIDIA Spectrum-X Ethernet Congestion Control with switch and NVIDIA BlueField DPU working in tandem.



Performance Isolation and Security

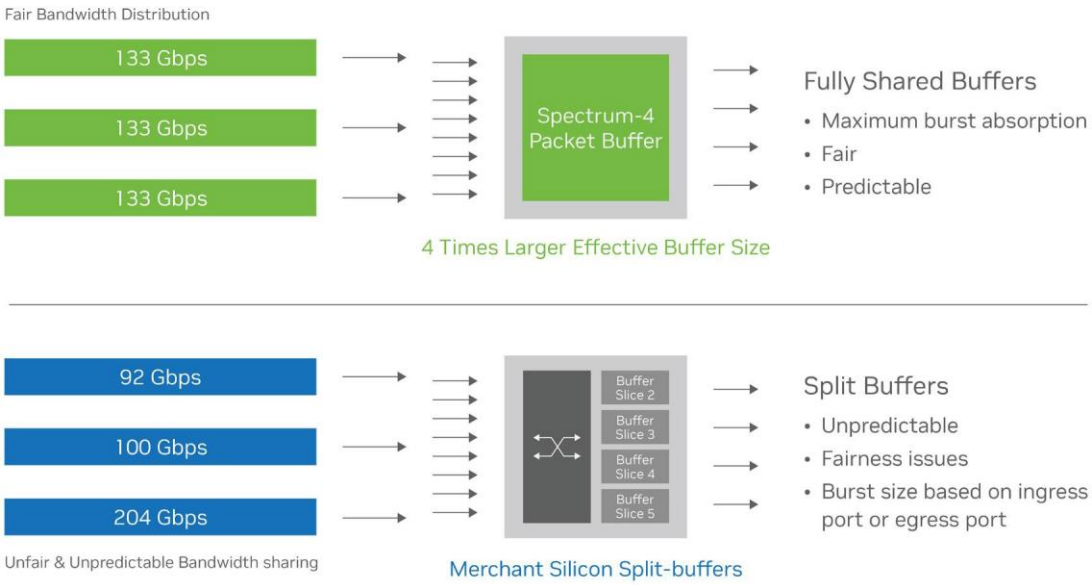
Multi-tenant environments such as AI Clouds require protection from other jobs running on the same infrastructure. Many Ethernet ASIC designs aren't built with ASIC-level job protection. This can lead to certain jobs being "victimized" with very low effective

bandwidth when a “noisy neighbor” (another adjacent job) sends network traffic to the same destination port.

Ethernet networks must also consider network fairness. AI Clouds must support a heterogeneous mix of applications on the same infrastructure. Different applications may use different data frame sizes, and without isolation optimizations, bigger data frames will use up a disproportionate share of the bandwidth when transmitting to the same destination port as a smaller data frame.

Shared packet buffers are key to delivering performance isolation and preventing noisy neighbors and network unfairness. A universal shared buffer that provides the same cache access to every port on the switch provides the predictability and consistent, low latency required for mixed AI Cloud workloads.

Figure 4 Diagram highlighting importance of universal shared packet buffer architecture versus split-buffer implementation



In addition to considering performance isolation from an effective bandwidth perspective, it’s important to recognize that performance isolation and zero-trust architecture are key to network security for multi-tenant environments. Data must be protected both at rest and in motion, with efficient encryption and authentication tools delivering security without compromising performance. The BlueField-3 DPU (compatible with both Ethernet and InfiniBand) features secure boot for hardware-based root-of-trust, and also supports MACsec and IPsec for data-in-motion encryption as well as AES-XTS 256/512 encryption for data-at-rest.

NVIDIA Quantum InfiniBand: Inherently Optimized for AI

Without question, NVIDIA Quantum InfiniBand has enabled many of the large-scale supercomputing deployments for complex distributed scientific computing. However, as a lossless network with ultra-low latencies, native RDMA architecture, and in-network computing capabilities, it's revered as the gold standard in performance, and pivotal in accelerating today's mainstream development and deployment of AI.

Geared more towards massive, large-scale workflows or large AI foundational model development, AI Factories have emerged. As foundation models serve as the starting point for developing more advanced and complex models, these models are trained on mountains of raw data that are fed across the data center. It is paramount that the AI Factory's networking provides scalable performance across many hundreds and thousands of GPUs working together on a single application, such as AI model training. NVIDIA Quantum InfiniBand gives developers and scientific researchers the fastest networking performance and feature-sets available. This includes in-network computing for hardware-based acceleration of collective communication operations that are used extensively with AI systems.

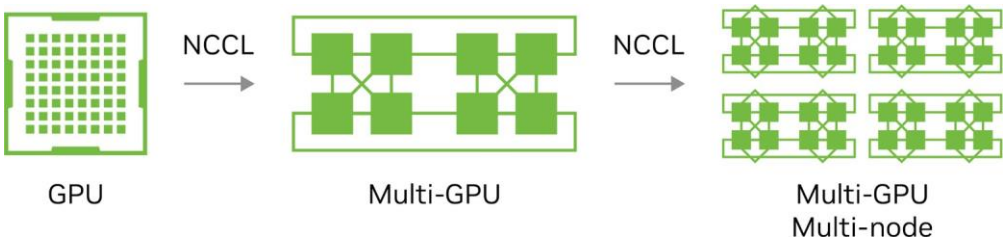
Collective Computational Power

Collective communication algorithms are instrumental for ensuring efficient and coordinated communication between distributed nodes during AI model training. They allow large-scale models to be trained effectively, improve training speed, reduce communication overhead, and enable distributed training to take advantage of the collective computational power of multiple nodes. This leads to accelerated model convergence and enhanced performance.

Collective communication libraries have been developed for deep learning frameworks to take advantage of GPUs within and across multiple nodes. NVIDIA Collective Communication Library (NCCL) is an example of such a library that implements communication algorithms for all-reduce, all-gather, reduce, broadcast, reduce-scatter, as well as any send/receive based communication pattern. It's optimized to achieve high bandwidth on any platform using PCIe and or NVLink, and scales across multiple machines, using NVSwitch, InfiniBand, or Ethernet.

.

Figure 5 Diagram showing single GPU and using NCCL to scale across multi-GPU and multi-node configurations.

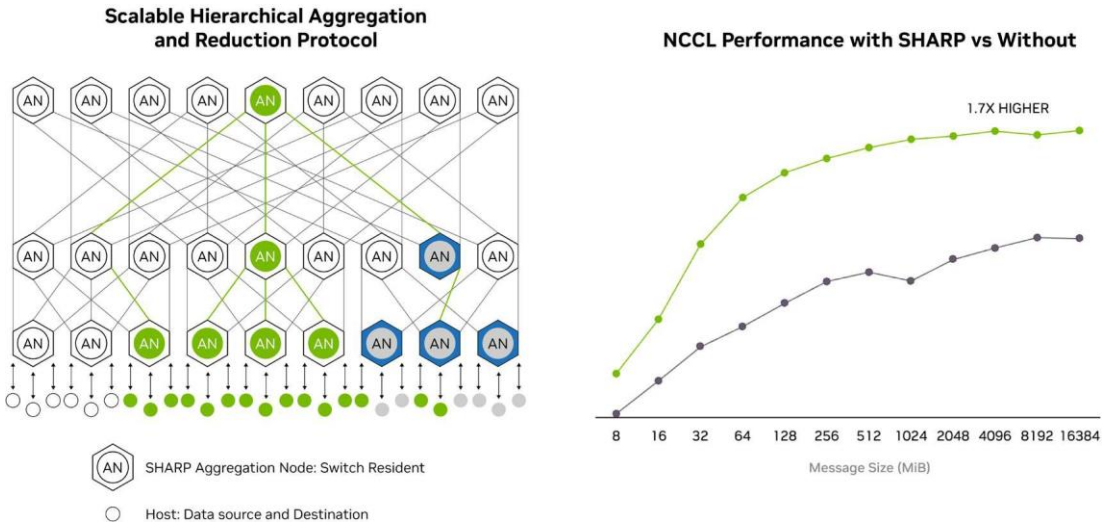


In-Network Computing

In-network computing is a unique performance capability developed specifically for the InfiniBand architecture. This feature enables hardware-based computing engines within the network for offloading complex operations at scale. In-network computing is implemented on the NVIDIA Quantum InfiniBand switch as NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™.

As an in-network tree-based aggregation mechanism, SHARP supports multiple simultaneous collective operations. With SHARP enabled, switches are identified as aggregation nodes and will perform such data reductions. NCCL takes advantage of this capability when performing communication algorithms across many multi-GPU nodes. As data is sent only once to perform the operation, it effectively doubles the bandwidth for data reductions. Thus, NCCL performance running on an end-to-end NVIDIA Quantum-2 400Gb/s InfiniBand network using SHARP, would perform better than an 800Gb/s end-to-end network without it.

Figure 6 Visual representation of Scalable Hierarchical Aggregation and Reduction Protocol Architecture (SHARP) on the left, and its performance when used with NCCL on the right



NVIDIA Quantum InfiniBand Adaptive Routing

NVIDIA Quantum InfiniBand operates as a full Software-Defined Network (SDN) and is managed by a software management utility called the Subnet manager (SM). This centralized entity configures the switches to choose routes based on the network conditions. The switch ASIC selects the least loaded output port (from a set of outgoing ports) that will achieve the best performance across the network. The selection between different outgoing switch ports is based on a grading mechanism that considers egress port queue depth and path priority, where the shortest path has higher priority.

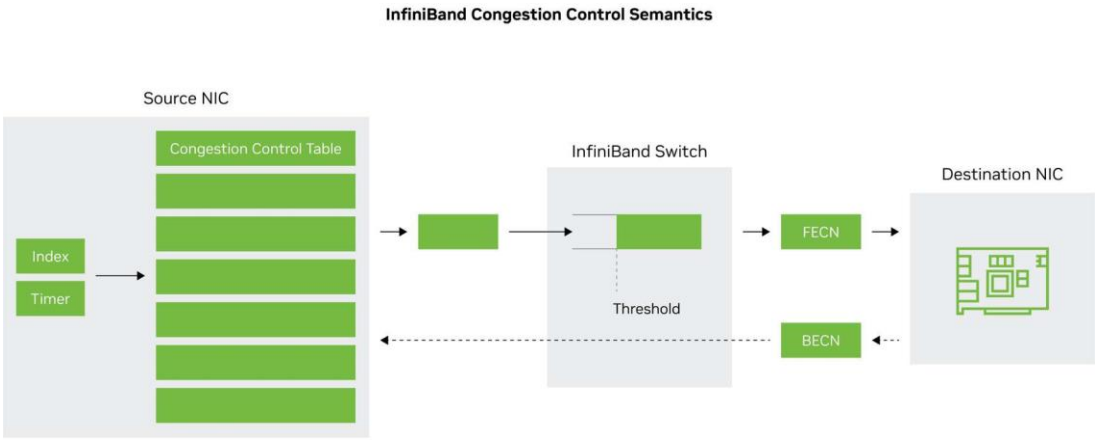
InfiniBand's adaptive routing maximizes overall performance by spreading the traffic across all network links and increasing link utilization and balance, thus optimizing link bandwidth. It's important to know that adaptive routing can lead to network packets arriving at their destination out-of-order. However, InfiniBand, as an end-to-end solution, natively includes in-hardware capabilities to manage out-of-order packet arrivals.

NVIDIA Quantum InfiniBand Congestion Control

NVIDIA Quantum InfiniBand includes a comprehensive and scalable method for Quality of Service (QoS) that guarantees deterministic bandwidth and latency using a credit-based flow control mechanism to regulate data flow between sender and receiver.

InfiniBand implements Congestion Control Architecture (CCA), a three-stage process to manage congestion events. When a switch detects congestion, it turns on a bit (in packets) known as Forward Explicit Congestion Notification (FECN). When the packet reaches the destination adapter, it responds to the source with packets having a different bit set called Backward Explicit Congestion Notification (BECN). When the sending or source adapter receives a BECN, it responds by throttling back its injection of packets.

Figure 7 Diagram of NVIDIA Quantum InfiniBand Congestion Control Mechanism



Avoiding Common Misconceptions

Avoiding common misconceptions is crucial when it comes to network design. One prevalent misconception is that changing the end-to-end link speed might be acceptable for AI deployments, but in reality, this can lead to latency and performance penalties. Other misconceptions when building a network for AI include:

- Continued development of emerging AI
- Switch radix is a critical metric
- Shallow versus deep buffer architecture
- Techniques for network resilience

The underlying networking ultimately defines the data center's class of operation and expected level of performance and efficiency; therefore, it’s essential to dispel these misconceptions and embrace a holistic approach that balances performance, security, and flexibility to align with the data center’s mission, whether it’s an AI Cloud or an AI Factory.

Continued Development of Emerging AI

InfiniBand addresses the growing demand for faster and more efficient scalable communication between GPU accelerators, servers, storage systems, and other components. InfiniBand's architecture allows for the introduction of new features and capabilities without requiring a complete overhaul of the technology. This adaptability makes it possible to incorporate new technologies and techniques as they emerge, making it well-suited to meet the challenges and requirements of future technology landscapes.

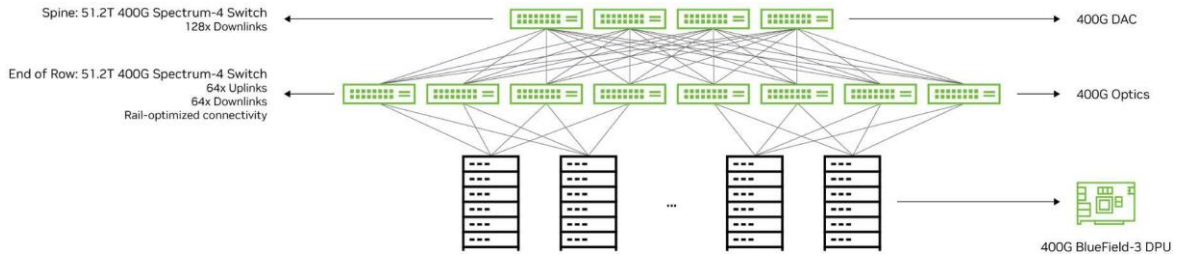
Cut-through Switching and End-to-end Link Speed

Ethernet employs two modes for data handling: store-and-forward switching and cut-through switching.

A store-and-forward switch waits for the entire data frame to be received before sending the data, while a cut-through switch sends the data to the destination immediately. For AI workloads, cut-through switching is preferred.

Cut-through switching requires the same link speed end-to-end. Network designs that change link speed (i.e., from 400Gb/s host-to-leaf to 4 x 100Gb/s from leaf-to-spine) require traffic splitting and store-and-forward switching. This introduces a latency penalty that becomes more significant when dealing with the large data frames usually seen in AI training. Spectrum-X uses end-to-end cut-through connectivity to optimize the network for AI.

Figure 8 Diagram showcasing an optimized 400Gb/s end-to-end AI Cloud Ethernet topology.



Switch Radix and AI Scalability

Switch radix, or the number of logical MACs a switch can support, has been traditionally used as a proxy for the scalability of a switch. A larger switch radix can connect more hosts for a given number of network tiers.

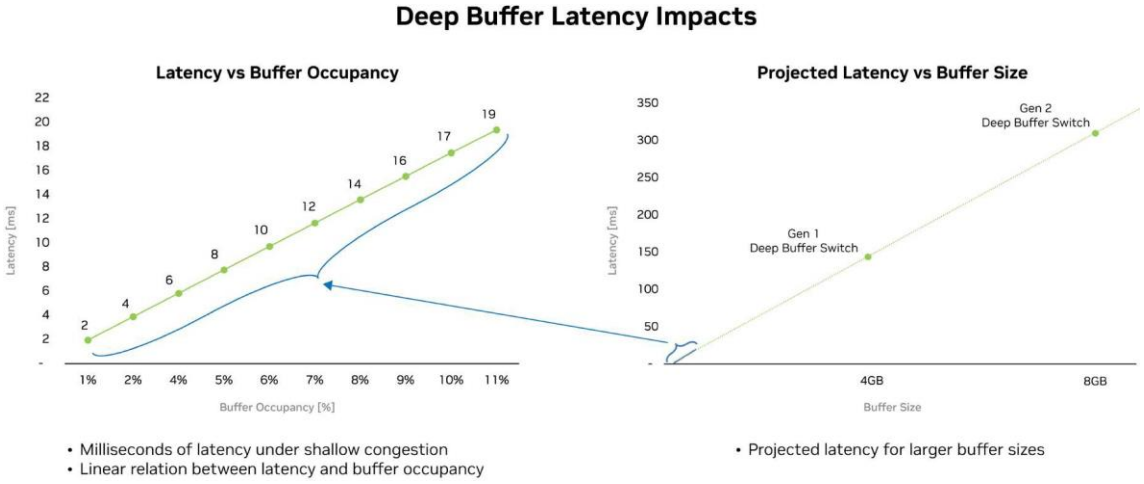
With AI, this paradigm has changed. Effective bandwidth, latency, and tail latency are vital to performance, and are independent of radix scale. A higher radix switch can connect a larger number of GPUs with the same number of network tiers for a lower network price point, but at the cost of decreased application performance and lower ROI. Collective operations such as NCCL All-to-All require the same high wire speed at every tier of the network, and splitting traffic to take advantage of radix scale isn't recommended.

Switch Buffer Architectures

While InfiniBand switches are “shallow” buffer switches by design, Ethernet switches can be broadly categorized as either “deep” or “shallow” buffer switches. Deep buffer switches have buffer sizes in the gigabytes (GB), while shallow buffer switches (such as Spectrum Ethernet switches) have buffer sizes in the megabytes (MB). Deep buffer switches were originally designed for different purposes, such as routing and WAN; as such, they have a very different architecture compared to traditional shallow buffer Ethernet switches. Deep buffer switch systems often utilize a modular design, featuring larger chassis switches populated with line cards.

While deep buffer switches are feature-rich and support the scale needed for Data Center Interconnect (DCI) and telco networking, they aren't optimized for AI networking. Deep buffer switches hold extra data traffic and are less sensitive to microbursts, but larger data capacity leads to higher tail latencies, causing increased average latency and high jitter. This directly impacts AI workloads which depend on worst-case latency, leading to higher job completion times and increasing time-to-train.

Figure 9 Graphs showing real and projected latency compared to buffer size and buffer occupancy



Resilience to Network Link Failures

NVIDIA Quantum InfiniBand switches are distinctively equipped with self-healing capabilities. Thanks to this self-healing autonomy, the speed with which communications can be corrected in the event of a link failure is fast enough to save communications from expensive retransmissions or absolute failure.

Unlike traditional application workloads with homogeneous traffic patterns that typically run on Ethernet, AI generates heterogeneous traffic that's bursty and highly sensitive to network failures. When a link from a leaf to a spine is down, for instance, this impacts multiple GPU nodes in multiple racks and significantly reduces performance for All-to-All. Popular Ethernet-based redundancy measures such as EVPN Multihoming or MLAG won't solve the performance issue.

Spectrum-X provides best in-class resiliency for AI workloads which require dual/multi-rail host designs (i.e additional NIC ports for full hardware redundancy) and intelligence at the switch that adjusts load balancing based on detection of link failure. Spectrum-X provides the most optimal environment to ensure time-sensitive data delivery.

AI Cloud Management

Managing an AI cloud data center that serves thousands of users necessitates a robust set of management tools to ensure efficient operations, monitoring, security, and resource allocation. While not typically a requirement for an AI Factory, the AI Cloud relies upon customized Cloud Management Platforms (CMPs) to manage and automate the cloud infrastructure. Though typically native to Ethernet, this management and provisioning ecosystem can also be developed and integrated with InfiniBand.

Deploying an AI Cloud doesn’t necessarily require reinvesting in the ecosystem for resource provisioning, workload orchestration, and user access control. It’s often practical to reuse the ecosystem that supports virtualization management, orchestrates containers and services across the cloud infrastructure, and monitors resource health and performance, etc. This approach is often considered when choosing the type of networking that will serve as the cornerstone in the AI data center.

Conclusion

AI workloads have introduced new challenges and requirements for data center network architectures. The network defines the data center and serves as the backbone of the AI infrastructure. It is essential to consider the network's capabilities and end-to-end implementation when deploying data centers for both generative AI and foundational models.

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. Neither NVIDIA Corporation nor any of its direct or indirect subsidiaries and affiliates (collectively: "NVIDIA") make any 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, NVIDIA Spectrum, BlueField, and Scalable Hierarchical Aggregation and Reduction Protocol Architecture (SHARP), 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.

Copyright

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