Solution Overview

NVIDIA Quantum-X800 InfiniBand Platform

Optimized for GPU computing and AI infrastructure at the trillion-parameter scale.

Accelerate the Next Generation of AI

The global shift toward ubiquitous AI is expanding rapidly, fueled by surging demand for AI solutions across various sectors. This demand is leading to significant investments aimed at streamlining productivity. Companies are enhancing their offerings with generative AI, while early adopters are seeing improvements in user experiences and business performance.

The race is on for an AI platform that maximizes performance for a given TCO, further accelerating AI adoption across all industries. In addition, the convergence of AI with traditional high-performance computing (HPC) is hyper-accelerating scientific discovery. As the landscape for AI evolves, the quest for ever-larger language models becomes central for researchers and organizations. This pursuit reveals the challenges and complexities of real-time inference as these models expand.

To maximize AI's benefits, data center architects must design networks tailored for AI workloads, focusing on networking considerations to unlock AI's full potential and drive data center innovation.

NVIDIA is pioneering innovations at data center scale, offering the most energy-efficient networking platforms with unparalleled bandwidth, ultra-low latency, and CPU utilization, setting industry benchmarks for performance and efficiency.

NVIDIA Quantum-X800 InfiniBand Platform

The NVIDIA Quantum-X800 platform is the next generation of NVIDIA Quantum InfiniBand. Unleashing 800 gigabits per second (Gb/s) of end-to-end connectivity with ultra-low latency, NVIDIA Quantum-X800 is purpose-built for training and deploying trillion-parameter-scale AI models. The NVIDIA Quantum-X800 Q3400/Q3450 InfiniBand switch at the core of the platform supports 2x faster speeds and 5x higher scalability for AI compute fabrics. Additionally, the platform includes the NVIDIA ConnectX SuperNIC, delivering 800G connectivity to the host, with advanced offload and quality-of-service enhancements.

Leveraging advanced hardware-based NVIDIA In-Network Computing with NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™ v4, adaptive routing, and telemetry-based congestion control for highest performance, NVIDIA Quantum-X800 is enabling a new frontier of AI innovation.

Features and Innovations

- **Innovative In-Networking Computing:** Advanced technologies like NVIDIA SHARP v4, Message Passing Interface (MPI) tag matching v2, PI_Alltoall, and programmable cores boost NVIDIA In-Network Computing.
- **Adaptive Routing:** The switch and NVIDIA® ConnectX® SuperNIC™, working together, maximize bandwidth and ensure network resilience for AI fabrics.
- **Telemetry-Based Congestion Control:** These techniques provide noise isolation for multi-tenant AI workloads.
- **Network Resiliency Improvements:** The platform proactively tackles hardware issues to maintain seamless application performance.
- **Acceleration Engines:** These engines cut latency and double data throughput, enhancing network efficiency.

NVIDIA Quantum-X800 Switch Systems

NVIDIA Quantum X-800 InfiniBand Q3400-RA 4U

The Q3400-RA 4U switch—the first to utilize 200 Gb/s-per-lane serializer/deserializer (SerDes) technology—significantly boosts network performance and bandwidth. It includes 144x 800 Gb/s ports distributed over 72 octal small-form-factor pluggable (OSFP) cages and a dedicated management port for NVIDIA UFM® (Unified Fabric Manager) connectivity.

With this very high radix, a two-level fat tree topology can connect up to 10,368 network interface cards (NICs) at lowest latency while keeping maximum job locality.

NVIDIA Quantum Infiniband Q3401-RD

The Q3401-RD delivers the same industry-leading capabilities as the Q3400-RA, featuring 144 800 Gb/s ports across 72 OSFP cages and integrated support for NVIDIA UFM via a dedicated management port. Designed for power-conscious deployments, the Q3401-RD replaces the standard AC inlet with a high-efficiency 48–54V DC busbar input. This streamlines power distribution and reduces energy loss in dense, high-performance environments.

NVIDIA Quantum-X InfiniBand Q3200-RA

For smaller platforms or existing infrastructures, the NVIDIA Quantum-X800 Q3200 2U air-cooled switch is ideal. It houses two independent switches, each with 36 ports at 800Gb/s. The Q3200 switches efficiently connect new compute clusters to previous-generation Quantum and Quantum-2 InfiniBand storage.

NVIDIA Quantum-X Photonics Q3450-LD

NVIDIA is advancing data center networking into the agentic AI era with co-packaged optics (CPO). By integrating silicon photonics directly with the InfiniBand switch ASIC, the Q3450-LD eliminates pluggable optical transceivers—reducing electrical loss, enhancing signal integrity, and improving overall power and thermal efficiency.

With 144 800 Gb/s ports connected via direct multi-fiber push-on (MPO) connectors, the Q3450-LD delivers unmatched port density and radix for high-performance AI fabrics. CPO technology shortens the high-speed electrical path to just a few millimeters within the substrate, slashing insertion loss to ~4 decibels (dB)—compared to 22 dB in traditional pluggable designs. This results in 63x better signal integrity, enabling higher data rates with lower digital signal processing (DSP) complexity and reduced power per bit.

As part of the NVIDIA Quantum-X800 InfiniBand platform, the Q3450-LD is purpose-built for AI workloads that demand ultra-low latency, high bandwidth, and deterministic performance across many thousands of GPUs. It also simplifies thermal design and cable management, accelerating the deployment and scaling of power-efficient, InfiniBand-based AI factories.

NVIDIA ConnectX-8 and ConnectX-9 SuperNIC

The ConnectX SuperNIC family delivers the high-performance, secure, and scalable networking foundation required for modern AI and HPC infrastructures. Designed to meet the demands of multi-tenant, generative AI clouds and gigascale training environments, ConnectX-8 and ConnectX-9 SuperNICs extend NVIDIA's leadership in network acceleration, in-network computing, and data center efficiency.

- **Self-Healing Interconnect:** The interconnect enhances the resilience and reliability of the NVIDIA Quantum-X800 InfiniBand network, ensuring consistent network integrity.
- **Full Offload Capabilities:** Remote direct-memory access (RDMA), NVIDIA GPUDirect® RDMA, and GPUDirect Storage maximize investment returns.
- **Advanced Power Efficiency:** Power capping and low-power state transition decrease power consumption during idle periods.
- **Co-Packaged Optics:** The NVIDIA Quantum-X Photonics switch provides better power efficiency, higher network resiliency, and faster time to deployment.

ConnectX-8 and ConnectX-9 SuperNICs leverage NVIDIA's next-generation adapter architecture to deliver unparalleled end-to-end 800 Gb/s networking with performance isolation, essential for efficiently managing multi-tenant, generative AI clouds. Both provide 800 Gb/s data throughput with PCIe Gen6, offering up to 48 lanes for various use cases such as PCIe switching inside NVIDIA GPU systems. Both SuperNICs support NVIDIA In-Network Computing, MPI_Alltoall, as well as fabric enhancement features like quality of service and congestion control. ConnectX-9 SuperNICs deliver up to 1.6 terabits per second (Tb/s) of throughput to NVIDIA Rubin GPUs.

ConnectX-8 and ConnectX-9 SuperNICs feature single-port OSFP224 and dual-port QSFP112 connectors, are compatible with various form factors, including OCP 3.0 and Card Electromechanical (CEM) PCIe x16, and support NVIDIA Socket Direct™ 16-lane auxiliary card expansion.

Cables and Transceivers

The NVIDIA Quantum-X800 platform connectivity options with the NVIDIA LinkX® interconnect portfolio of products provide the maximum flexibility to build a preferred network topology. This is achieved by using connectorized twin-port single-mode 2xDR4 and 2xFR4 transceivers with passive fiber cables, as well as linear active copper cables (LACCs).

Advanced UFM Management

In addition to the operational disruption of security threats, keeping a data center intact and running smoothly is critical. The UFM platform includes the InfiniBand subnet manager (SM) that acts as the software-defined network (SDN) controller of the InfiniBand cluster. It enables data center operators to effectively set up, monitor, manage, and proactively diagnose issues with their InfiniBand data center fabric. The UFM platform has a comprehensive feature set that can satisfy the widest range of modern, scale-out data center needs to achieve the highest usage of fabric resources.

NVIDIA is pioneering innovations at data center scale, offering the most energy-efficient networking platforms with unparalleled bandwidth, ultra-low latency, and CPU utilization, setting industry benchmarks for performance and efficiency.

Ready to Get Started?

Learn more by contacting an NVIDIA sales representative:
nvidia.com/contact-sales

NVIDIA