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 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®-8 SuperNIC™, delivering 800G connectivity to the host, with advanced offload and quality-of-service enhancements.

Leveraging advanced hardware-based 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, MPI_Alltoall, and programmable cores boost NVIDIA In-Network Computing.
- **Adaptive routing:** The switch and ConnectX-8 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 efficency.
- **Self-healing interconnect:** The interconnect enhances the resilience and reliability of the NVIDIA Quantum-X800 InfiniBand network, ensuring consistent network integrity.

NVIDIA Quantum-X800 Q3400-RA 4U

The groundbreaking Q3400-RA 4U switch—the first to utilize 200Gb/s-per-lane serializer/deserializer (SerDes) technology—significantly boosts network performance and bandwidth. It includes 144 800Gb/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. The Q3400 is air-cooled and compatible with standard 19-inch rack cabinets. A parallel liquid-cooled system, Q3400-LD, fitting an Open Compute Project (OCP) 21-inch rack, is offered as well.

NVIDIA ConnectX-8 SuperNIC

The NVIDIA ConnectX-8 SuperNIC leverages NVIDIA's next-generation adapter architecture to deliver unparallelled end-to-end 800Gb/s networking with performance isolation, essential for efficiently managing multi-tenant generative AI clouds. It provides 800Gb/s data throughput with PCI Express (PCIe) Gen6, offering up to 48 lanes for various use cases such as PCIe switching inside NVIDIA GPU systems. It also supports advanced NVIDIA In-Network Computing, MPI_Alltoall, and MPI tag-matching hardware engines, as well as fabric enhancement features like quality of service and congestion control.

The ConnectX-8 SuperNIC, featuring single-port OSFP224 and dual-port quad small form-factor pluggable (QSFP) 112 connectors for the adapters, is compatible with various form factors, including OCP 3.0 and Card Electromechanical (CEM) PCIe x16. ConnectX-8 SuperNIC also supports 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.

> **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.

> "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/en-us/contact/sales