Datasheet

NVIDIA Quantum-X800 InfiniBand Switches

Accelerate AI workloads with 800G InfiniBand.

NVIDIA Quantum-X800 InfiniBand switches deliver 800 gigabits per second (Gb/s) throughput, with ultra-low latency and advanced NVIDIA In-Network Computing, which is essential for handling trillion-parameter-scale generative AI.

These switches incorporate advanced features, including remote direct-memory access (RDMA), the fourth-generation NVIDIA® Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™, adaptive routing, telemetry-based congestion control, and self-healing technologies. Such enhancements elevate overall application performance within high-performance computing (HPC) and AI data centers.

Key Benefits

- **Highest scale for AI:** NVIDIA Quantum-X800 switches enable 2X faster speeds and 5X higher scalability for AI compute fabrics. A two-tier Quantum-X800 fat-tree topology can support over 10,000 800Gb/s host connections.
- **Next-generation In-Network Computing:** Quantum-X800 switches support SHARP for efficient offloading of compute operations to the network, boosting performance by up to 9X. The fourth-generation of SHARP adds support for FP8 precision and new collective operations such as ReduceScatter and ScatterGather.
- **Higher effective bandwidth and performance isolation:** Quantum-X800 switches support enhanced adaptive routing and telemetry-based congestion control. This enables nearly perfect effective bandwidth, as well as performance isolation for multi-tenant and multi-job environments.
- **Enhanced software operations:** Quantum-X800 switches include NVIDIA Networking OS Software (NVOS) for comprehensive chassis management and system configuration. NVOS supports various interfaces, including a command-line interface (CLI), REST APIs, Simple Network Management Protocol (SNMP), and gRPC Network Management Interface (gNMI) telemetry.

Key Features

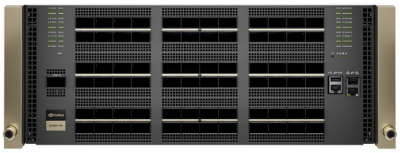
- 800Gb/s speed
- Ultra-low latency
- 4th generation NVIDIA SHARP
- Self-healing technology
- Enhanced adaptive routing
- Telemetry-based congestion control
- Communications frameworks offloads

> "NVIDIA Quantum-X800 InfiniBand switches are pivotal for achieving trillion-parameter-scale generative AI."

NVIDIA Quantum-X800 Switches

The NVIDIA Quantum-X800 Q3400-RA 4U switch, the first to leverage 200Gb/s-per-lane serializer/deserializer (SerDes) technology, significantly enhances network performance and bandwidth. It features 144 ports at 800Gb/s distributed across 72 octal small form-factor pluggable (OSFP) cages. The switch's high radix supports a two-level fat-tree topology capable of connecting up to 10,368 network interface cards (NICs) with minimal latency and optimal job locality, as well as other topologies providing connectivity to tens of thousands of GPUs. The Q3400 is air-cooled. Data centers that use liquid cooling can opt for the liquid-cooled system, Q3400-LD, as well.

NVIDIA Quantum-X800 Q3200 InfiniBand switch



NVIDIA Quantum-X800 Q3400-RA InfiniBand switch

For smaller-scale platforms or integration with existing infrastructures, the NVIDIA Quantum-X800 Q3200 2U air-cooled configuration switch is ideal. This system houses two independent Switches within a single enclosure, each providing 36 ports at 800Gb/s. The Q3200 fixed-configuration switches are well-suited for connecting new compute clusters to previous-generation Quantum and Quantum-2 InfiniBand storage infrastructure.

Both the Q3400 and Q3200 switches include a dedicated InfiniBand in-band management port specifically for NVIDIA Unified Fabric Manager (UFM®) management, separated on the front panel from the other ports. This separation allows the full set of standard ports to be used for data network connectivity, simplifying port allocation and streamlining topology design.

Additionally, NVIDIA Quantum-X800 switches feature optional router capabilities, facilitating the expansion of InfiniBand clusters to support a large scale of nodes located across multiple sites.

NVIDIA Quantum-X800 InfiniBand Platform

The Quantum-X800 InfiniBand platform includes the Q3400 and Q3200 switches, the NVIDIA ConnectX®-8 SuperNIC and NVIDIA LinkX® interconnect portfolio of transceivers and cables. The platform achieves end-to-end throughput of 800Gb/s from switch to host. For fabric-scale platform management and monitoring, Quantum-X800 features UFM, which enables true software-defined networking with powerful visibility and insights into the performance and health of the network. This end-to-end network platform is purpose-built to deliver the highest performance for the scale-out compute fabrics, enabling massive-scale AI.

System Specifications			
	Q3200-RA	**Q3400-LD**	**Q3400-RA**
Performance	Two switches, each of 28.8Tb/s throughput	115.2Tb/s throughput	115.2Tb/s throughput
Switch radix	Two switches, each of 36 800Gb/s non-blocking ports	144 800Gb/s non-blocking ports	144 800Gb/s non-blocking ports
Connectors and cabling	Two groups of 18 OSFP connectors	72 OSFP connectors	72 OSFP connectors
Management ports	Separate 400Gb/s InfiniBand in-band management port (UFM)	Separate 400Gb/s InfiniBand in-band management port (UFM)	Separate 400Gb/s InfiniBand in-band management port (UFM)
CPU	Intel CFL 4 Cores i3-8100H 3GHz	Intel CFL 4 Cores i3-8100H 3GHz	Intel CFL 4 Cores i3-8100H 3GHz
Security	CPU/CPLD/Switch IC based on IRoT	CPU/CPLD/Switch IC based on IRoT	CPU/CPLD/Switch IC based on IRoT
Software	NVOS	NVOS	NVOS
Rack mount	2U	4U	4U
Cooling Mechanism	Air-cooled	Liquid-cooled	Air-cooled
EMC (emissions)	CE, FCC, VCCI, ICES, and RCM	CE, FCC, VCCI, ICES, and RCM	CE, FCC, VCCI, ICES, and RCM
Product safety compliant/certified	RoHS, CB, cTUVus, CE, and CU	RoHS, CB, cTUVus, CE, and CU	RoHS, CB, cTUVus, CE, and CU
Warranty	One year	One year	One year

Ready to Get Started?

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

NVIDIA