Datasheet



NVIDIA Quantum-2 QM9700 Series

Scaling out data centers with 400G InfiniBand smart switches.

Accelerate Research and Product Innovation with Enhanced Data and In-Network Computing

As high-performance computing (HPC) and artificial intelligence (AI) applications increase in complexity, the demand for advanced high-speed networking is critical for extreme-scale systems. NVIDIA Quantum-2, a premier switch platform, excels in power and density, offering 400 gigabits per second (Gb/s) of InfiniBand throughput. This high networking performance is essential for AI developers and scientific researchers taking on the world's most challenging problems.

Advanced Computing Needs Advanced Networking

The NVIDIA Quantum-2-based QM9700 series of switch systems deliver 64 ports of 400Gb/s InfiniBand per port, packed into a 1U standard chassis. Each switch carries an impressive 51.2 terabits per second (Tb/s) of aggregated bidirectional throughput and a landmark capacity exceeding 66.5 billion packets per second (BPPS). Supporting NVIDIA's advanced 400Gb/s interconnect technology, NVIDIA Quantum-2 is a high-speed, extremely low-latency, and scalable networking solution. Key features incorporated include state-of-the-art technologies such as remote direct-memory access (RDMA), adaptive routing, and NVIDIA® Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™.

Unlike other networking solutions, NVIDIA InfiniBand incorporates self-healing network capabilities, quality of service, enhanced virtual lane mapping, and advanced congestion control, maximizing overall application throughput. As ideal rack-mounted InfiniBand solutions, the QM9700 series of 400Gb/s InfiniBand fixed-configuration switches offer exceptional flexibility, with support for various topologies like fat tree, Slim Fly, Dragonfly+, multi-dimensional torus, and others. They maintain backward compatibility with previous generations and support a broad software ecosystem.

The Era of Data-Driven Computing

Today's complex research demands ultra-fast processing for high-resolution simulations, massive datasets, and highly parallelized algorithms with real-time information exchanges. The QM9700 series of 400Gb/s InfiniBand switches, enhanced with NVIDIA In-Network Computing technologies, incorporate the third generation of NVIDIA SHARP technology, SHARPv3. This innovation allows virtually unlimited scalability for small and large data aggregation across the network. SHARPv3 delivers AI acceleration that's 32x higher than its predecessor, significantly enhancing the performance of complex computational applications as data moves through the data center network. It actively participates in runtime processes and reduces the volume of data needed across the network.

Streamlining Network Design and Topologies

The NVIDIA Quantum-2 QM9700 family of InfiniBand switches supports 64 ports of 400Gb/s NDR InfiniBand or up to 128 ports of 200Gb/s via port-splitting in a 1U chassis. This density and performance allows deployments, from small to extremely large scale, to utilize various topologies including a two-level fat tree (often for smaller to medium scale) while providing benefits like reduced power consumption, ultra-low latency, and high space efficiency.

> "NVIDIA SHARPv3 delivers AI acceleration that's 32X higher than its predecessor, significantly enhancing the performance of complex computational applications as data moves through the data center network."

Integrated Router Capabilities

NVIDIA Quantum-2 InfiniBand switches with optional router capabilities enable clusters to scale to over 40K nodes in a single fabric. This capability significantly surpasses the scaling limits of the previous generation to sustain the peak performance and reliability demands of research, simulations, AI, and data processing for cloud applications.

Enhanced Management

The internally managed QM9700 and QM9701 switches feature an on-board subnet manager that enables simple, out-of-the-box bringup for up to 2,000 nodes. Running the NVIDIA MLNX-OS® software package, the subnet manager delivers full chassis management through command-line interface (CLI), web-based user (WebUI), Simple Network Management Protocol (SNMP), or JavaScript Object Notation (JSON) interfaces.

The externally managed QM9790 switch can utilize the advanced NVIDIA Unified Fabric Manager (UFM®) feature sets to empower data center operators to efficiently provision, monitor, manage, preventatively troubleshoot, and maintain the modern data center fabric, realizing higher utilization and reducing overall operational costs.

Energy Efficiency

NVIDIA Quantum-2 9700 series switches can be grouped and connected with short-reach cables to create a virtual modular switch. NVIDIA systems use copper instead of optical transceivers for better energy efficiency, cutting power usage from 30 watts to 0–3 watts. This reduction improves latency and reliability, lowering the total cost of ownership. The layout includes InfiniBand switches linked by copper cables in a fat tree topology, between either levels 1 and 2 or levels 2 and 3.

System Specifications*

Performance	400Gb/s per port
Switch radix	64 400Gb/s non-blocking ports with aggregate data throughput up to 51.2Tb/s
Connectors and cabling	32 octal small form-factor pluggable (OSFP) connectors; passive or active copper or active fiber cable; optical module
Power supply	QM9700/QM9790: > Input range: 200–240VAC > US min: 2 phases of 100–100V-in total, at least 208V > 1+1 redundant and hot-swappable power > 80 Gold+ and ENERGY STAR certified QM9701: > DC bus bar > Input: 48VDC
Cooling	Front-to-rear or rear-to-front Cooling option: 6+1 hot-swappable fan unit
Management ports	> 1x USB 3.0 x1 > 1x USB for I2C channel 1x RJ45 > 1x RJ45 > 1x RJ45 (UART)
CPU	x86 Coffee Lake i3
System memory	Single 8GB, 2,666 mega transfers per second (MT/s), DDR4 SO-DIMM
Storage	M.2 SSD SATA 16GB 2242 FF
Software	MLNX-OS
System weight	14.5kg
System dimensions	> Height: 1.7in (43.6 mm) > Width: 17.0in (438 mm) > Depth: 26.0in (660.4 mm)
Rack mount	1U rack mount
Environmental conditions	Temperature: > Operational: Forward air flow: 0–35°C Reverse air flow: 0–40°C Non-operational: -40–70°C > Humidity: Operating 10–85% non-condensing Non-operating 10–90% non-condensing > Altitude: Up to 3,050m
EMC (emissions)	CE, FCC, VCCI, ICES, and RCM
Product safety/compliant/certified	RoHS, CB, cTUVus, CE, and CU
Warranty	One year

*This section describes hardware features and capabilities. For feature availability, refer to the firmware release notes.

Enterprise Support and Services

A minimum one-year Enterprise Business-Standard Support entitlement is required when purchasing NVIDIA Quantum-2 InfiniBand switches.

- Enterprise Support provides access to NVIDIA experts, the NVIDIA Enterprise Support Portal, advanced return material authorization (RMA), and more.
- Add-on services—including installation, configuration, a technical account manager, a four-hour on-site engineer, expedited RMA, media retention, and more—are available.

For details, visit the NVIDIA Enterprise Support and Services User Guide.

Product Specifications

Switches Portfolio

For details on the NVIDIA Quantum-2 InfiniBand series switches, along with their ordering part numbers, please refer to the QM97XX 1U NDR 400Gbps InfiniBand Switch Systems User Manual.

Transceivers and Cables

For details on NVIDIA cables and transceivers, visit the Interconnect documentation hub.

Ready to Get Started?

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

© 2025 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, MLNX-OS, Scalable Hierarchical Aggregation and Reduction Protocol (SHARP), and UFM are trademarks and/or registered trademarks of NVIDIA Corporation 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. 4519551. NOV25