NVIDIA Quantum-2 Takes Supercomputing to New Heights, Into the Cloud

New 400Gbps InfiniBand Switch and Networking Platform Features Secure, Cloud-Native, Multi-Tenant, Bare-Metal Performance for AI, Data Analytics, HPC Applications

GTC—NVIDIA today announced NVIDIA Quantum-2, the next generation of its InfiniBand networking platform, which offers the extreme performance, broad accessibility and strong security needed by cloud computing providers and supercomputing centers.

The most advanced end-to-end networking platform ever built, NVIDIA Quantum-2 is a 400Gbps InfiniBand networking platform that consists of the NVIDIA Quantum-2 switch, the ConnectX-7® network adapter, the BlueField-3® data processing unit (DPU) and all the software that supports the new architecture.

The introduction of NVIDIA Quantum-2 comes as supercomputing centers are increasingly opening to multitudes of users, many from outside their organizations. At the same time, the world's cloud service providers are beginning to offer more supercomputing services to their millions of customers.

NVIDIA Quantum-2 includes key features required for demanding workloads running in either arena. Supercharged by cloud-native technologies, it provides high performance with 400 gigabits per second of throughput and advanced multi-tenancy to accommodate many users.

"The requirements of today's supercomputing centers and public clouds are converging," said Gilad Shainer, senior vice president of Networking at NVIDIA. "They must provide the greatest performance possible for next-generation HPC, AI and data analytics challenges, while also securely isolating workloads and responding to varying demands of user traffic. This vision of the modern data center is now real with NVIDIA Quantum-2 InfiniBand."

NVIDIA Quantum-2 Performance and Cloud-Native Capabilities
With 400Gbps, NVIDIA Quantum-2 InfiniBand doubles the network speed and triples the number of network ports. It accelerates performance by 3x and reduces the need for data center fabric switches by 6x, while cutting data center power consumption and reducing data center space by 7 percent each.

The multi-tenant performance isolation of NVIDIA Quantum-2 keeps the activity of one tenant from disturbing others, utilizing an advanced telemetry-based congestion control system with cloud-native capabilities that ensure reliable throughput, regardless of spikes in users or workload demands.

NVIDIA Quantum-2 SHARPv3™ In-Network Computing technology provides 32x more acceleration engines for AI applications compared with the previous generation. Advanced InfiniBand fabric management for data centers, including predictive maintenance, is enabled with the NVIDIA UFM® Cyber-AI platform.

A nanosecond-precision timing system integrated into NVIDIA Quantum-2 can synchronize distributed applications, like database processing, helping to reduce the overhead of wait and idle times. This new capability allows cloud data centers to become part of the telecommunications network and host software-defined 5G radio services.

NVIDIA Quantum-2 InfiniBand Switch
At the heart of the NVIDIA Quantum-2 platform is the new Quantum-2 InfiniBand switch. With 57 billion transistors on 7-nanometer silicon, it is slightly bigger than the NVIDIA A100 GPU with 54 billion transistors.

It features 64 ports at 400Gbps or 128 ports at 200Gbps and will be offered in a variety of switch systems up to 2,048 ports at 400Gbps or 4,096 ports at 200Gbps — more than 5x the switching capability over the previous generation, Quantum-1.

The combined networking speed, switching capability and scalability is ideal for building the next-generation of giant HPC systems.

The NVIDIA Quantum-2 switch is now available from a wide range of leading infrastructure and system vendors around the world, including Atos, DataDirect Networks (DDN), Dell Technologies, Excelero, GIGABYTE, HPE, IBM, Inspur, Lenovo, NEC, Penguin Computing, QCT, Supermicro, VAST Data and WekaIO.

NVIDIA Quantum-2, ConnectX-7 and BlueField-3
The NVIDIA Quantum-2 platform provides two networking end-point options, the NVIDIA ConnectX-7 NIC and NVIDIA BlueField-3 DPU InfiniBand.

ConnectX-7, with 8 billion transistors in a 7-nanometer design, doubles the data rate of the world's current leading HPC

networking chip, the NVIDIA ConnectX-6. It also doubles the performance of RDMA, GPUDirect® Storage, GPUDirect RDMA and In-Networking Computing. The ConnectX-7 will sample in January.

BlueField-3 InfiniBand, with 22 billion transistors in a 7-nanometer design, offers sixteen 64-bit Arm CPUs to offload and isolate the data center infrastructure stack. BlueField-3 samples in May.

About NVIDIA
NVIDIA's (NASDAQ: NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market and has redefined modern computer graphics, high performance computing and artificial intelligence. The company's pioneering work in accelerated computing and AI is reshaping trillion-dollar industries, such as transportation, healthcare and manufacturing, and fueling the growth of many others. More information at https://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, features, performance and availability of NVIDIA Quantum-2, the NVIDIA Quantum-2 switch, the ConnectX-7 network adapter and the BlueField-3 data processing unit; supercomputing centers increasingly opening to multitudes of users; cloud service providers beginning to offer more supercomputing services to their millions of customers; and the requirements of today's supercomputing centers and public clouds and their convergence are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

Alex Shapiro
Enterprise Networking
1-415-608-5044
ashapiro@nvidia.com