NVIDIA Announces Mellanox InfiniBand for Exascale AI Supercomputing

Global Ecosystem of Server and Storage Partners to Offer Systems with NVIDIA Mellanox 400G, World's Only Fully In-Network Acceleration Platform

SC20—NVIDIA today introduced the next generation of NVIDIA® Mellanox® 400G InfiniBand, giving AI developers and scientific researchers the fastest networking performance available to take on the world's most challenging problems.

As computing requirements continue to grow exponentially in areas such as drug discovery, climate research and genomics, NVIDIA Mellanox 400G InfiniBand is accelerating this work through a dramatic leap in performance offered on the world's only fully offloadable, in-network computing platform.

The seventh generation of Mellanox InfiniBand provides ultra-low latency and doubles data throughput with NDR 400Gb/s and adds new NVIDIA In-Network Computing engines to provide additional acceleration.

The world's leading infrastructure manufacturers — including Atos, Dell Technologies, Fujitsu, GIGABYTE, Inspur, Lenovo and Supermicro — plan to integrate NVIDIA Mellanox 400G InfiniBand into their enterprise solutions and HPC offerings. These commitments are complemented by extensive support from leading storage infrastructure partners including DDN and IBM Storage, among others.

"The most important work of our customers is based on AI and increasingly complex applications that demand faster, smarter, more scalable networks," said Gilad Shainer, senior vice president of networking at NVIDIA. "The NVIDIA Mellanox 400G InfiniBand's massive throughput and smart acceleration engines let HPC, AI and hyperscale cloud infrastructures achieve unmatched performance with less cost and complexity."

Today's announcement builds on Mellanox InfiniBand's lead as the industry's most robust solution for AI supercomputing. The NVIDIA Mellanox NDR 400G InfiniBand offers 3x the switch port density and boosts AI acceleration power by 32x. In addition, it surges switch system aggregated bi-directional throughput 5x, to 1.64 petabits per second, enabling users to run larger workloads with fewer constraints.

Expanding Ecosystem for Expanding Workloads
Early interest in the next generation of Mellanox InfiniBand is coming from some of the world's largest scientific research organizations.

"Microsoft Azure's partnership with NVIDIA Networking stems from our shared passion for helping scientists and researchers drive innovation and creativity through scalable HPC and AI. In HPC, Azure HBv2 VMs are the first to bring HDR InfiniBand to the cloud and achieve supercomputing scale and performance for MPI customer applications with demonstrated scaling to eclipse 80,000 cores for MPI HPC," said Nidhi Chappell, head of product, Azure HPC and AI at Microsoft Corp. "In AI, to meet the high-ambition needs of AI innovation, the Azure NDv4 VMs also leverage HDR InfiniBand with 200Gb/s per GPU, a massive total of 1.6Tb/s of interconnect bandwidth per VM, and scale to thousands of GPUs under the same low-latency InfiniBand fabric to bring AI supercomputing to the masses. Microsoft applauds the continued innovation in NVIDIA's Mellanox InfiniBand product line, and we look forward to continuing our strong partnership together."

"High-performance interconnects are cornerstone technologies required for exascale and beyond. Los Alamos National Laboratory continues to be at the forefront of HPC networking technologies," said Steve Poole, chief architect for next-generation platforms at Los Alamos National Laboratory. "The Lab will continue their relationship working with NVIDIA in evaluating and analyzing their latest 400Gb/s technology aimed at solving the diverse workload requirements at Los Alamos."

"Amid the new age of exascale computing, researchers and scientists are pushing the limits of applying mathematical modeling to quantum chemistry, molecular dynamics and civil safety," said Professor Thomas Lippert, head of the Jülich Supercomputing Centre. "We are committed to leveraging the next generation of Mellanox InfiniBand to further our track record of building Europe's leading, next-generation supercomputers."

"InfiniBand continues to maintain its pace of innovation and performance, underlining the differentiation that has made it the most commonly used high-performance server and storage interconnect for HPC and AI systems," said Addison Snell, CEO of Intersect360 Research. "As applications continue to demand increased network throughput, the need for high-performance solutions, such as NVIDIA Mellanox 400G InfiniBand, has the potential to keep expanding into new use cases and markets."

Product Specifications and Availability
Offloading operations is crucial for AI workloads. The third-generation NVIDIA Mellanox SHARP technology allows deep

learning training operations to be offloaded and accelerated by the InfiniBand network, resulting in 32x higher AI acceleration power. When combined with NVIDIA Magnum IO™ software stack, it provides out-of-the-box accelerated scientific computing.

Edge switches, based on the Mellanox InfiniBand architecture, carry an aggregated bi-directional throughput of 51.2Tb/s, with a landmark capacity of more than 66.5 billion packets per second. The modular switches, based on Mellanox InfiniBand, will carry up to an aggregated bi-directional throughput of 1.64 petabits per second, 5x higher than the last generation.

The Mellanox InfiniBand architecture is based on industry standards to ensure backwards and future compatibility and protect data center investments. Solutions based on the architecture are expected to sample in the second quarter of 2021.

Learn more about NVIDIA Mellanox InfiniBand in the live NVIDIA SC20 Special Address at 3 p.m. PT today.

About NVIDIA
NVIDIA's (NASDAQ: NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics and revolutionized parallel computing. More recently, GPU deep learning ignited modern AI — the next era of computing — with the GPU acting as the brain of computers, robots and self-driving cars that can perceive and understand the world. More information at https://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the benefits, performance and abilities of the NVIDIA Mellanox InfiniBand architecture, NVIDIA NDR 400Gb/s, and NVIDIA Mellanox SHARP technology; what our technology enables; NDR 400Gb/s InfiniBand allowing infrastructures to achieve unmatched performance with less cost and complexity; the early interest for NDR 400Gb/s InfiniBand; NDR 400Gb/s InfiniBand being a needle mover for Microsoft Azure and Microsoft's reliance on InfiniBand to deliver the best performance and cost; Los Alamos National Laboratory being at the forefront of HPC networking technologies and it continuing its relationship working with NVIDIA and its products; researchers and scientists pushing the limits of mathematical modeling; the commitment to leverage the next generation of NVIDIA NDR 400Gb/s InfiniBand; the networking manufacturers planning to integrate NVIDIA NDR 400Gb/s InfiniBand solutions and support from storage infrastructure partners; the potential for high-performance solutions to keep expanding into new use cases and markets; and the availability of solutions based on the architecture 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