Mellanox HDR 200G InfiniBand Deep Learning Acceleration Engines Demonstrates Two Times Higher Performance for Artificial Intelligence (AI) Platforms with NVIDIA

Mellanox In-Network Computing "Hierarchical Aggregation and Reduction Protocol" (SHARP)â„¢ Technology in Combination with NVIDIA Collective Communications Library (NCCL) Delivers Performance Breakthrough to AI

Mellanox Technologies, Ltd., a leading supplier of high-performance, end-to-end smart interconnect solutions for data center servers and storage systems, today announced that its HDR 200G InfiniBand with the "Scalable Hierarchical Aggregation and Reduction Protocol" (SHARP)™ technology has set new performance records, doubling deep learning operations performance. The combination of Mellanox In-Network Computing SHARP with NVIDIA® V100 Tensor Core GPU technology and Collective Communications Library (NCCL) deliver leading efficiency and scalability to deep learning and artificial intelligence applications.

The combination of the state-of-the-art NVIDIA GPUs, Mellanox's InfiniBand, GPUDirect RDMA and NCCL to train neural networks has already become a de-facto standard when scaling out deep learning frameworks, such as Caffe, Caffe2, Chainer, MXNet, TensorFlow, and PyTorch. With the Mellanox SHARP technology and HDR InfiniBand, deep learning training's data aggregation operations can be offloaded and accelerated by the InfiniBand network, resulting in improving their performance by two times.

The joint effort with NVIDIA and testing performed in Mellanox's performance labs, using the Mellanox HDR InfiniBand Quantum connecting four system hosts, each with eight NVIDIA V100 Tensor Core GPUs with NVLink interconnect technology and a single ConnectX-6 HDR adapter per host, have achieved an effective reduction bandwidth of 19.6GB/s by integrating SHARP's native streaming aggregation capability with NVIDIA's latest NCCL 2.4 library, which now takes full advantage of the bi-directional bandwidth available from the Mellanox interconnect. This implementation is effectively two times higher bandwidth than NVIDIA's current tree-based implementation using the same hardware configuration.

In the more common setup for this configuration, four HCAs in each system host are used for balanced performance across a variety of workloads where the initial SHARP and NCCL results yielded an expected 70.3GB/s. For more densely populated GPU-based systems, like NVIDIA DGX-2, which houses 16 NVIDIA V100 Tensor Core GPUs with NVLink in each system node, the in-network capabilities and available bidirectional bandwidth of the Mellanox fabric can be fully leveraged.

"Our long-standing collaboration with NVIDIA has again delivered a robust solution that takes full advantage of the best-of-breed capabilities from Mellanox InfiniBand, including GPUDirect RDMA and now extending in-network computing to NCCL, which delivers two times better performance for AI," said Gilad Shainer, Vice President of Marketing at Mellanox Technologies. "HDR InfiniBand in-network computing acceleration engines, including the SHARP technology, provide the highest performance and scalability for HPC and AI workloads."

"Mellanox solutions amplify NVIDIA's unmatched CUDA-X acceleration libraries using NCCL, our open source collective communication library," said Ian Buck, vice president and general manager of Accelerated Computing at NVIDIA. "Together, we offer solutions that ensure the most demanding AI applications in the data center benefit from cutting-edge performance and scaling efficiency."

Supporting Resources:

- Learn More about Mellanox SHARP™
- Learn more about Mellanox Quantum™ HDR 200Gb/s InfiniBand Smart Switches
- Follow Mellanox on: Twitter, Facebook, LinkedIn, and YouTube
- Join the Mellanox Community

About Mellanox
Mellanox Technologies is a leading supplier of end-to-end Ethernet and InfiniBand smart interconnect solutions and services for servers and storage. Mellanox interconnect solutions increase data center efficiency by providing the highest throughput and lowest latency, delivering data faster to applications, unlocking system performance and improving data security. Mellanox offers a choice of fast interconnect products: adapters, switches, software and silicon that accelerate application performance and maximize business results for a wide range of markets including cloud and hyperscale, high performance

computing, artificial intelligence, enterprise data centers, cyber security, storage, financial services and more.

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