



Mellanox Works with VMware and NVIDIA to Enable High Performance Virtualized Machine Learning Solutions

Advanced Paravirtualization over RDMA (PVRDMA) Networking Technology and NVIDIA's Latest vGPU Solution Enable Efficient Machine Learning Deployments

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 RDMA (Remote Direct Memory Access) networking solutions for VMware vSphere enable virtualized Machine Learning solutions that achieve higher GPU utilization and efficiency. Benchmarks demonstrate that the NVIDIA® Virtual Compute Server (vCS) for virtualized GPUs achieve two times better efficiency by using VMware's paravirtualized RDMA (PVRDMA) technology than when using traditional networking protocols. The benchmark was performed on a four-node cluster running vSphere 6.7 equipped with NVIDIA T4 GPUs with vCS software and Mellanox ConnectX-5 100 GbE SmartNICs, all connected by a Mellanox Spectrum SN2700 100 GbE switch.

The PVRDMA Ethernet solution enables VM-to-VM communication over RDMA, which boosts data communication performance in virtualized environments while achieving significantly higher efficiency compared with legacy TCP/IP transports. Additionally, PVRDMA retains core virtual machine capabilities such as vMotion. This translates to real-world customer advantages including optimized server and GPU utilization, reduced machine learning training time and improved scalability. Using PVRDMA also shrinks backup times, improves data center simplicity, simplifies consolidation, lowers power consumption and reduces total cost of ownership.

"As Moore's Law has slowed, traditional CPU and networking technologies are no longer sufficient to support the emerging machine learning workloads," said Kevin Deierling, vice president marketing, Mellanox Technologies. "Using hardware compute accelerators such as NVIDIA T4 GPUs and Mellanox's RDMA networking solutions has proven to boost application performance in virtualized deployments."

NVIDIA T4 GPUs supercharge the world's most trusted mainstream servers, easily fitting into standard data center infrastructures. Their low-profile, 70-watt design is powered by NVIDIA Turing™ Tensor Cores, delivering revolutionary multi-precision performance to accelerate a wide range of modern applications, including machine learning, deep learning, and virtual desktops. With the latest Virtual Compute Server software for GPU virtualization, it also provides maximum performance and manageability for AI, ML and data science workloads in a virtualized server environment.

"Machine learning has become extremely important and every company, regardless of size, must leverage its power to remain competitive," said Bob Pette, vice president, Professional Visualization, NVIDIA. "Our collaboration with VMware and Mellanox creates a high-performance GPU platform that enables acceleration for compute-intensive workloads in the most efficient way."

Machine learning workloads are extremely resource intensive, often relying on hardware acceleration to achieve the performance necessary to solve large, complex problems in a timely manner. Interconnect acceleration – special hardware that delivers extremely high bandwidth and low latency, and compute acceleration – often delivered through exploitation of very highly-parallel GPU compute engines, are the most common forms of such acceleration. While both types of acceleration have long been available on vSphere, it is now possible with vSphere to combine these technologies to support advanced machine learning applications that allow applications to combine the compute power of NVIDIA GPUs with the high-performance data transfer capabilities of Mellanox RDMA-capable adapters, enabling linear scalability.

"Modern data center infrastructures need to keep pace with the compute and efficiency requirements for the exceedingly complex machine learning computational models," said Sudhanshu (Suds) Jain, Product Management, Cloud Platform Business Units, VMware. "The ability to virtualize GPUs using the latest NVIDIA Virtual Compute Server product and Mellanox's high-speed networking solutions over vSphere makes it possible to meet those requirements while keeping the cost intact."

Availability

VMware vSphere is fully qualified with Mellanox ConnectX 10/25/40/50/100G adapters today. All Mellanox adapters support PVRDMA over RoCE (RDMA over Converged Ethernet), enabling advanced capabilities like GPU virtualization, and making data center infrastructure RoCE-Ready as new technologies over RDMA become generally available. PVRDMA will also be supported by the latest ConnectX-6 Dx and BlueField-2 SmartNICs announced today at VMworld.

Supporting Resources:

- View demonstrations and presentations at VMworld 2019 U.S. (Aug. 25 – Aug. 28)
- Visit Mellanox at booth #1463 to learn about the benefits of using Mellanox's 10/25/40/50/100/200 Gb/s Ethernet and RoCE with VMware vSphere and NVIDIA GPUs
- Visit NVIDIA at booth #1351 to learn about GPU virtualization for every workload

About Mellanox

Mellanox Technologies is a leading supplier of end-to-end InfiniBand and Ethernet 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 and unlocking system performance capability. Mellanox offers a choice of fast interconnect products: adapters, switches, cables and transceivers, software and silicon that accelerate application runtime and maximize business results for a wide range of markets including high performance computing, enterprise data centers, Web 2.0, cloud, storage and financial services.

Mellanox and ConnectX are registered trademarks of Mellanox Technologies, Ltd. LinkX is a trademark of Mellanox Technologies, Ltd. All other trademarks are property of their respective owners.

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