SOLUTION BRIEF

High Performance Virtualized Machine Learning

Mellanox, VMware and NVIDIA enable virtualized Machine Learning solutions that achieve higher GPU utilization and efficiency

Overview

As Moore's Law continues to slow, new methods of accelerating compute processes are necessary to boost application performance. Virtualization technologies have proven to be a cost-effective approach to achieving continued, efficient scaling. However, until now, machine learning solutions had not evolved to support virtualization and was bound by dedicated physical resources. Recently, VMWare developed technologies to effectively share accelerators across virtualized compute and networking to enable explicit runtimes and maximize performance in virtualized environments. VMWare, NVIDIA, and Mellanox collaborated on integrating NVIDIA vGPU within VMware vSphere environments that allows sharing of GPUs across multiple virtual machines. With the assistance of NVIDIA and Mellanox, resources in vSphere machine learning environments can now be virtualized including multiple GPUs while persevering critical features like vMotion.

GPU and RDMA Virtualization

VMware vSphere has embraced Mellanox RDMA technology in the past on Mellanox ConnectX-5 adapters where they have been successfully used to enable various use cases, such as VMware vSAN and across business-critical applications in cloud deployments. Recently, VMware vSphere extended virtualization support to the latest GPU hardware from NVIDIA. By combining the benefits of vSphere with the capabilities of high-performance hardware accelerators from Mellanox and NVIDIA, the team was able to design a compelling machine learning solution that is capable of provisioning multiple GPUs to a single VM where GPU clusters are managed by vCenter.

Improve Training with GPUs

In deep learning, reducing neural-network training is in high demand. GPUs provide a most of the horsepower in the neural training. NVIDIA has been optimizing their GPU for deep learning since 2012. Their latest GPU architecture is Turing and is available from the T4 as well as RTX 6000

KEY FEATURES:

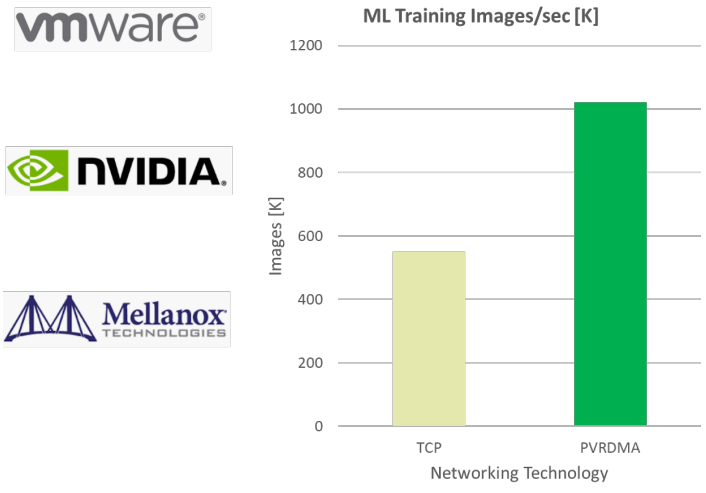
- VMware introduces out of the box ready solutions to virtualize compute and networking accelerators
- Multiple GPUs can be provision to a single VM, enabling maximum GPU acceleration
- GPUs can be successfully shared across multiple virtual machines
- Compute and Networking accelerators can be used to fuel compute horsepower after the end of Moore’s Law
- Mellanox support for PVRDMA permits linear scalability as more GPU are added
- PVRDMA boosts communication performance and efficiency in virtualized environments
- Critical vSphere features like HA and vMotion can be preserved
- Ideal for deployment with Lenovo ThinkSystem Servers

and RTX 8000 GPUs, each of which supports NVIDIA's vGPU technology for virtualization which is available through NVIDIA's vComputeServer. vComputeServer software enables virtualized NVIDIA GPUs to power more than 600 GPU accelerated applications for AI, machine learning, and HPC. With GPU sharing, multiple VMs can be powered by a single GPU to maximizing utilization, or multiple virtual GPUs can power a single VM. GPU sharing enables vSphere to power the most compute-intense machine learning workloads.

High-Speed Networking with PVRDMA

Mellanox Intelligent ConnectX-5 EN adapters enable application acceleration through paravirtualized RDMA (PVRDMA) technology to facilitates VM-to-VM communication. This boosts data transfer performance in vSphere environments and allows for significantly higher efficiency compared to legacy TCP/IP transports. Additionally, it allows retention of core virtual machine capabilities such as vMotion. The use of PVRDMA enables real-world customer advantages, including optimized server and GPU utilization, reduced machine learning training time, improved scalability and can shrink backup times.

2X HIGHER TRAINING EFFICIENCY



Distributed Machine Learning

There is significant pressure to reduce the deployment time for machine learning models, and as the datasets grow in size, this is increasing. There is an escalating need to have distributed machine learning as it stands to reduce training time and model development. Horovod is an open-source distributed training framework that supports popular machine learning applications such as TensorFlow, Keras, PyTorch, and MXNet. VMware chose Horovod as it requires minimal modification to the user code and therefore stands to reduce model development time. Benchmarks demonstrated that the NVIDIA® vComputeServer (vCS) for virtualized GPUs achieved 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.

Conclusion

Traditional CPU and TCP/IP technologies are no longer sufficient to support emerging machine learning workloads. Virtualization and hardware accelerators are one of the most efficient ways to overcome these deficiencies. VMware's support for vGPU and PVRDMA within vShpere has proven through the collectively developed proof of concept that performance and efficiency can be increased for a virtualized Horovod-based machine learning solution. The solution demonstrated excellent scalability while leveraging PVRDMA and vGPU to boost application performance while preserving critical features like vMotion. Further testing validated that even under heavy load virtual machines using vGPUs and PVRDMA can be migrated successfully. This capability improves utilization and availability while enabling maximum GPU acceleration and aids in decreasing training times in machine learning solutions.

Want To Learn More?

View the Reference Design Guide for Virtualized ML and HPC Workloads utilizing NVIDIA vGPU and VMware PVRDMA: https://docs.mellanox.com/pages/releaseview.action?pageId=18482697

Version 9.25.19