Distributed Machine Learning on vSphere leveraging NVIDIA vGPU and Mellanox PVRDMA

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001
www.vmware.com

Table of Contents

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

1 Introduction

While virtualization technologies have proven themselves in the enterprise with cost effective, scalable and reliable IT computing, High Performance Computing (HPC) however has not evolved and is still bound to dedicating physical resources to obtain explicit runtimes and maximum performance. VMWare has developed technologies to effectively share accelerators for compute and networking.

VMWare, NVIDIA and Mellanox have collaborated on NVIDIA vGPU integration with VMware vSphere that enables sharing of GPU across multiple virtual machines, while preserving critical vSphere features like vMotion. It is also possible to provision multiple GPUs to a single VM, enabling maximum GPU acceleration and utilization.

vSphere enables RDMA based high performance network sharing using Paravirtualized RDMA. PVRDMA also supports vSphere features like HA & vMotion.

2 Solution Components

2.1 GPUs for Machine Learning

With the impending end to Moore's law, the spark that is fueling the current revolution in deep learning is having enough compute horsepower to train neural-network based models in a reasonable amount of time

The needed compute horsepower is derived largely from GPUs, which NVIDIA began optimizing for deep learning since 2012. The latest GPU architecture from NVIDIA is Turing, available with T4 as well as the RTX 6000 and RTX 8000 GPUs, which all support virtualization. .

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001
www.vmware.com

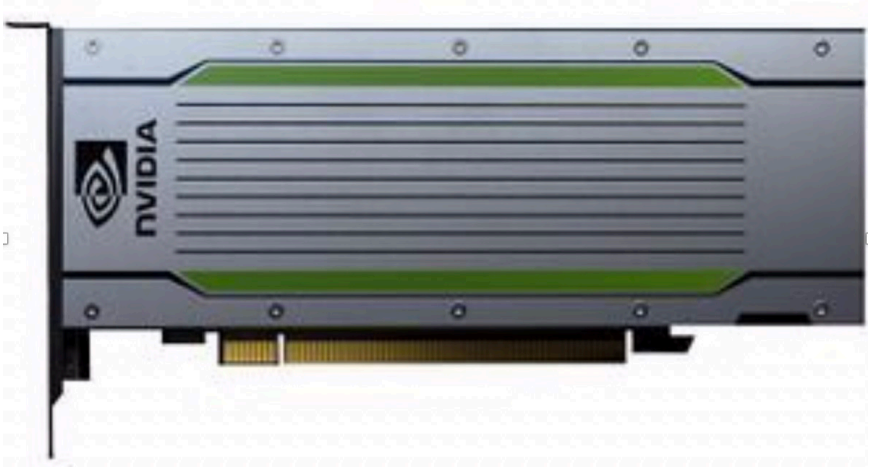


Figure 1: The NVIDIA T4 GPU

The NVIDIA® T4 GPU accelerates diverse cloud workloads, including high-performance computing, deep learning training and inference, machine learning, data analytics, and graphics. Based on the new NVIDIA Turing™ architecture and packaged in an energy-efficient 70-watt, small PCIe form factor, T4 is optimized for mainstream computing environments and features multi-precision Turing Tensor Cores and new RT Cores. Combined with accelerated containerized software stacks from NGC, T4 delivers revolutionary performance at scale. (Source: NVIDIA)

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

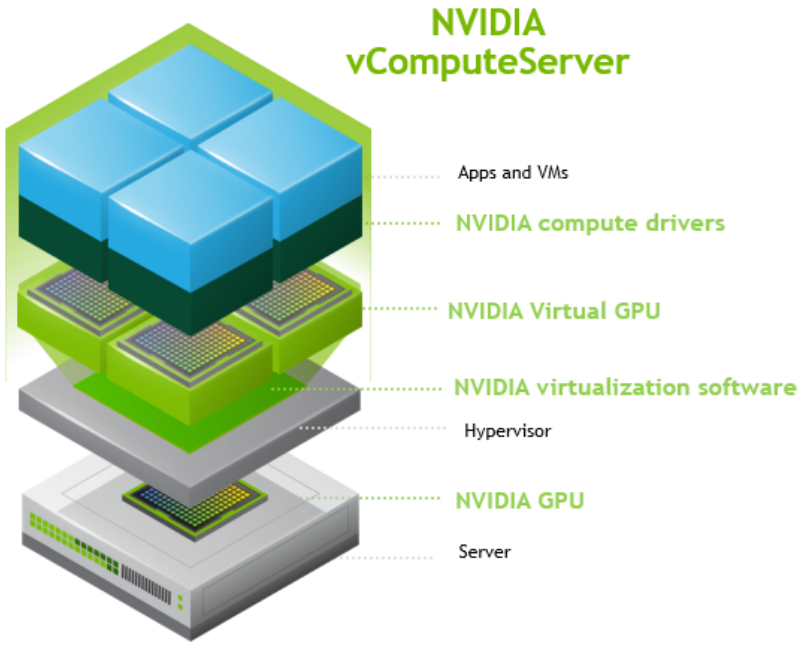


Figure 2: Layered model showing NVIDIA vGPU components

NVIDIA vGPU technology enables GPU virtualization for any workload and is available through software licenses such as the NVIDIA vComputeServer. NVIDIA vComputeServer software, enables virtualize NVIDIA GPUs such as the T4 to power the more than 600 GPU accelerated applications for AI, deep machine learning, and high-performance computing (HPC) as well as the NGC containers. With GPU sharing, multiple VMs can be powered by a single GPU, maximizing utilization and affordability, or a single VM can be powered by multiple virtual GPUs, making even the most compute-intensive workloads possible. With vSphere integration, GPU clusters for compute can be managed by vCenter, maximizing GPU utilization and ensuring security.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001
www.vmware.com

2.2 High Speed Networking with PVRDMA & RoCE

Remote Direct Memory Access (RDMA) provides direct memory access from the memory between hosts bypassing the Operating System and CPU. This can boost network and host performance with reduced latency & CPU load while providing higher bandwidth. RDMA compares favorably to TCP/IP, which adds latency and consume significant CPU and memory resources.

High Performance Computing (HPC) workloads have been traditionally run on bare-metal, non-virtualized clusters. Virtualization was often seen as an additional layer that leads to performance degradation. Performance studies have shown that virtualization has minimal impact on HPC applications.

RDMA over Converged Ethernet (**RoCE**) is a network protocol that allows remote direct memory access (RDMA) over an Ethernet network. There are two RoCE versions, RoCE v1 and RoCE v2. RoCE v1 is an Ethernet link layer protocol and hence allows communication between any two hosts in the same Ethernet broadcast domain. RoCE v2 is an internet layer protocol which means that RoCE v2 packets can be routed. Although the RoCE protocol benefits from the characteristics of a converged Ethernet network, the protocol can also be used on a traditional or non-converged Ethernet network. (Source: Wikipedia)

2.3 Mellanox ConnectX®-5

Intelligent ConnectX-5 EN adapter cards introduce new acceleration engines for maximizing High Performance, Web 2.0, Cloud, Data Analytics and Storage platforms. ConnectX-5 supports dual ports of 100Gbs Ethernet connectivity, sub-700 nanosecond latency, and very high message rate, plus PCIe switch and NVMe over Fabric offloads, providing the highest performance and most flexible solution for the most demanding applications and markets. (Source: Mellanox)

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

2.4 Need for Distributed Machine Learning with Horovod:

There is a lot of time pressure to reduce the time develop a new machine learning model even as the datasets grow in size. There is an increasing need to have distributed machine learning to reduce training time and model development. Horovod is an open source distributed training framework that supports popular machine learning frameworks such as TensorFlow, Keras, PyTorch and MXNet. Horovod distributed deep learning leverages a technique called ring-allreduce, while requiring minimal modification to the user code.

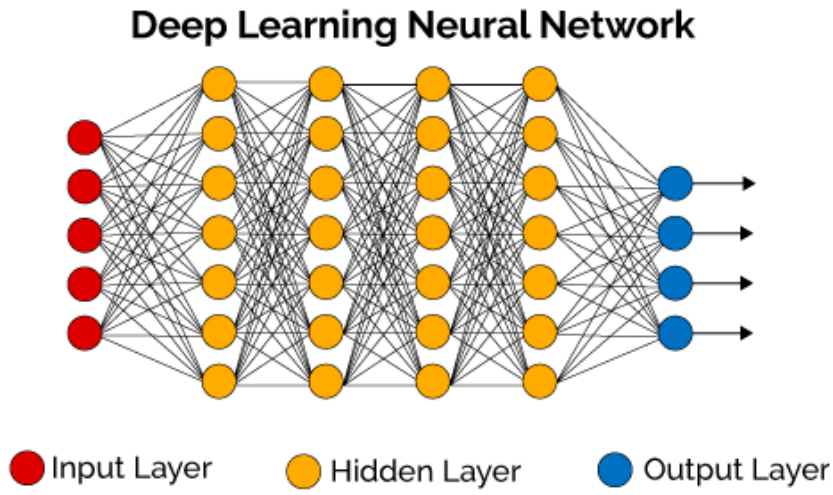


Figure 3: Neural Networks can benefit from the use of GPUs

3 High Performance Virtual Infrastructure for Distributed ML

vSphere supports virtualization of the latest hardware from NVIDIA the T4 GPUs and Mellanox with their Connect X-5 RoCE. There is a potential to combine the benefits of vSphere with the capabilities of these type of high-performance hardware accelerators for Horovod based machine learning and build a compelling solution. VMware, NVIDIA & Mellanox teamed together to develop and create a proof of concept for a High Performance Horovod based machine learning environment.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

3.1 Infrastructure Components of the solution:

The infrastructure components are as shown. The cluster contains four Dell R740 vSphere hosts with one NVIDIA T4 GPU and a 100 Gbps Mellanox Connect X-5 Ethernet card each.

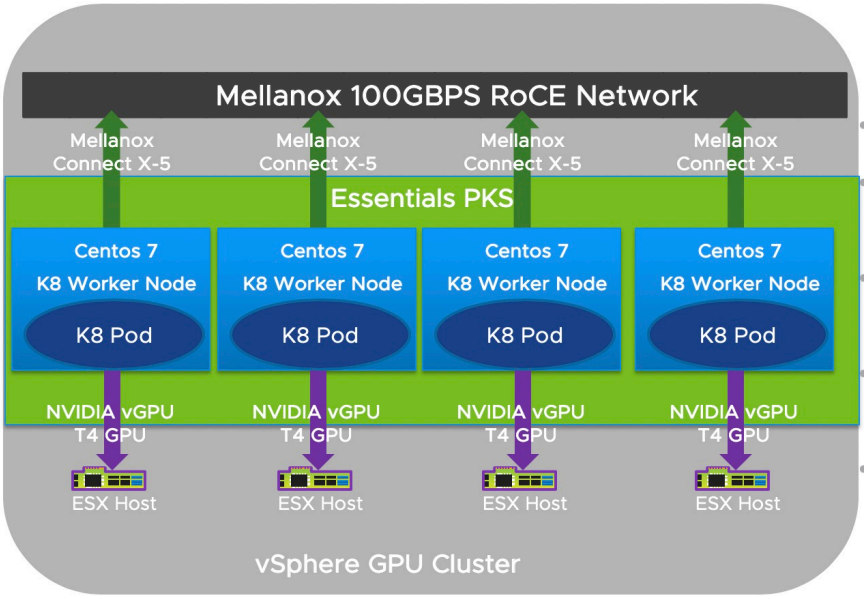


Figure 4: Logical schematic showing solution components

The hosts have been setup with NVIDIA vComputeServer software for GPU virtualization.. PVRDMA based high speed networking also has been enabled.

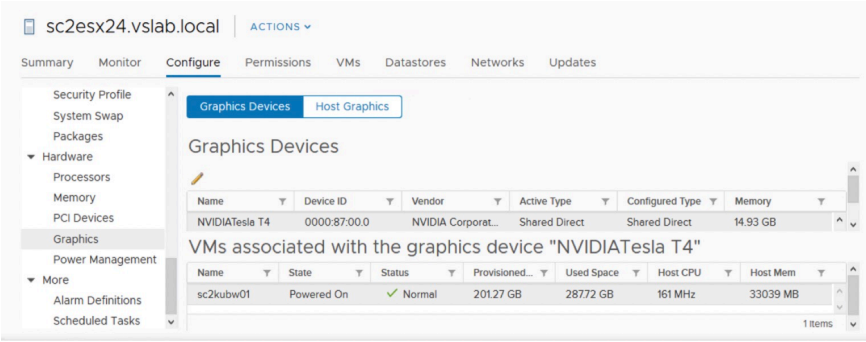


Figure 5: vSphere Node with virtual machines using NVIDIA vGPU

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

CentOS 7.6 based virtual machines were setup and a PKS essentials cluster with one master node and four worker nodes was established. Each of the worker nodes has a vGPU representing an entire NVIDIA T4 processor and are separated across the four physical nodes. PVRDMA based networking has been setup to provide for 100 Gbps capabilities for the PKS essentials virtual machines

Other important features of the solution include

- One Kubernetes pod per VM
- Pods leverage full GPUs allocated to the virtual machine
- Pods packaged with TensorFlow & CUDA Libraries etc.
- NVIDIA vComputeServer for GPU virtualized of the T4 GPU
- High Speed Interconnect with Mellanox X-5 100 Gbps RoCE

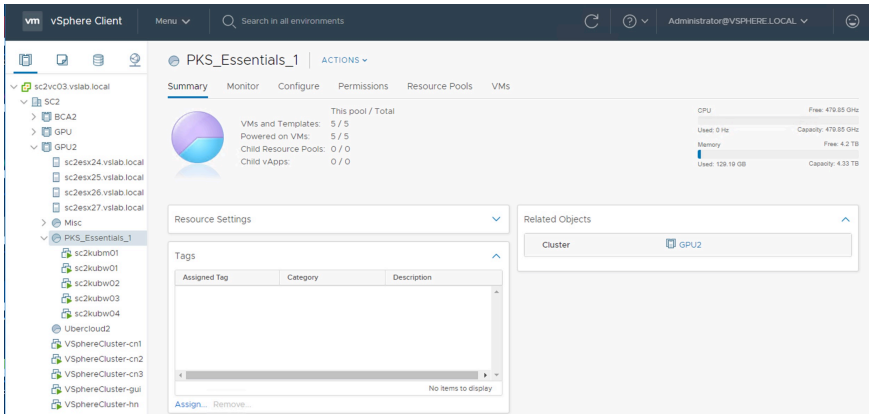


Figure 6: Resource pool with PKS Essential master and worker nodes

The solution used a readily available docker image with the following components to instantiate and run Horovod on the Kubernetes infrastructure.

Name	Version
Docker Image	*NVIDIA/cuda:10.0-devel-ubuntu18.04*
Docker	*18.09.1*
nvidia-docker	*v 2*

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

TF version	*1.13.1*
Benchmark version	*cnn_tf_v1.13_compatible*
Nvidia driver	*430.3*
NCCL	*2.4.7-1+cuda10.0*
CUDA	*10*
Cudnn	*7.6.0.64-1+cuda10.0*
OpenMPI	*4.0.0*
Ethernet	*100 Gb/s*
Docker per VM	*Docker CE 18.09.7*

Table 1: Container & Software components used in the solution

4 Testing

4.1 Running the CNN TensorFlow benchmark with Horovod

Kubernetes, Kubeflow and MPI Operator have to installed and running. The following is a summary of the steps used to deploy and run the solution.

- Create an MPI job using Tensorflow benchmark example. Below is an example of how to run a distributed TensorFlow training job with Horovod framework and RoCE acceleration.

- The job was run with RDMA and with TCPIP. To toggle job from RDMA to TCP the following parameter should be changed

  NCCL_IB_DISABLE=0 to NCCL_IB_DISABLE=1

  HOROVOD_MPI_THREADS_DISABLE=1 to HOROVOD_MPI_THREADS_DISABLE=0

- The MPIJob resource has to be deployed to start the training. Once the MPIJob resource is created, it can be monitored from the status section.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

- The training epoch was initially run for 100 steps and this takes a few minutes on a GPU cluster. In later phases the number of steps was increased to have the process run over many hours. The logs were inspected to monitor the progress of the jobs.

4.2 vMotion Testing:

One of the major features of the vSphere platform is vMotion. The solution leverages specialized hardware and software like the vGPU and PVRDMA that provides specialized compute and networking capabilities to the worker virtual machines.

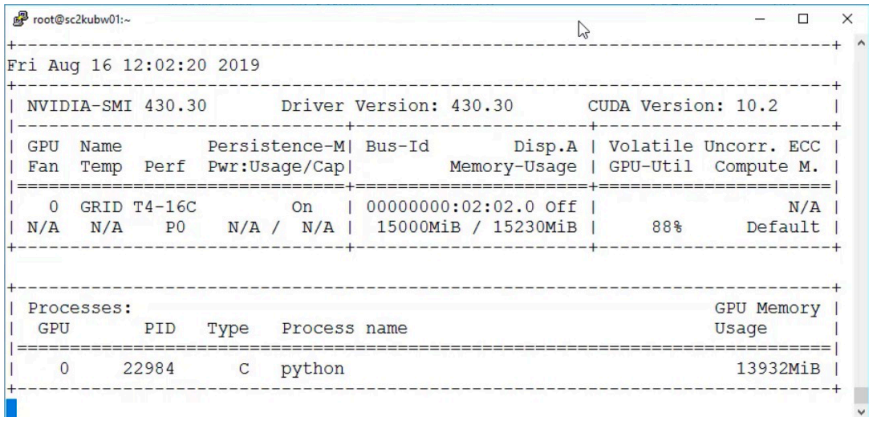
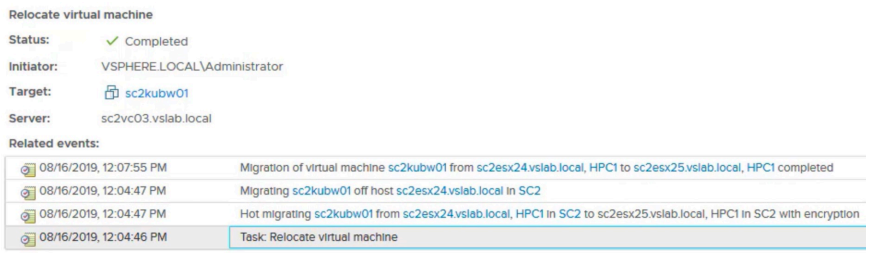


Figure 7: Worker node with the GPU being fully loaded during vMotion

With Horovod running and using 88% of the GPU memory resources, a vMotion was initiated and as shown below successfully completed in three minutes. The vMotion moved the VM to a different host attached it to a different physical GPU. The PVRDMA interface was also migrated successfully between two nodes while it was heavily used for coordination in distributed machine learning.



VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

Figure 8: vMotion statistics after successful completion

5 Results:

The infrastructure was used to run Horovod and the TensorFlow machine learning benchmarks by varying the total number of nodes and vGPUs used. The tests were run separately for TCPIP and PVRDMA and the results were compared.

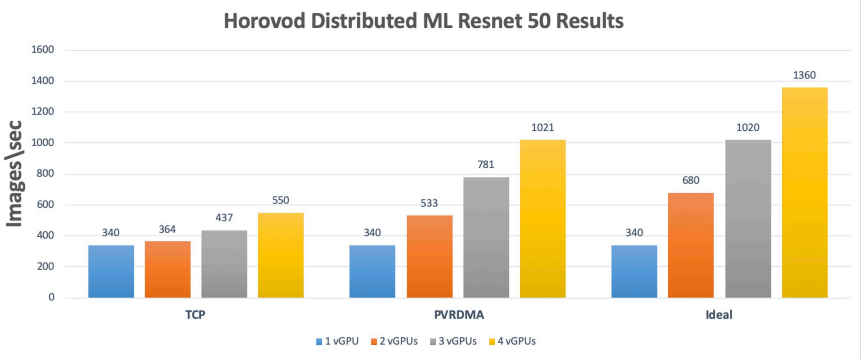


Figure 9: Image processing throughput with Horovod based ML

The results show that when the nodes communicate via PVRDMA there is linear scalability in performance as seen in the images/sec processed. The overhead associated with TCP results in very poor scalability as the number of GPUs increase as seen.

# Nodes	vGPUs	TCP		PVRDMA	
		Images/sec	Speedup	Images/sec	Speedup
1	1	340	1	340	1
2	2	364	1.07	533	1.57
3	3	437	1.29	781	2.30
4	4	550	1.62	1021	3.00

Table 2: Scalability and Speedup with added GPUs

The table represents the same data and shows the speedup. The ideal scenario is calculated by multiplying the images/sec for 1 vGPU by the number of vGPUs. This number can never be achieved even if all the GPUs existed on the same node.

.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

6 Conclusion:

This solution clearly demonstrated the value of distributed machine learning with Horovod on the vSphere platform. VMware essential PKS was successfully deployed with NVIDIA vComputeServer for vGPU and PVRDMA to provide for a high-performance container based machine learning platform. By running machine learning benchmarks across multiple worker pods with independent GPUs, the solution showed excellent scalability while leveraging PVRDMA. vMotion testing validated that even under heavy load virtual machine using vGPUs and PVRDMA can be migrated successfully. This capability provides flexibility improves availability of high-performance machine learning environments

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

Appendix A: YAML File used for the Solution

```
apiVersion: kubeflow.org/v1alpha2
kind: MPIJob
metadata:
  name: tensorflow-benchmarks
spec:
  slotsPerWorker: 1
  cleanPodPolicy: Running
  mpiReplicaSpecs:
    Launcher:
      replicas: 1
      template:
         spec:
           containers:
           - image: mpioperator/tensorflow-benchmarks:latest
             name: tensorflow-benchmarks
             command:
             - mpirun
             - --allow-run-as-root
             - -np
             - "4"
             - -bind-to
             - none
             - -map-by
             - slot
             - -x
             - NCCL_DEBUG=INFO
             - -x
             - NCCL_IB_DISABLE=0
             - -x
             - NCCL_IB_GDR_LEVEL=0
             - -x
             - HOROVOD_MPI_THREADS_DISABLE=1
             - -x
             - LD_LIBRARY_PATH
             - -x
             - PATH
             - -mca
             - pml
             - ob1
             - -mca
             - btl
             - ^openib
             - python
             -
scripts/tf_cnn_benchmarks/tf_cnn_benchmarks.py
             - --model=resnet50
             - --batch_size=32
             - --variable_update=horovod
             - --use_fp16
```

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

```
        - --xla=True
        - --num_batches=10000
Worker:
  replicas: 4
  template:
    spec:
      containers:
      - image: mpioperator/tensorflow-benchmarks:latest
        name: tensorflow-benchmarks
        securityContext:
          privileged: true
        volumeMounts:
        - mountPath: /dev/infiniband
          name: infiniband
        resources:
          limits:
            nvidia.com/gpu: 1
     volumes:
      - name: infiniband
        hostPath:
          path: /dev/infiniband
```

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com