TECHNICAL WHITE PAPER - MAY 2020

VIRTUALIZING HIGH PERFORMANCE COMPUTING (HPC) ENVIRONMENTS

DELL TECHNOLOGIES REFERENCE ARCHITECURE

Table of Contents

1 Audience 3

2 Introduction 3

3 What is High Performance Computing? 3

4 Major Types of HPC Workloads 4

4.1 Parallel Distributed Applications: 4
4.2 Throughput Workloads: 5

5 Virtualizing High Performance Computing 5

5.1 Virtual Machines 5
5.2 Security 6
5.3 Resilience and Redundancy 6
5.4 Performance 6

6 Design 7

6.1 Traditional HPC 7
6.2 Virtual HPC (vHPC) 7
6.2.1 Virtual HPC (vHPC) Components 8
6.3 Management Cluster 9
6.4 Compute Clusters 10
6.4.1 MPI 10
6.4.2 Throughput 11
6.4.3 Hybrid 11
6.5 Hardware Compute Accelerators 11
6.5.1 DirectPath I/O pass-through 13
6.5.2 NVIDIA GRID vGPU mode 13
6.5.3 Accelerators in Cluster Design 13

7 Sample Architectures 15

7.1 Configuration A – Parallel Distributed Applications (MPI) 15
7.1.1 VM sizing 16
7.2 Configuration B – Throughput Workloads 17
7.2.1 VM sizing 18

Conclusion 19

Glossary 19

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

The virtual High-Performance Computing Reference Architecture guide describes the infrastructure and configuration of a High-Performance Computing deployment based on VMware technologies. In addition, it also provides information around the components of virtualization and traditional HPC environments.

1 Audience

This document is intended for Virtualization Architects, IT Infrastructure Administrators and High-Performance Computing (HPC) Systems Administrators who are looking to design, deploy and maintain virtualized HPC workloads within the Enterprise. These organizations generally are either new to High Performance Computing (HPC) or Virtualization environments and looking to combine and deliver virtualized HPC workloads.

2 Introduction

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

This paper is to help identify how virtualization can aid in the delivery of HPC technologies by detailing a system design that both virtualization and HPC admins can interpret, showing how these two technology segments can come together and deliver an elastic, fully managed and self-service, secure virtual HPC environment, ready for the enterprise. We will review the requirements for the physical hardware, the management and operational software and how a virtualized HPC environment can be architected and operated in the datacenter without compromising the performance.

3 What is High Performance Computing?

High-performance computing (HPC) is the use of multiple computers and parallel processing techniques for solving complex computational problems. HPC technology focuses on developing parallel processing algorithms and systems by incorporating both throughput based and parallel computational techniques.

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

High-performance computing is typically used for solving advanced technical problems and performing research activities through computer modeling, simulation and analysis. HPC systems have the ability to deliver sustained performance through the concurrent use of computing resources.

High-performance computing (HPC) evolved to meet increasing demands for processing capabilities. HPC brings together several technologies such as computer architecture, algorithms, programs and electronics, and system software under a single canopy to solve advanced problems effectively and quickly. HPC technology is implemented in multidisciplinary areas including:

- Biosciences
- Geographical data
- Oil and gas industry modeling
- Electronic design automation
- Climate modeling
- Media and entertainment

4 Major Types of HPC Workloads

This design is targeted for the following types of workloads:

4.1 Parallel Distributed Applications:

Parallel distributed applications consist of multiple simultaneously running processes that need to communicate with each other, often with extremely high frequency, making their performance sensitive to interconnect latency and bandwidth. Parallel distributed applications are also called Message Passing Interface (MPI) applications because they are enabled by MPI parallel programming on distributed-memory systems. MPI is a standard for multiprocessor programming of HPC codes. Typical parallel distributed applications include weather forecasting, molecular modeling, and design of jet engine, spaceship, airplane automobile, where each program is able to run in parallel on a distributed memory system and communicate through MPI processes.

Numerous scientific applications that run in a distributed way use the MPI library and take advantage of communication primitives such as point-to-point operations (e.g., MPI_Send and MPI_Receive) and collective operations (e.g., MPI_Bcast, MPI_Reduce). MPI libraries are designed to

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

use the best available pathways between communicating endpoints, including shared memory, Ethernet, and – for high performance – RDMA-capable interconnects. Data-intensive workloads require parallel file system with high performance I/O. Data-intensive workloads with large amount of data require effective storage technologies. Parallel file system is designed to achieve high performance for handling of large datasets. It can distribute file data across multiple hosts and provide concurrent access for applications with parallel I/O implementation with MPI.

4.2 Throughput Workloads:

Throughput workloads require a large number of individual jobs to be run in order to complete a task with each job running independently and no communication between the jobs. Typical throughput workloads include Monte Carlo simulations in financial risk analysis, digital movie rendering, electronic design automation and genomics analysis, where each program run in a long-time scale or have hundreds or thousands even millions of executions with varying inputs.

5 Virtualizing High Performance Computing

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 often has minimal impact on HPC application performance.

5.1 Virtual Machines

- Adds the ability to run multiple server configurations, operating systems and configurations of HPC all on the same physical hardware at the same time allowing for greater flexibility.
- Allows infrastructure administrators and HPC users greater control by being able to dynamically resize, pause, take snapshots, back up, replicate to other virtual environments or simply wipe and redeploy based on role-based permissions.
- Greater control of resource prioritization and balancing with resource scheduling across pools of physical servers.
- Remove the physical location as a limitation to expanding resources. For example, multiple sites including cloud.

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

5.2 Security

Security rules and policies can be defined and applied based on environment, workflow, virtual machine, physical server or operator.

- All actions controlled via user permissions and logged for audit reporting. For example, Root access privilege is only granted as needed and based on the specified virtual machines preventing compromise of other HPC workflows.
- Truly isolated workflows where sensitive data is not shared with other HPC environments, workflows or other users while running on the same pool of resources.
- If running in remote or isolated locations, optional virtualization solutions can be used to provide secure remote access to consoles for monitoring, operating and deploying HPC workflows as if the users was physically present.

5.3 Resilience and Redundancy

Providing fault isolation, dynamic recovery and other capabilities not available in traditional HPC environments.

- Deliver maintenance without impacting operational HPC workflows or serviceability.
- When a physical server fails, the HPC virtual machines can be restarted automatically on other physical servers within the cluster to maintain operations.
- When resources are stretched with a given host, being able to automatically migrate to another physical host without interruption to service.

5.4 Performance

Many studies have found that virtualization technologies can often match and sometimes even better HPC bare metal performance based on throughput workloads and parallel distributed applications. More details are available in the HPC Resources page.

Virtualization continues to evolve with increasing performance characteristics and one such example is the ability to map physical PCIe resources such as high-speed interconnects and General-Purpose Graphics Processing Units (GPGPU) directly to virtual machines providing similar performance to that of bare metal environments.

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

6 Design

Designing the characteristics of the HPC system for a virtual environment requires an understanding of how the architecture will be compared to that of traditional bare metal HPC environments as well as defining the workload parameters required for a successfully deployed virtual HPC.

6.1 Traditional HPC

With traditional HPC environments the entire solution is made up of multiple physical servers. Each physical server is dedicated to a specific role that is managed independently. Compute Nodes are controlled by additional servers running cluster management software and are connected via high speed interconnects.

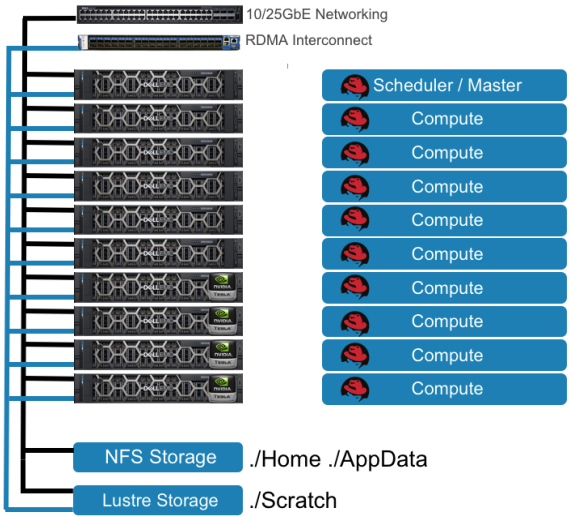

Figure 1: Traditional Bare Metal HPC Cluster

6.2 Virtual HPC (vHPC)

When virtualizing a traditional HPC environment design, a significant benefit for someone familiar with HPC is the ability to maintain and recreate the same familiar environment simply by using VMware software to create resource clusters (grouping of servers) and virtual machines (software-based servers). This simplifies the time to get up and running and allow the existing procedures to be maintained. Many of the HPC components can be virtualized as shown below.

Those more familiar with virtualization will view virtualized HPC as a dedicated application environment, built using their existing virtualization

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

skills. This would be similar to how other applications are deployed but needs a deeper awareness of the components and workflows required for the HPC solution.

6.2.1 Virtual HPC (vHPC) Components

Where traditional High-Performance Compute environments are built with physical servers and physical high speed devices, a virtualized High-Performance Compute environment utilizes VMware solutions and technologies.

- **VMware vSphere (ESXi)** is an Enterprise class, type 1 hypervisor that the underlying physical assets are still present, however much of the complexity is reduced by standardization of physical resources which become encapsulated by the hypervisor to where compute, storage and networking are presented spanning all physical resources and sub systems within a cluster. This highly optimized layer forms the foundation to the virtual HPC infrastructure and allows for software to define HPC clusters and virtual machines where the HPC workloads will be scheduled and executed.
- **VMware vCenter Server Appliance (VCSA)** provides the centralized management of all virtualized infrastructure and delivers a single management interface to interact, manage and monitor the virtualization configuration, settings and services.
- **VMware NSX** provides software defined networking (SDN) to where network functions like switching, routing, firewalling and load balancing are attached to your virtual High-Performance Compute applications creating optimized networking and security policies distributed throughout the environment.

Optional Components

- **VMware Tanzu Kubernetes Grid (TKG)** provides capabilities to deploy multiple Kubernetes clusters on top of vSphere. These clusters are all upstream-compliant; they run containers on Linux hosts. This can be integrated with **Singularity** which is a popular container runtime for HPC applications.
- **VMware vRealize Automation** accelerates the deployment and management of applications and compute services, embowering IT to quickly standardize the deployment of automated virtual High

Performance Compute platforms that can be requested or delivered on demand as needed.

- **VMware vRealize Operations** delivers tools to optimize the operating physical, virtual and application environment with intelligent alerting, policy-based automation and unified management.
- **VMware Horizon** delivers the ability to connect and operate your virtual High-Performance Compute environment securely from a remote location without any data or communications leaving the virtualized datacenter.

The use of these VMware technologies and solutions are designed to be modular so that only the functions and features desired can be utilized. Further information can be found at http://www.vmware.com

6.3 Management Cluster

Management cluster runs the virtual machines that manage the virtual High Performance Compute environment. These virtual machines host vCenter Server, vSphere Update Manager, NSX Manager, NSX Controllers, vRealize Operations Manager, vRealize Automation as well as the administrative virtual High Performance Compute services like the master node and workload schedulers. All management, monitoring, and infrastructure services are provisioned to a vSphere cluster which provides high availability for these critical services. Permissions on the management cluster limit access only to administrators. This limitation protects the virtual machines that are running the management, monitoring, and infrastructure services from unauthorized access.

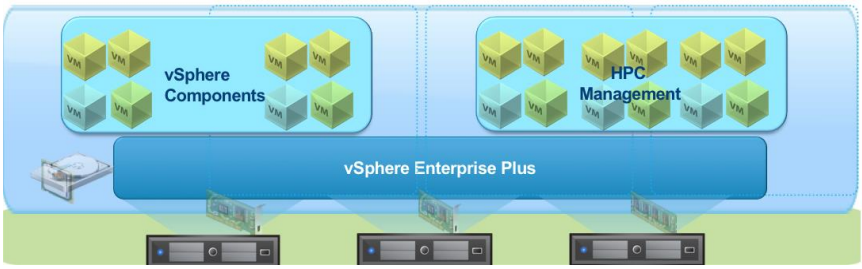

Figure 2: HPC Management Cluster

The management components for vSphere and HPC can be combined and deployed in this cluster. If no such cluster already exists, a new management cluster with a minimum cluster of 3 nodes is recommended.

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

The new cluster should be sized based on projected management workload and some headroom for growth. Capacity Analysis should be performed on existing management cluster and adjusted to ensure that there is enough capacity to add HPC management components.

Due to the critical nature of the workloads with many single points of failure, it is recommended that this cluster is licensed with vSphere Enterprise Plus for high availability and other advanced features. vSphere Enterprise Plus provides vSphere HA, DRS and other advanced capabilities that help reduce downtime for these critical workloads.

6.4 Compute Clusters

The compute cluster runs the actual HPC workload components. The vSphere Scale Out licensing can be leveraged for these compute clusters.

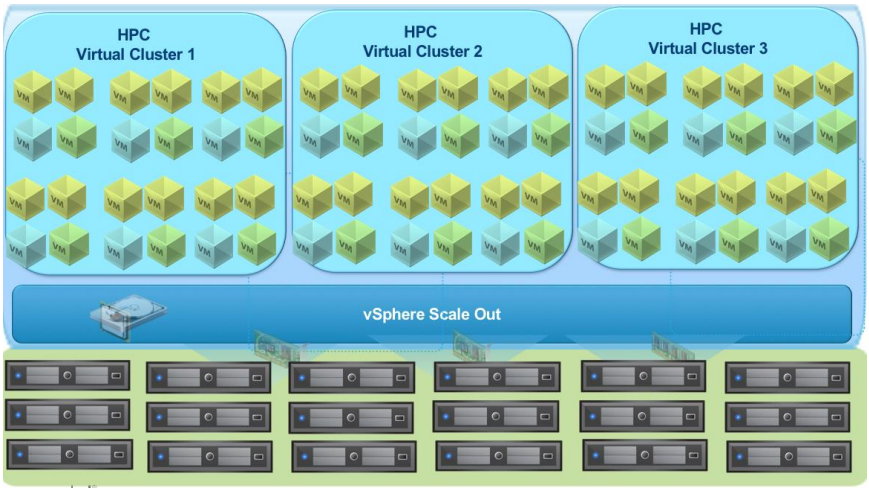

Figure 3: HPC compute cluster with vSphere Scale Out

6.4.1 MPI

MPI environments are dedicated as they have unique requirements with the need for low latency communications between nodes. The nodes are connected via high speed interconnect and are not very amenable to sharing with other workloads. MPI applications leverage the entire high-performance interconnects using Pass-Through mode in virtualized environments. Storage for MPI nodes is usually a parallel file system like Lustre also accessed via the high-speed interconnect.

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

6.4.2 Throughput

Throughput workloads are horizontally scalable with little dependency between individual tasks. Job schedulers divvy up the work across nodes and coordinate the activity. NFS is the typical shared storage across the nodes which is accessed via TCP/IP networks.

6.4.3 Hybrid

Different types of HPC workloads can co-exist and potentially leverage some of the capabilities of each other. Both MPI and Throughput can require the use of accelerators such as GPUs that can be shared by these workloads. Another type of Hybrid is where the GPUs are used for desktop graphics during the day and for HPC with Deep Learning during nights and weekends in a concept called cycle harvesting.

6.5 Hardware Compute Accelerators

A new area in which HPC techniques are becoming critically important is Machine learning. Machine Learning is increasingly applied in many areas like health science, finance, and intelligent systems, among others.

In recent years, the emergence of deep learning and the enhancement of accelerators like GPUs has brought the tremendous adoption of machine learning applications in a broader and deeper aspect of our lives. Some application areas include facial recognition in images, medical diagnosis in MRIs, robotics, automobile safety, and text and speech recognition.

The use of accelerators for handling massive floating-point computation that can be parallelized on many cores is a strong trend in HPC virtualization and cloud area, particularly for accelerating deep learning workloads.

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

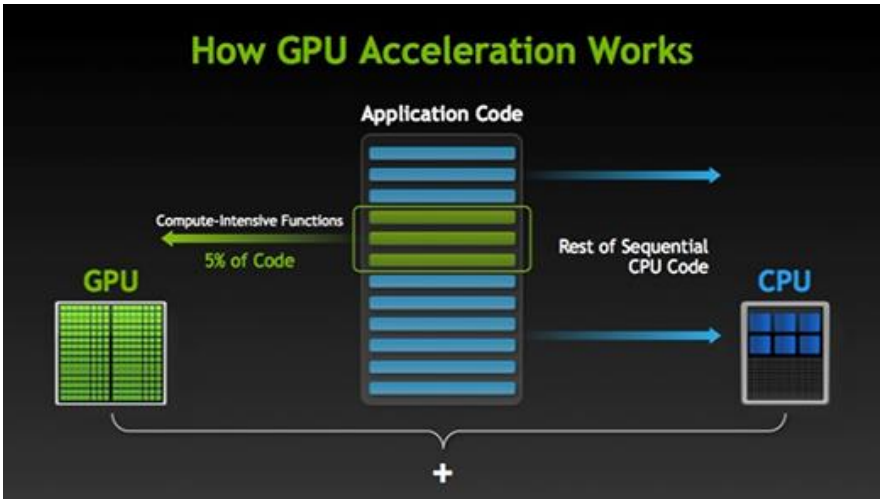

Figure 4: How GPU Acceleration Works (Source: NVIDIA)

Typical hardware compute accelerators include GPGPU, Intel Xeon Phi, field-programmable gate array (FPGA). NVIDIA, AMD MxGPU, Intel and Xilinx provide compute accelerators.

Unlike NVIDIA vGPU, which time slices GPU across multiple VMs, AMD MxGPU enables VMs sharing GPU with physical slices and gives users 100% time of the assigned physical GPU slice.

In general, compute accelerators can be configured in DirectPath I/O (passthrough) mode on vSphere, which allows a guest OS to directly access the device, essentially bypassing the hypervisor. Because of the shortened access path, performance of applications accessing accelerators in this way can be very close to that of bare-metal systems.

For NVIDIA GRID GPU cards, with vGPU technology in vSphere, vSphere supports the use of general-purpose GPU through direct access to the underlying hardware with no overhead, while also providing the ability to share GPUs across multiple virtual machines. vSphere 6.7 also provides the capability to do vMotion and suspend, resume those vGPU enabled VMs prioir to a vMotion.

The use of GPGPU for accelerating computing performance for deep learning workloads is a strong trend within HPC. GPUs reduce the time it takes for a machine learning or deep learning algorithm. Performance studies have shown minimal difference in performance between bare metal and virtualized workloads with pass through GPU.

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

You can configure the GPUs in two modes:

6.5.1 DirectPath I/O pass-through

In this mode, the host can be configured to have multiple GPUs in a DirectPath I/O pass-through mode. When using this mode only one VM can use a GPU device at a time. All accelerators work in this mode as long as the drivers exist at the virtual machine level.

6.5.2 NVIDIA GRID vGPU mode

This is specific to NVIDIA GRID GPUs. The NVIDIA GRID vGPU enables multiple virtual machines to share a single physical GPU. For NVIDIA grid drivers are available at the ESX level in the form of a VIB and at the virtual machine level (Windows or Linux). By this mode, vGPUs are created using profiles which specify how much GPU memory has been assigned to a vGPU. Currently, all vGPUs on the same physical GPU must use the same profile. Note that while GPU memory is partitioned to create vGPUs, each vGPU is given time-sliced access to all of a GPU's compute resources. The time slicing is managed using one of several available scheduling algorithms, each of which is appropriate for different shared GPU use-cases.

6.5.3 Accelerators in Cluster Design

Due to the specialized nature of the Accelerator hardware and their servers, dedicated clusters are recommended for these workloads. In general, the accelerator components such as GPU cards should be equally distributed across all nodes in the cluster and shared with the virtual machines.

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

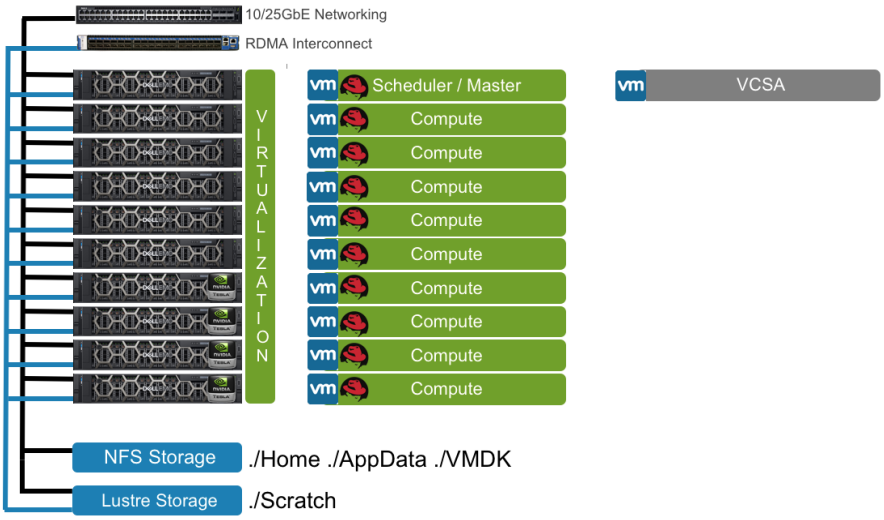

Figure 5:Virtualized HPC Cluster with single virtual machine per node

One key benefit is that while the utilization of virtualization technologies will incorporate additional virtualization management and operational solutions, the HPC solution deployed virtually requires no additional hardware to that used in an existing physical cluster. Physical disks in the VMware domain are represented by VMDKs and used for operating system boot disks.

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

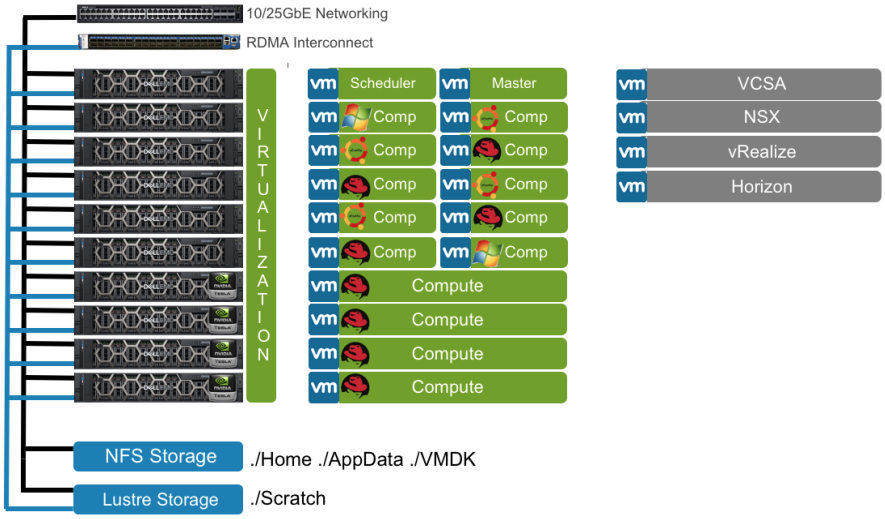

Figure 6: Hybrid Multi-Purpose HPC Cluster

When virtualizing, many of these servers become hosts for virtual machines and so a set of virtual machine templates must be architected.

Virtual machine templates are a way of defining different types of virtual machine resources and include such hardware resources as CPU, Memory, Storage capacity for a given use case and software components such as OS. Templates are typically based on a sizing order offering small, medium and large-scale resources.

7 Sample Architectures

With all of the HPC components understood and how they match to virtualization, the following is a sample of several architectures to show how HPC environments can be deployed with virtualization technologies based on scenario.

7.1 Configuration A – Parallel Distributed Applications (MPI)

Parallel Distributed Applications based HPC environments are designed to allow complex research or design simulations, such as handling massive amounts of simulation, machine-generated data, structural analysis and computational fluid dynamics; all requiring complex algorithms for modeling, rendering and analysis. To best achieve this, a virtual HPC environment will be based on multiple compute virtual machines offering a denser environment with fast interconnects to keep up with demanding research or manufacturing workloads.

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

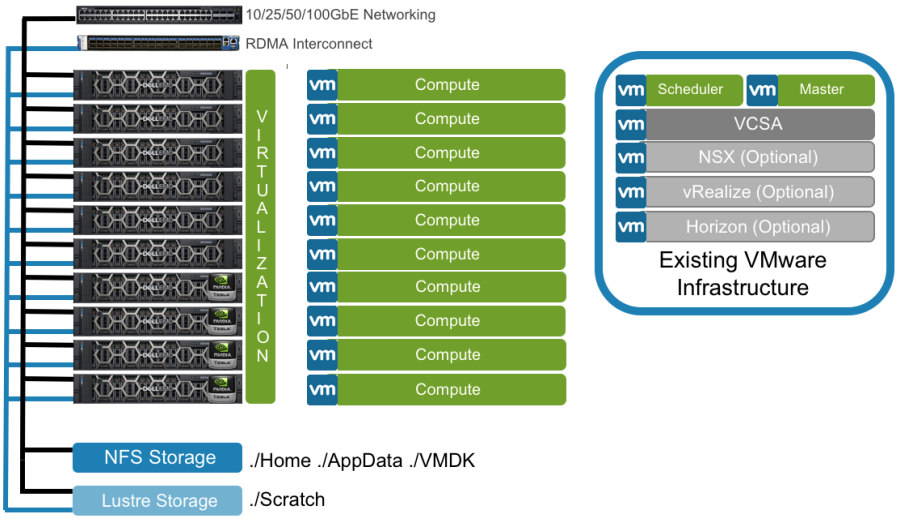

Figure 7: Sample Architecture for MPI

7.1.1 VM sizing

- Single Compute Virtual machine per node

Dell	Components
Compute Nodes	Intel: R640, R840, C6420 AMD: R6525, R7525, C6525
Management Nodes	Intel: R640
Accelerator Nodes	C4140, R740

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

Software	VMware VCF with vSphere, NSX, Bitfusion VMware TKG with Singularity (Optional) VMware vROPS & vRA (Optional) vHPC Toolkit NVIDIA CUDA Automation with HashiCorp Terraform vSphere Provider
InfiniBand/ROCE	Mellanox Connect X-3, X-4 & X-5
Networking	Dell EMC PowerSwitch S Series 25/40/100GbE Switches
Storage	NFS: Dell EMC HPC NFS Storage Lustre: Dell EMC HPC Lustre Storage Isilon F800 Scale-out NAS

Table 1: Dell Technologies HPC MPI Workload components

7.2 Configuration B – Throughput Workloads

This type of HPC enables high-throughput and fast turnaround of workflows in diverse fields. For a virtual HPC for Life Sciences the following is recommended. Throughput workloads can also be potentially deployed in VMware Cloud on AWS with vSphere virtual machines used for compute and Amazon EFS used for shared NFS like storage.

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

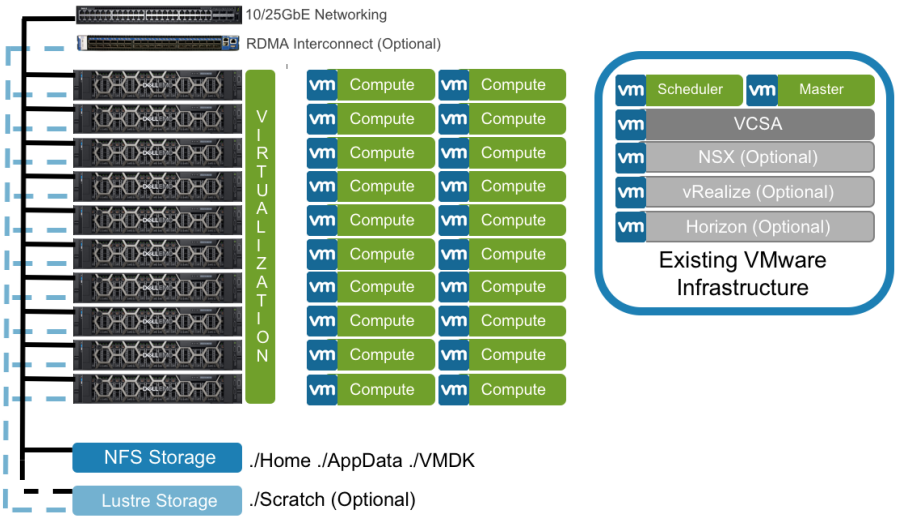

Figure 8: Sample Architecture for Throughput applications

7.2.1 VM sizing

- Multiple Compute Virtual machines per host

Dell	Components
Compute Nodes	Intel: R640, R840, C6420 AMD: R6525, R7525, C6525
Accelerator Nodes	C4140, R740
Software	VMware VCF with vSphere, NSX, vSphere Bitfusion VMware TKG with Singularity (Optional) VMware vROPS & vRA (Optional)

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

	vHPC Toolkit NVIDIA CUDA Automation with HashiCorp Terraform vSphere Provider
Networking	Dell EMC PowerSwitch S Series 10/25GbE Switches
Storage	NFS: Dell EMC HPC NFS Storage Lustre: Dell EMC HPC Lustre Storage Isilon F800 Scale-out NAS

Table 2: Dell Technologies HPC Throughput Workload components

Conclusion

Virtualization offers tremendous benefits for HPC Solutions. With a simple understanding of HPC components and technologies, an enterprise organization can follow HPC best practices and guidelines to adopt a VMware virtualized infrastructure. The scale out licensing for vSphere and HPC helps compute nodes leverage virtualization at a low cost. Virtualization provides the building blocks that can be leveraged to provide for individualized or hybrid clusters for HPC applications.

With the addition of VMware optional VMware Management solutions, HPC can benefit from increased security, self-service, multi workflow environments and accessible from remote console. With the HPC community eying cloud computing, virtualizing HPC is a must to future proof your infrastructure and tune for the future.

Glossary

HPC -- High Performance Computing

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

OS -- Operating System

MPI -- Message Passing Interface

FDR -- Fourteen Data Rate

EDR -- Enhanced Data Rate

HDR -- High Data Rate

RDMA -- Remote Direct Memory Access

RoCE -- RDMA cover Converged Ethernet

iWARP -- Internet Wide-area RDMA Protocol

PVRDMA -- Paravirtual RDMA

SR-IOV -- Single Root I/O Virtualization

GPGPU -- General Purpose Computing On Graphics Processing Unit

VMM -- Virtual Machine Monitor

BARs -- Base Area Registers

MMIO -- Memory-Mapped I/O

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