



NetApp Verified Architecture

VMware End-User Computing with NetApp HCI and NVIDIA GPUs

NetApp Verified Architecture Design

Suresh Thoppay, NetApp
January 2019 | NVA-1129-DESIGN | Version 1.0

Abstract

VMware End-User Computing with NetApp® HCI is a prevalidated, best-practice data center architecture for deploying graphics-intensive workloads at an enterprise scale. This document describes the architectural design and best practices for deploying the solution at production scale in a reliable and risk-free manner.

TABLE OF CONTENTS

- 1 Executive Summary..... 5**
- 2 NetApp HCI and VMware End-User Computing Architecture 5**
 - 2.1 NetApp HCI.....6
 - 2.2 VMware End User Computing.....7
 - 2.3 NetApp HCI and VMware End-User Computing Design Principles.....9
- 3 Solution Overview 9**
 - 3.1 Target Audience.....9
 - 3.2 NetApp HCI Use Cases10
 - 3.3 Solution Use Case Summary10
- 4 Technology Overview..... 11**
 - 4.1 NetApp HCI.....11
 - 4.2 VMware vSphere12
 - 4.3 Virtual Dedicated Graphics Acceleration (vDGA).....13
 - 4.4 Virtual Shared Graphics Acceleration14
 - 4.5 Virtual Shared Pass-Through Graphics Acceleration with the NVIDIA vGPU16
 - 4.6 VMware Horizon 7 Enterprise20
 - 4.7 VMware Identity Manager20
 - 4.8 VMware Horizon21
 - 4.9 VMware ThinApp25
 - 4.10 VMware App Volumes25
 - 4.11 User Environment Manager26
 - 4.12 Just-In-Time Management Platform.....27
 - 4.13 VMware Mirage.....28
 - 4.14 VMware NSX for vSphere.....28
 - 4.15 VMware vRealize Operations Manager29
 - 4.16 VMware vRealize Orchestrator Plug-In for Horizon31
 - 4.17 VMware vRealize Log Insight.....32
 - 4.18 VMware Workstation.....34
 - 4.19 VMware Fusion34
- 5 Solution Design 34**
- 6 Technology Components 36**
 - 6.1 Hardware Components36
 - 6.2 Software Components36

7 Solution Verification	37
8 Conclusion	37
Where to Find Additional Information	38
NetApp	38
NVIDIA	38
VMware	38
Version History	39

LIST OF TABLES

Table 1) vSphere maximums	13
Table 2: Horizon sizing	24
Table 3) Desktop VM configuration	35
Table 4) Datastore estimates	35
Table 5) Hardware components	36
Table 6) Software requirements	36

LIST OF FIGURES

Figure 1) VMware Horizon Cloud architecture	6
Figure 2) VMware Workspace One and NetApp HCI	8
Figure 3) VMware Horizon JMP technology	8
Figure 4) Solution architecture	11
Figure 5) Common Element OS	Error! Bookmark not defined.
Figure 6) SolidFire-to ONTAP disaster recovery solution	12
Figure 7) Enable DirectPath	14
Figure 8) Pass-through PCI device for VMs	14
Figure 9) Virtual Shared Graphics Acceleration	15
Figure 10) gpvm CLI	15
Figure 11) NVIDIA vGPU architecture	16
Figure 12) Host graphics settings	17
Figure 13) vGPU profiles for Tesla M10 (from NVIDIA documentation)	17
Figure 14) Homogenous virtual GPU profile combination	18
Figure 15) nvidia-smi vgpu cli	18
Figure 16) CUDA-Z screenshot	19
Figure 17) Web launcher portal (from VMware documentation)	21
Figure 18) Provisioning Instant Clones	22
Figure 19) Blast protocol features	Error! Bookmark not defined.
Figure 20) Unified Access Gateway logical diagram	24
Figure 21) VMware App Volumes	25

Figure 22) AppStacks and Writable Volumes	26
Figure 23) UEM file shares.....	27
Figure 24) JMP technologies.....	28
Figure 25) NSX microsegmentation.	29
Figure 26) Five key features of VMware vRealize Operations Manager.	29
Figure 27) NetApp HCI vROPS integration.	30
Figure 28) NVIDIA vROPS integration.	31
Figure 29) GPU metrics.....	31
Figure 30) Horizon vRealize Orchestrator Plug-In architecture.	32
Figure 31) VMware vRealize Log Insight architecture.	33
Figure 32) Resource block scalability.....	34

1 Executive Summary

The right information at the right time becomes more crucial than ever for business success. Smartphones and tablets can retrieve information easily from the internet, but securely retrieving data from workspaces is always challenging. VMware Horizon virtual desktops and hosted apps along with VMware Workspace ONE have helped address this issue by providing secure access to virtual desktop environments or to specific applications connected to end devices.

Modern operating systems and other applications such as Microsoft Office have now become more graphics intensive, which adds CPU cycles to virtual desktops and applications and reduces user density on host servers. Having a graphics acceleration card on the server offloads graphics processing and rendering work from the CPU, creates a better user experience, and increases user density.

NetApp® HCI provides scale-out all-flash storage and guaranteed quality of service (QoS), which makes it easier to run virtual desktops and applications along with other workloads. With the NetApp HCI vCenter Plug-In, administrators can provision datastores, expand clusters, and perform other vital functions directly from vCenter. NetApp HCI makes it the ideal platform for consolidating your workloads.

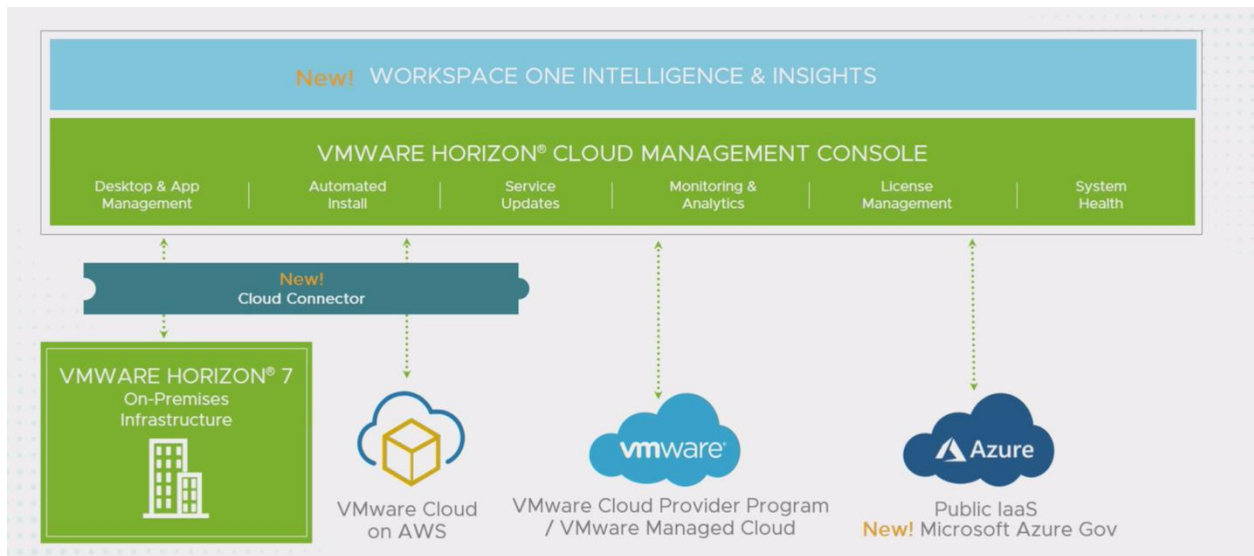
2 NetApp HCI and VMware End-User Computing Architecture

This combination of technologies from NetApp and VMware allows customers to experience the benefits of the End-User Computing ecosystem. This NetApp Verified Architecture details the design decisions made to deploy VMware End-User Computing on NetApp HCI.

NetApp HCI is a hybrid cloud infrastructure solution capable of transforming and empowering organizations to move faster, promote operational efficiencies, and reduce costs. NetApp HCI is the foundation of the End-User Computing strategy, which can run multiple applications with the predictable performance that enterprises and customers demand. NetApp HCI enables the independent scaling of compute and storage resources so that systems are rightsized. With VMware Horizon, NetApp HCI can provision desktops for users in minutes along with the required applications, eliminating the complex management of traditional architectures. Integration into the NetApp Data Fabric means that you can easily integrate your infrastructure with the cloud by using the required data services.

NetApp HCI frees you from the limitations of current infrastructure solutions that are complex, that cannot consolidate all workloads, that force customers to scale in ways that strand resources, and that throttle the performance required by next-generation applications. With VMware Horizon and NetApp HCI, customers can quickly deploy virtual desktops and applications on an infrastructure that can be deployed quickly and scaled as needs change. Figure 1 illustrates how VMware Horizon running on-premises with NetApp HCI can integrate with cloud providers like AWS, Azure, and so on using the VMware Horizon Cloud service.

Figure 1) Horizon Cloud architecture.



2.1 NetApp HCI

NetApp HCI offers various benefits to consumers seeking a hybrid cloud infrastructure by combining industry best practices and the VMware vSphere Hypervisor. NetApp HCI delivers features and capabilities that first-generation HCI vendors could not. NetApp HCI is predictable, flexible, and scalable; provides simple administration and deployment; and is integrated into the NetApp Data Fabric.

Predictable

One of the biggest challenges for anyone managing infrastructure is delivering predictable performance, especially in the face of proliferating applications and workloads. Dedicated platforms and massive overprovisioning are not economically viable. However, when multiple applications share infrastructure, one application might interfere with the performance of another. NetApp HCI alleviates this concern with QoS limits available natively with NetApp Element® software. Element enables the granular control of each application and volume, eliminates noisy neighbors, and satisfies all performance SLAs. All applications can be deployed on a shared platform, predictably and with confidence. The multitenancy capabilities of NetApp HCI can help eliminate more than 90% of traditional performance-related problems.

Flexible and Scalable

Hyper converged infrastructures have required fixed resource ratios, limiting deployments to configurations between four and eight nodes. NetApp HCI, however, scales compute and storage resources independently. Independent scaling avoids costly and inefficient overprovisioning, eliminates the 10% to 30% "HCI tax" from controller VM overhead, and simplifies capacity and performance planning.

With NetApp HCI, licensing costs are reduced. NetApp HCI is available in mixtures of small, medium, and large storage and compute configurations. The architectural design choices enable customers to confidently scale on their terms, making HCI viable for core data center applications and platforms.

No data center scales linearly, because business needs change constantly, and each application has different infrastructure requirements. NetApp HCI enables independent scaling of compute and storage resources, allowing on-demand scaling, avoiding costly and inefficient overprovisioning, and simplifying capacity and performance planning.

NetApp HCI is architected in building blocks at either the chassis or the node level. Each chassis can hold four nodes that are made up of storage nodes, compute nodes, or both. A minimum configuration is two chassis with six nodes, consisting of four storage nodes and two compute nodes. Two more blank spots can be used for expansion. Compute and storage nodes can be mixed if best practices are followed. Resources can be scaled nondisruptively through a simple GUI-driven process.

Simple

An imperative within the IT community is to automate all routine tasks, eliminating the risk of user error while freeing up resources to focus on more interesting, higher-value projects. NetApp HCI allows IT departments to become more agile and responsive by simplifying deployment and ongoing management.

The new NetApp Deployment Engine (NDE) eliminates most manual steps needed to deploy infrastructure, such as assigning names, network settings, and IP addresses, and provisioning ESXi hosts and VMware datastores. You can expect the infrastructure to be functional in less than 30 minutes.

The VMware vCenter Plug-In simplifies management in an intuitive way. Additionally, NetApp HCI uses a robust suite of APIs to promote integration into higher-level management, orchestration, backup, and disaster recovery tools.

NetApp Data Fabric

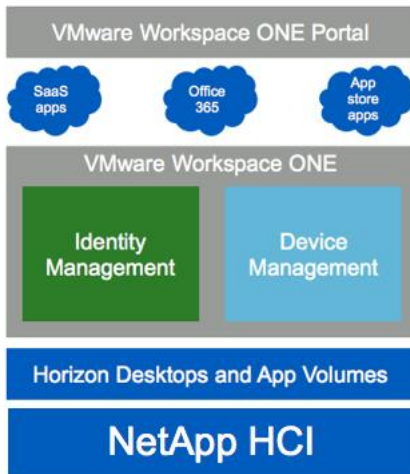
Traditional HCI platforms often involved the introduction of a silo of resources into an existing data center. Such platforms have little in common with other infrastructure-consumption choices that consumers might have made already or would like to make in the future. This approach is not efficient in the long term.

In contrast, NetApp HCI integrates into the NetApp Data Fabric for enhanced data portability, visibility, and protection of workloads, whether they reside on premises, in near-cloud storage, or in a public cloud. The NetApp Data Fabric removes lock-in and provides you with new choices. It allows the full potential of your data to be unleashed across cloud environments.

2.2 VMware End-User Computing

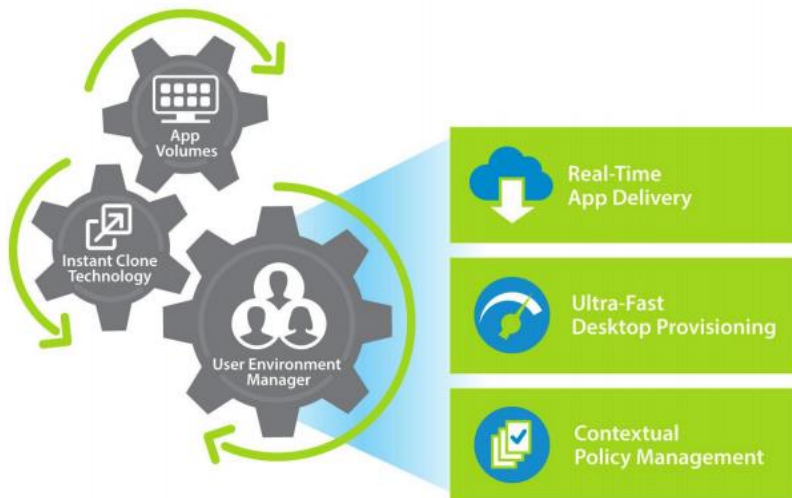
VMware Horizon enables IT organizations to provide virtual desktops and hosted applications to their users. As users start using smartphones, tablets, and other portable devices to access information, managing such devices becomes a challenge. VMware Workspace ONE provides unified management for all endpoints. It also provides single sign-on to hosted apps, desktops, intranet applications, and SaaS-based applications, as is depicted in Figure 2.

Figure 2) VMware Workspace ONE and NetApp HCI.



The VMware Horizon Just-In-Time Management Platform (JMP) allows customers to perform ultrafast desktop or Remote Desktop Session Host (RDSH) provisioning with VMware Instant Clones, real-time application delivery with VMware App Volumes, and contextual policy management with VMware User Environment Manager.

Figure 3) Horizon JMP technology.



Horizon Instant Clones, App Volumes, and User Environment Manager are all part of Horizon 7 Enterprise Edition, which also includes other components such as the following:

- vRealize Operations Manager for performance monitoring of desktop and application pools
- vRealize Log Insight for centralized log collection and analysis
- VMware NSX for vSphere for software-defined networking; securing desktops with microsegmentation; and edge services such as load balancers and the Dynamic Host Configuration Protocol (DHCP).

2.3 NetApp HCI and VMware End-User Computing Design Principles

NetApp HCI and VMware End-User Computing products provide an integrated system that offers all the benefits of VMware Horizon and the scalability and granularity of NetApp HCI. The underlying NetApp HCI system allows you to expand or resize a data center according to CPU, memory, storage capacity, and storage IOPS requirements.

NetApp HCI also lets you add and repurpose compute and storage nodes of various capacities to expand or contract any of the compute or storage parameters according to a data center's needs. This scaling is managed through vCenter and the NDE.

The NDE manages hardware configuration and deployment of the NetApp HCI environment, so compute and NetApp HCI storage nodes can be added or deleted easily in any configuration. To add compute nodes to a VMware cloud configuration, you simply add them to the vCenter data center and compute clusters. You add storage nodes to the NetApp HCI cluster and provision the datastores from the vCenter Plug-In. To add capacity and throughput, add more datastores to the desktop pools.

3 Solution Overview

Media content usage has significantly increased recently with users watching training videos, using the latest office applications, or using image-editing tools. An increased media streaming workload is a challenge for a virtualization environment, because it adds load to the environment and degrades performance. If the environment is not virtualized, it can become difficult for administrators to provide security updates for workstations spread across the enterprise, because they can push updates only when mobile devices are connected to the network.

NetApp HCI with a graphics adapter allows users to stream media (even at 4K resolution) with a virtual desktop or in a hosted app environment. With VMware Horizon, customers can securely host their virtual desktop and apps in the data center, so users can use any supported device to access desktops or applications in the enterprise or from remote locations.

As you start designing graphics-intensive workloads on virtual desktops, you face several questions:

- What guest operating systems can I use?
- How can I provide high availability for the solution components?
- What versions of DirectX or OpenGL does it support?
- Does the environment support CUDA?
- How many users can I host with each server?
- How do I build a cost-effective solution to meet customer demands?
- What security features are available?
- How do I scale the infrastructure for workload demands?
- How do I manage the infrastructure?

This solution address all of these questions.

3.1 Target Audience

The target audience for this technical report includes the following groups:

- Field consultants: to help with design decisions with the VMware End-User Computing environment
- Executives and sales engineers: to understand the value of the solution
- Professional services and IT managers: to understand and identify the components of the solution
- Partners: to learn and assist the customers who face similar challenges

3.2 NetApp HCI Use Cases

In addition to the previously mentioned benefits, NetApp HCI is ideal for the use cases described in this section. For customers performing the following tasks, NetApp HCI is architected to deliver exceptional value:

- Deploying private clouds
- Designing end-user computing environments that include hardware-accelerated graphics applications
- Considering workload consolidation

Private Cloud

NetApp HCI is an optimal foundation for an enterprise private cloud model, whether you choose OpenStack, VMware, or a solution developed in-house. This is because NetApp HCI uses native NetApp Element APIs that allow the on-demand provisioning of workloads through storage drivers and management plug-ins.

As an example, NetApp HCI integrates with VMware Virtual Volumes, enabling VMware administrators to achieve the most granular control over storage performance on a per-VM basis. With this functionality, you can set minimum, maximum, and burst IOPS levels, confirming exact amounts of capacity and performance for even the most sensitive virtual machines (VMs). You can change capacity and performance dynamically without migrating data or affecting system performance.

End-User Computing

NetApp HCI is optimal for end-user computing environments because capacity and performance are allocated independently for every virtual desktop and every application. The allocations can be easily adjusted as workloads shift or needs evolve without complexity. If an application needs more performance, the initial configuration can become a bottleneck, but NetApp HCI eliminates the penalty for underestimating requirements. You can modify the QoS policies to easily change the settings for minimum, maximum, and burst. The new settings take effect immediately.

Workload Consolidation

NetApp HCI eliminates workload silos, allowing customers to predictably run multiple applications on the same infrastructure. Traditionally, when multiple applications share infrastructure, all performance resources, both IOPS and bandwidth, are freely available to all applications all the time across the shared resources. Without a more precise resource allocation, one application or “noisy neighbor” can easily consume an unfair share of the resources, leaving little available for others. This “first-come, first-served” allocation methodology can have a huge negative effect on all the other applications on the system.

Performance expectations on an application-by-application basis are erratic and unpredictable. One misbehaving application can cripple the entire system. To keep these variances in check, customers must constantly monitor and manage which applications share resources. Often, alleviating resource contention requires migration of either the “noisy neighbor” or the unhappy customer to a new system.

The NetApp HCI QoS settings eliminate resource contention and the variable application performance caused by noisy neighbors. Each volume on the system is assigned its own minimum, maximum, and burst settings. Thus, performance becomes predictable for each application, but you avoid the capacity sprawl and low utilization that are common in today’s modern infrastructures.

3.3 Solution Use Case Summary

This solution applies to the following use cases:

- Hardware-based graphics acceleration for multimedia workloads
- On-demand desktop and application deployment for end users

- Self-service of user profile management for end users
- Ease of management for security patches
- Integration with an existing private cloud
- Secure multitenant infrastructure for enterprises

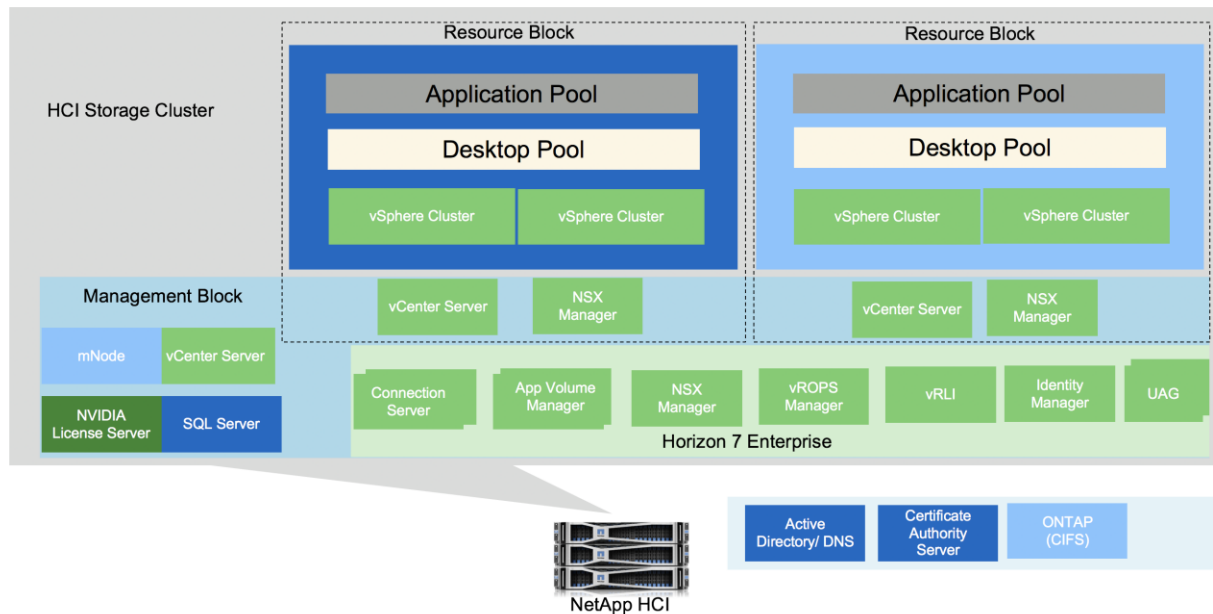
4 Technology Overview

VMware Horizon 7 Enterprise Edition provides components for deploying and managing virtual desktops and hosted apps. To deploy virtual desktops for users in minutes, VMware introduced the JMP technology. JMP uses Horizon Instant Clones for quick virtual desktop provisioning, application assignments from App Volumes, and the association of virtual desktop policies managed by User Environment Manager.

NetApp HCI H610C includes the NVIDIA Tesla M10 graphics adapter, which helps accelerate media streaming for virtual desktops and apps hosted in VMware Horizon.

Figure 4 shows the logical diagram of the components of the VMware Horizon solution on NetApp HCI.

Figure 4) Solution architecture.



4.1 NetApp HCI

NetApp HCI is scale-out architecture that permits the independent expansion of storage or compute capacity. It can scale up to 40 storage nodes and 64 compute nodes per cluster. The form factor of NetApp HCI varies by model. Mixed nodes (nodes of various NetApp HCI models) are supported in a cluster, thus protecting your investment. NetApp HCI provides APIs that provide deep ecosystem integration with various automation tools. The NetApp HCI H610C is a 2U compute node with two NVIDIA Tesla M10 graphics adapters, which accommodates a greater number of task and knowledge workers who use graphics-intensive media. An NVIDIA Tesla M10 graphics card is a dual PCIe slot device. Each board is equipped with the following components:

- Four mid-tier NVIDIA Maxwell GPUs
- 32GB of GDDR5 memory (8GB per GPU)

- 2560 NVIDIA CUDA cores (640 per GPU)
- 28 H.264 1080p30 streams

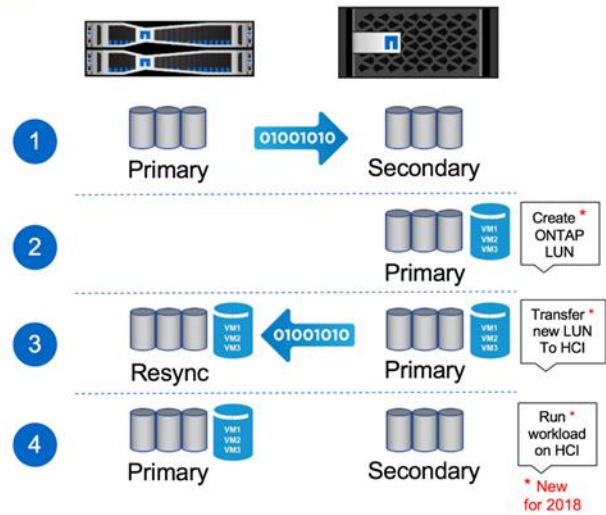
NetApp HCI provides the NDE to minimize the number of inputs and steps required to configure the system for a VMware vSphere environment. The NDE creates the storage cluster and the compute cluster, configures networks including iSCSI, and creates an initial datastore. Additional datastores can be provisioned and managed from the vCenter plug-in that NetApp HCI provides. NetApp HCI storage supports hardware acceleration (VAAI, or vSphere Storage APIs—Array Integration), and no further tuning is required for VMFS6 volumes.

Figure 5) NetApp SolidFire to NetApp ONTAP disaster recovery solution.

SolidFire-to-ONTAP disaster recovery solution

New functionality available with ONTAP 9.4 and Element 10.3

- ONTAP LUNs can now be replicated to Element OS
- Create new LUNs while failed over to ONTAP and restore to HCI on failback
- Migrate volumes to NetApp HCI using SnapMirror
- Increased fan-in ratio of up to 32 HCI clusters to one ONTAP cluster
- SnapMirror to ONTAP Select now supported
- ONTAP Cloud support planned for Element 11.0



NetApp HCI storage provides inline duplication, QoS, and RESTful APIs to integrate into any automation framework. NetApp HCI can perform synchronous and asynchronous replication between NetApp HCI systems or to NetApp ONTAP® systems on premises or in the cloud.

NetApp HCI uses 10/25Gb ports for the iSCSI, virtual machine, and vMotion traffic and 1Gb for the Intelligent Platform Management Interface (IPMI) and management ports. Although the NDE supports both virtual LAN (VLAN) untagged and tagged networks, it is easier to work with tagged networks. Configure any uplink ports used in vSphere distributed switches as trunk ports on the network switches, thus creating appropriate VLANs.

4.2 VMware vSphere

Separate vSphere clusters are recommended for management, desktop pools, and application pools to provide logical separation and fault isolation between the components. NetApp recommends that you enable high availability and the Dynamic Resources Scheduler (DRS) on those vSphere clusters. High availability provides fault tolerance to host failures, which results in only a short outage before VMs are automatically brought back online.

Enable host monitoring and admission control so that at least one host failure or maintenance operation can be tolerated while sufficient resources are still provided to run the entire workload of the cluster. More capacity can be reserved to provide greater headroom for concurrent host failures or maintenance.

DRS automatically balances CPU and memory workloads across the cluster members. vSphere supports up to 64 nodes per cluster. VMware Horizon supports up to 32 nodes for linked clones and up to 64 nodes for the Instant Clones.

vCenter Server provides the control pane for the vSphere environment. Up to 15 vCenter instances can be joined in linked mode. The NDE deploys one vCenter Server, or it can join to any existing vCenter Server. The NetApp HCI vCenter Plug-In supports linked mode and can provision datastores for hosts within that vCenter data center. A vSphere host can exist in only one vCenter data center.

Table 1) vSphere maximums.

Item	Maximum	Description
VMs per host	1,024	For H610C nodes, with vGPU: 128.
Virtual disks per host	2,048	
VMFS volume size	64TB	
Volumes per host	1,024	
Hosts per volume	64	
Powered-on VMs per VMFS volume	2,048	
Maximum file size on VMFS5/VMFS6	62TB	
Shared GPUs per ESXi host	16	
Hosts per vCenter Server	2,000	
Linked vCenter Servers	15	
Hosts in linked vCenter Servers	5,000	
Powered-on VMs in linked vCenter Servers	50,000	
Latency between vCenter instances in linked mode	100ms	

Graphics adapters in vSphere hosts can be used in three ways:

- Virtual Dedicated Graphics Acceleration (vDGA)
- Virtual Shared Graphics Acceleration (vSGA)
- Virtual Shared Pass-Through Graphics Acceleration (with NVIDIA, referred to as “vGPU”; with AMD, referred to as “multiuser GPU” or “MxGPU”)

4.3 Virtual Dedicated Graphics Acceleration (vDGA)

VMs access the graphics adapter using a PCI pass-through device. You must enable DirectPath for that device on the vSphere host (see Figure 6). Using the pass-through device, the VMs cannot be suspended, be migrated with vMotion, or use VM snapshot features.

Figure 6) Enable DirectPath.

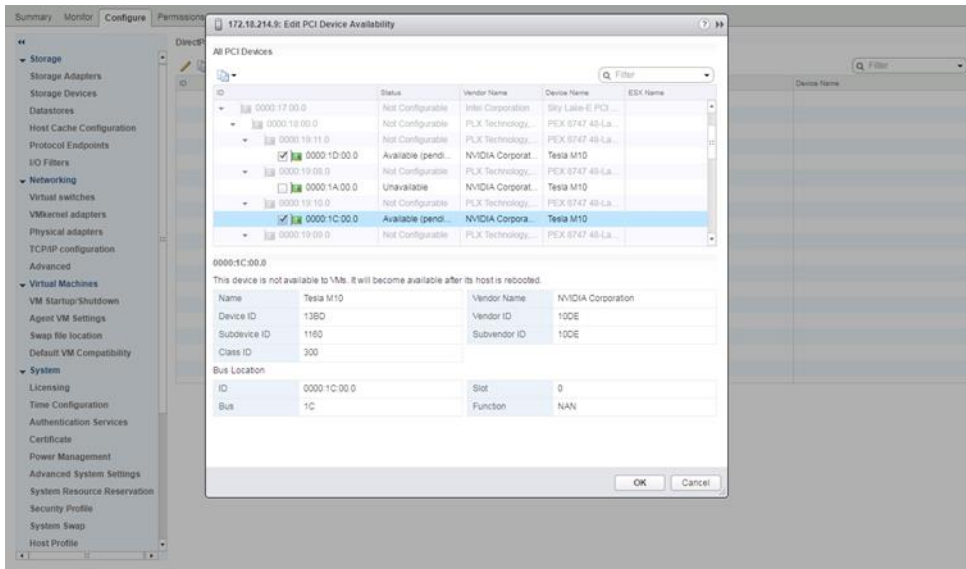
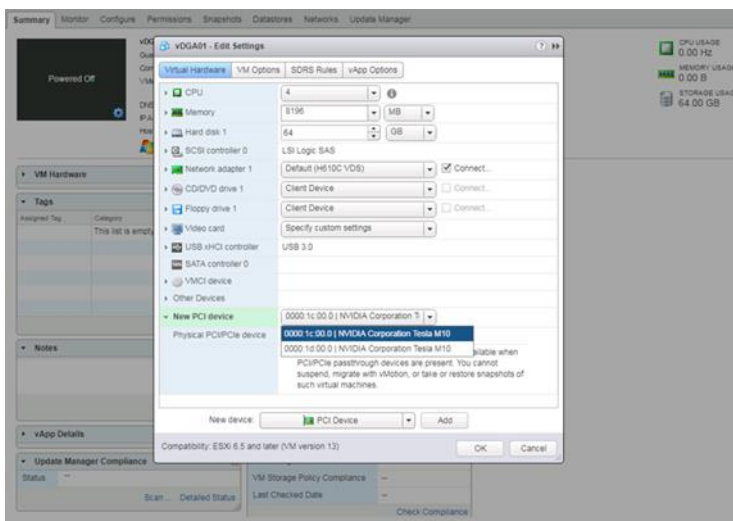


Figure 7) Pass-through PCI device for VMs.



A native graphics adapter driver must be installed on the VM, but no drivers are required on the hypervisor. The VM has access to the complete GPU and GPU memory even if it is idle or powered off. NetApp doesn't recommend using vDGA, because it doesn't support many vSphere features and it provides low resource utilization.

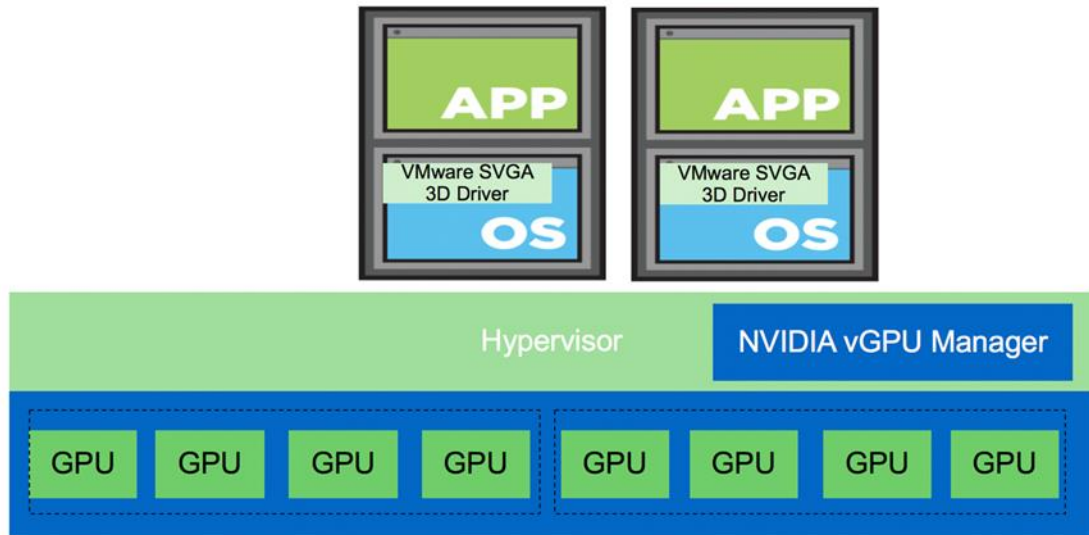
Partners might need to demonstrate the NVIDIA vGPU vMotion feature on a single server and use vSphere nested virtualization. NetApp hasn't tested this functionality, and you should know its limitations before you consider it. With NVIDIA, it requires a Quadro Virtual Data Center Workstation license or a GRID Virtual Apps license.

4.4 Virtual Shared Graphics Acceleration (vSGA)

vSGA provides hardware 3D acceleration by sharing GPUs across multiple VMs. This solution is attractive for users who require a GPU's full potential during brief periods. The NVIDIA vGPU Manager or the GRID software driver is installed on hypervisor hosts, and the VMware 3D driver, which is part of

VMware Tools, is installed on the guest OS. Graphics API support is limited to vSGA, and only selective versions of DirectX and OpenGL are supported. There is no CUDA support.

Figure 8) vSGA.



To install the NVIDIA driver on multiple hypervisor hosts, VMware vSphere Update Manager can be used with the host extension type. You can monitor VM GPU association by using the vSphere web client or the `gpuvvm` CLI command.

Figure 9) gpuvvm CLI.

VM GPU Info

Using vSphere Web Client / gpuvvm CLI

Name	Vendor	Active Type	Configured Type	Memory
NVIDIA7510 M10	NVIDIA Corporation	Shared	Shared	7.56 GB
NVIDIA7510 M10	NVIDIA Corporation	Shared	Shared	7.56 GB
NVIDIA7510 M10	NVIDIA Corporation	Shared	Shared	7.56 GB
NVIDIA7510 M10	NVIDIA Corporation	Shared	Shared	7.56 GB
NVIDIA7510 M10	NVIDIA Corporation	Shared	Shared	7.56 GB
NVIDIA7510 M10	NVIDIA Corporation	Shared	Shared	7.56 GB

Name	Power	Mode	Reserved Space	Used Space	Head CPU	Head Mem
HE100-L0015	Powered On	Normal	80.20 GB	16.28 GB	20.96%	1.92GB MB
HE100-L0033	Powered On	Normal	80.20 GB	16.30 GB	20.94%	1.74GB MB
HE100-L0024	Powered On	Normal	80.33 GB	16.38 GB	20.94%	1.76GB MB
HE100-L0041	Powered On	Normal	80.10 GB	16.22 GB	20.94%	2.07GB MB
HE100-L0045	Powered On	Normal	80.23 GB	16.29 GB	20.94%	1.39GB MB
HE100-L0052	Powered On	Normal	80.24 GB	16.3 GB	20.94%	2.05GB MB
HE100-L0056	Powered On	Normal	80.18 GB	16.25 GB	20.94%	2.04GB MB
HE100-C000	Powered On	Normal	80.26 GB	16.32 GB	20.94%	1.80GB MB

```

root@R204107:~# gpuvvm
Reserved: vmx15, PCI ID 0:26:00:0, vSGA mode, GPU maximum memory 0360112KB
pid 108804, VM "HE100-L0015", reserved 131072KB of GPU memory.
pid 108807, VM "HE100-L0022", reserved 131072KB of GPU memory.
pid 111100, VM "HE100-L0033", reserved 131072KB of GPU memory.
pid 109448, VM "HE100-L0036", reserved 131072KB of GPU memory.
pid 111259, VM "HE100-L0043", reserved 131072KB of GPU memory.
pid 111264, VM "HE100-L0058", reserved 131072KB of GPU memory.
pid 111264, VM "HE100-L0067", reserved 131072KB of GPU memory.
pid 111268, VM "HE100-L0031", reserved 131072KB of GPU memory.
GPU memory left: 7819852KB.

Reserved: vmx17, PCI ID 0:27:00:0, vSGA mode, GPU maximum memory 0360112KB
pid 108808, VM "HE100-L0019", reserved 131072KB of GPU memory.
pid 109449, VM "HE100-L0020", reserved 131072KB of GPU memory.
pid 109448, VM "HE100-L0036", reserved 131072KB of GPU memory.
pid 111320, VM "HE100-L0043", reserved 131072KB of GPU memory.
pid 111269, VM "HE100-L0058", reserved 131072KB of GPU memory.
pid 111353, VM "HE100-L0026", reserved 131072KB of GPU memory.
pid 111270, VM "HE100-L0037", reserved 131072KB of GPU memory.
GPU memory left: 7819852KB.

Reserved: vmx12, PCI ID 0:28:00:0, vSGA mode, GPU maximum memory 0360128KB
pid 111280, VM "HE100-L0015", reserved 131072KB of GPU memory.
pid 108809, VM "HE100-L0022", reserved 131072KB of GPU memory.
pid 112927, VM "HE100-L0041", reserved 131072KB of GPU memory.
pid 112830, VM "HE100-L0052", reserved 131072KB of GPU memory.
pid 111262, VM "HE100-L0058", reserved 131072KB of GPU memory.
pid 112931, VM "HE100-L0056", reserved 131072KB of GPU memory.
pid 121015, VM "HE100-L0040", reserved 131072KB of GPU memory.
pid 112947, VM "HE100-L0040", reserved 131072KB of GPU memory.
GPU memory left: 7819852KB.

Reserved: vmx18, PCI ID 0:29:00:0, vSGA mode, GPU maximum memory 0360128KB
pid 108810, VM "HE100-L0022", reserved 131072KB of GPU memory.
pid 108807, VM "HE100-L0036", reserved 131072KB of GPU memory.
pid 111326, VM "HE100-L0030", reserved 131072KB of GPU memory.
pid 111262, VM "HE100-L0039", reserved 131072KB of GPU memory.
pid 111268, VM "HE100-L0058", reserved 131072KB of GPU memory.
pid 111271, VM "HE100-L0058", reserved 131072KB of GPU memory.
pid 111268, VM "HE100-L0068", reserved 131072KB of GPU memory.
pid 111260, VM "HE100-L0040", reserved 131072KB of GPU memory.
GPU memory left: 7819852KB.

Reserved: vmx16, PCI ID 0:26:00:0, vSGA mode, GPU maximum memory 0360128KB
pid 108810, VM "HE100-L0022", reserved 131072KB of GPU memory.
pid 108861, VM "HE100-L0034", reserved 131072KB of GPU memory.
pid 111269, VM "HE100-L0058", reserved 131072KB of GPU memory.
pid 111252, VM "HE100-L0056", reserved 131072KB of GPU memory.
pid 111267, VM "HE100-L0062", reserved 131072KB of GPU memory.
pid 111269, VM "HE100-L0069", reserved 131072KB of GPU memory.
pid 111259, VM "HE100-L0054", reserved 131072KB of GPU memory.
GPU memory left: 7450624KB.
    
```

The vSGA mode provides higher user density with little management overhead. This mode is the default option when you install the NVIDIA driver. All the vSphere features—such as vMotion, Suspend, and snapshots—are supported with a wide range of NVIDIA driver versions.

When no graphics adapter driver is installed on the hypervisor, it shows the memory size as zero on the vSphere web client.

4.5 Virtual Shared Pass-Through Graphics Acceleration with the NVIDIA vGPU

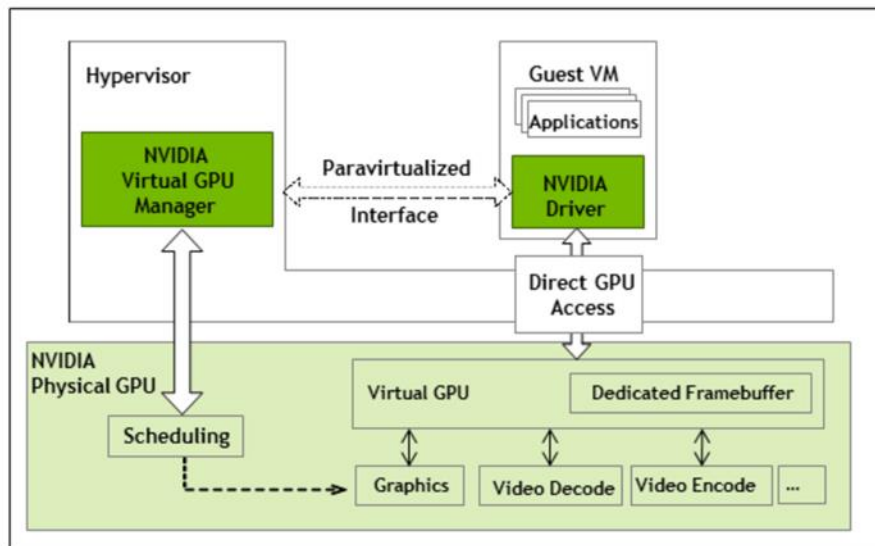
When users demand more graphics performance than the consolidation ratio provides in a given environment, the NVIDIA vGPU provides the balance of dedicated resource features, such as vDGA and higher consolidation ratio like vSGA.

In a manner similar to how vSphere virtualizes hardware resources, the NVIDIA vGPU Manager and the GRID software virtualize the GPU hardware.

Figure 10) NVIDIA vGPU architecture ([source](#)).

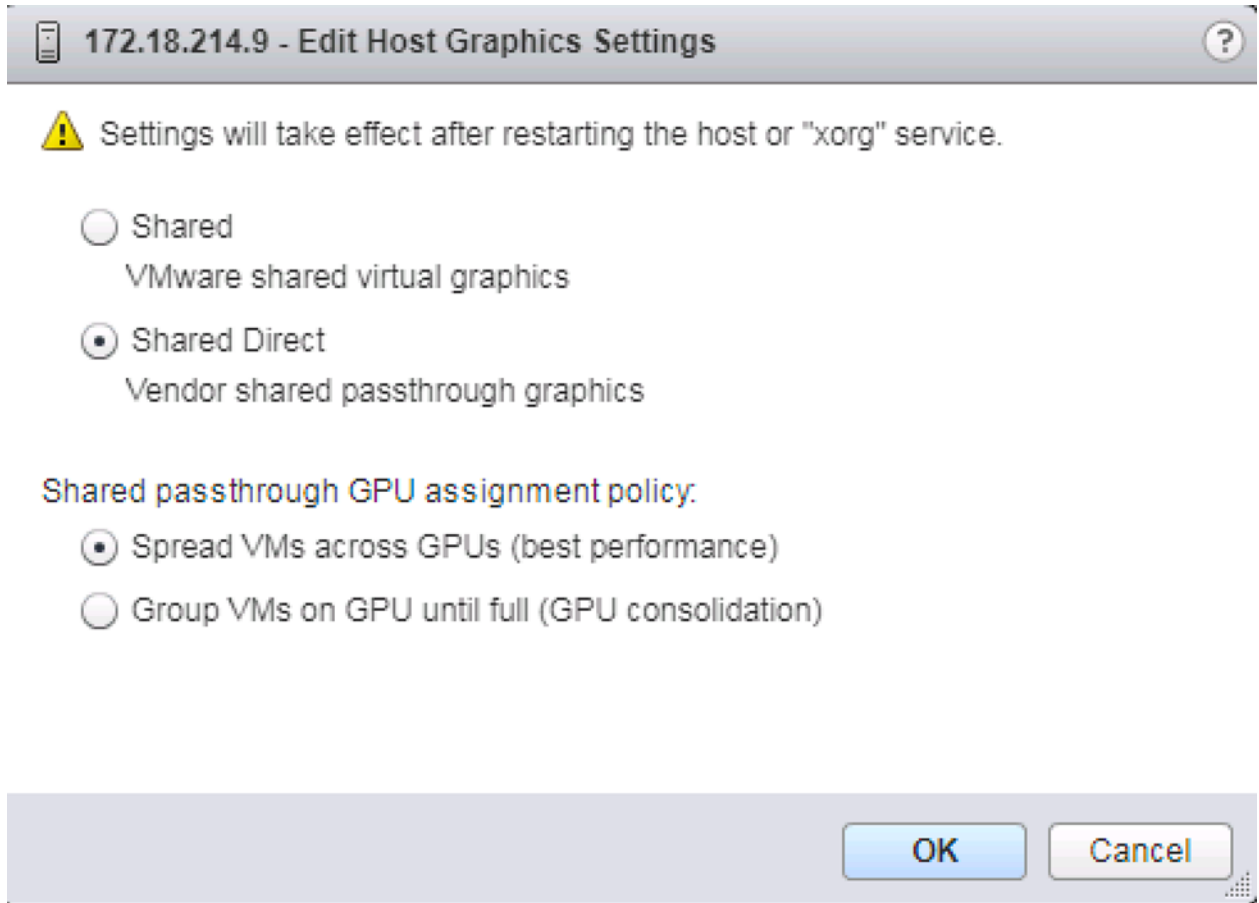
NVIDIA vGPU Architecture

From their user guide



To use this mode after installing the NVIDIA driver on the hypervisor host, switch the graphics adapter mode, as Figure 11 shows.

Figure 11) Host graphics settings.



A shared PCI device is added to the VM, and the appropriate GPU profile is selected. The GPU profile list is provided in Figure 12.

Figure 12) vGPU profiles for Tesla M10 ([source](#)).

Physical GPUs per board: 4

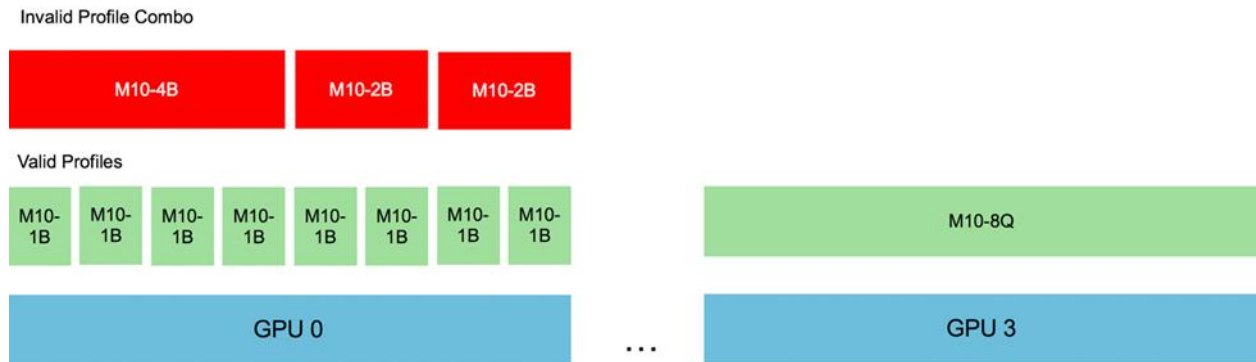
Virtual GPU Type	Intended Use Case	Frame Buffer (Mbytes)	Virtual Display Heads	Maximum Resolution per Display Head	Maximum vGPUs per GPU	Maximum vGPUs per Board	Required License Edition
M10-8Q	Designer	8192	4	4096x2160	1	4	Quadro vDWS
M10-4Q	Designer	4096	4	4096x2160	2	8	Quadro vDWS
M10-2Q	Designer	2048	4	4096x2160	4	16	Quadro vDWS
M10-1Q	Power User, Designer	1024	2	4096x2160	8	32	Quadro vDWS
M10-0Q	Power User, Designer	512	2	2560x1600	16	64	Quadro vDWS
M10-2B	Power User	2048	2	4096x2160	4	16	GRID Virtual PC or Quadro vDWS
M10-2B4	Power User	2048	4	2560x1600	4	16	GRID Virtual PC or Quadro vDWS
M10-1B	Power User	1024	4	2560x1600	8	32	GRID Virtual PC or Quadro vDWS
M10-0B	Power User	512	2	2560x1600	16	64	GRID Virtual PC or Quadro vDWS
M10-8A	Virtual Application User	8192	1	1280x1024 ¹	1	4	GRID Virtual Application
M10-4A	Virtual Application User	4096	1	1280x1024 ¹	2	8	GRID Virtual Application
M10-2A	Virtual Application User	2048	1	1280x1024 ¹	4	16	GRID Virtual Application
M10-1A	Virtual Application User	1024	1	1280x1024 ¹	8	32	GRID Virtual Application

A GRID Virtual PC (profiles typically end with B) is used for virtual desktops, and a GRID Virtual App (profiles end with A) is used for hosted apps. Most of the Q profiles (Quadro Virtual Data Center Workstation) support the 4K resolution, which provides an enhanced user experience for image-editing tools.

According to the policy set, the vGPU Manager schedules a VM to the GPU to either spread the VMs (best performance) or to group VMs (GPU consolidation), as Figure 13 shows.

This release of NVIDIA vGPU supports only homogeneous virtual GPUs. At any moment, the virtual GPUs resident on a single physical GPU must all be of the same type. However, this restriction doesn't extend across physical GPUs on the same card. Different physical GPUs on the same card can host different types of virtual GPUs at the same time, as long as the vGPU types on any one physical GPU are the same.

Figure 13) Homogenous virtual GPU profile combination.

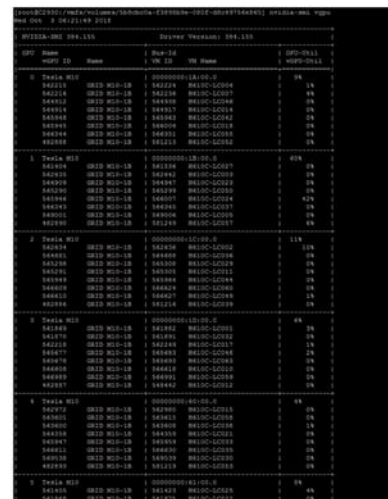
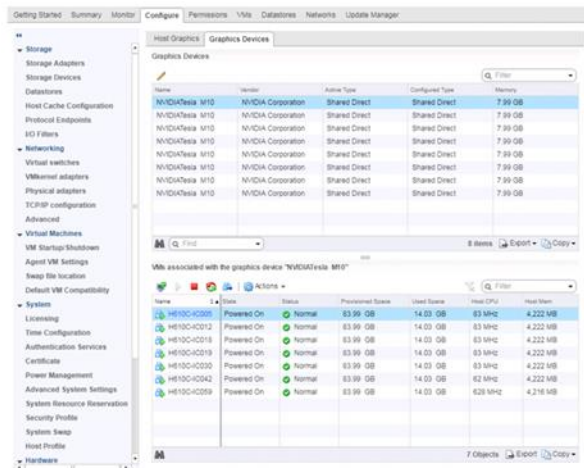


To monitor this relationship, either the vSphere web client or the `nvidia-smi vgpu` CLI can be used.

Figure 14) nvidia-smi vgpu CLI.

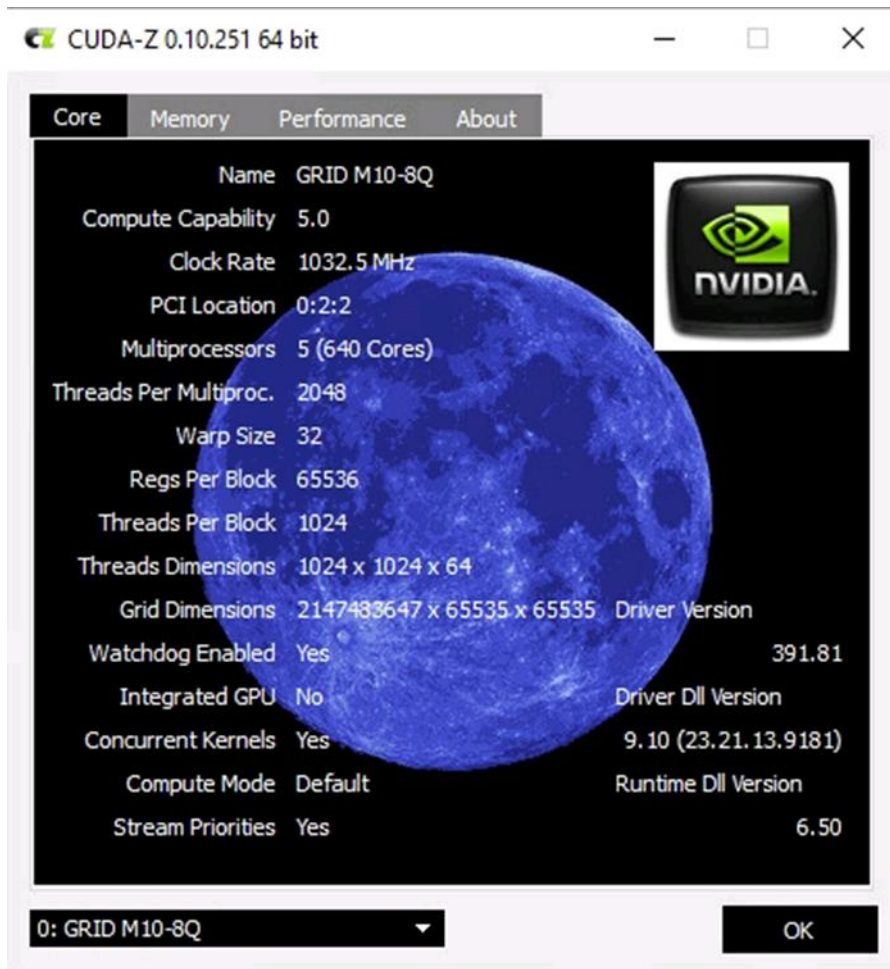
VM GPU Info

With vSphere Web Client/ [nvidia-smi vgpu cli](#)



The vGPU 8Q profile supports CUDA applications. Figure 15 shows a screenshot of CUDA-Z for the 8Q profile.

Figure 15) CUDA-Z screenshot.



The combination of VMware vSphere 6.7 Update 1 and NVIDIA GRID software 7.0 supports vMotion, making it easier to manage the host.

Using the frame buffer for the NVIDIA hardware-based H.264/HEVC video encoder (NVENC) can cause memory exhaustion with vGPU profiles that have 512MB or less of frame buffer. To reduce the risk of memory exhaustion, NVENC is disabled on profiles that have 512MB or less of frame buffer.

Application GPU acceleration remains fully supported and available for all profiles, including profiles with 512MB or less of frame buffer. NVENC support from both Citrix and VMware is a recent feature, so you should experience no change in functionality if you are using an older version.

The following vGPU profiles have 512MB or less of frame buffer:

- Tesla M6-0B, M6-0Q
- Tesla M10-0B, M10-0Q
- Tesla M60-0B, M60-0Q

If you require NVENC, use a profile that has at least 1GB of frame buffer.

To use vGPU, the NVIDIA driver must be installed on the guest OS. You must also have other ways to access the VM, such as VNC or Horizon View Agent Direct-Connection. The vSphere web console can't be used to access the VM console session after driver installation and reboot.

Use of vGPU requires an NVIDIA license for the appropriate profile.

4.6 VMware Horizon 7 Enterprise Edition

With VMware, you can pick a subscription-based license or a perpetual license for desktop pools or application pools. Subscription pricing options are available for Horizon 7, Horizon Cloud Service with Hosted Infrastructure, and Horizon Cloud on Microsoft Azure. Horizon 7 subscription licenses can be used on premises or with VMware Cloud on AWS to burst to the cloud. Horizon 7 provides IT with a new streamlined approach to deliver, protect, and manage Windows and Linux desktops and applications while containing costs and making sure that end users can work anytime, anywhere, and on any device.

VMware Horizon 7 Enterprise Edition includes the following components:

- VMware Identity Manager
- VMware Horizon (Connection Server, Agent, Client, and Unified Access Gateway [UAG])
- VMware ThinApp
- VMware App Volumes
- User Environment Manager
- Just-in-Time Management Platform (JMP)
- VMware Mirage
- VMware vSphere Hypervisor
- VMware vCenter Server
- VMware NSX for vSphere
- VMware vRealize Operations Manager (including for vRealize Operations for Published Applications)
- VMware vRealize Orchestrator Plug-In for Horizon
- VMware vRealize Log Insight
- VMware Workstation
- VMware Fusion

4.7 VMware Identity Manager

VMware Identity Manager provides single sign-on (SSO) for virtual desktops and applications, which includes SaaS, web, ThinApp, and mobile applications. With SSO, users do not need to remember multiple user names and passwords. It provides a central location to instantly disable user access to all resources, which protects systems from data leakage. It also acts as an app store for the company, and it provides a self-service catalog that can be customized to add branding.

Figure 16) Web launcher portal ([source](#)).

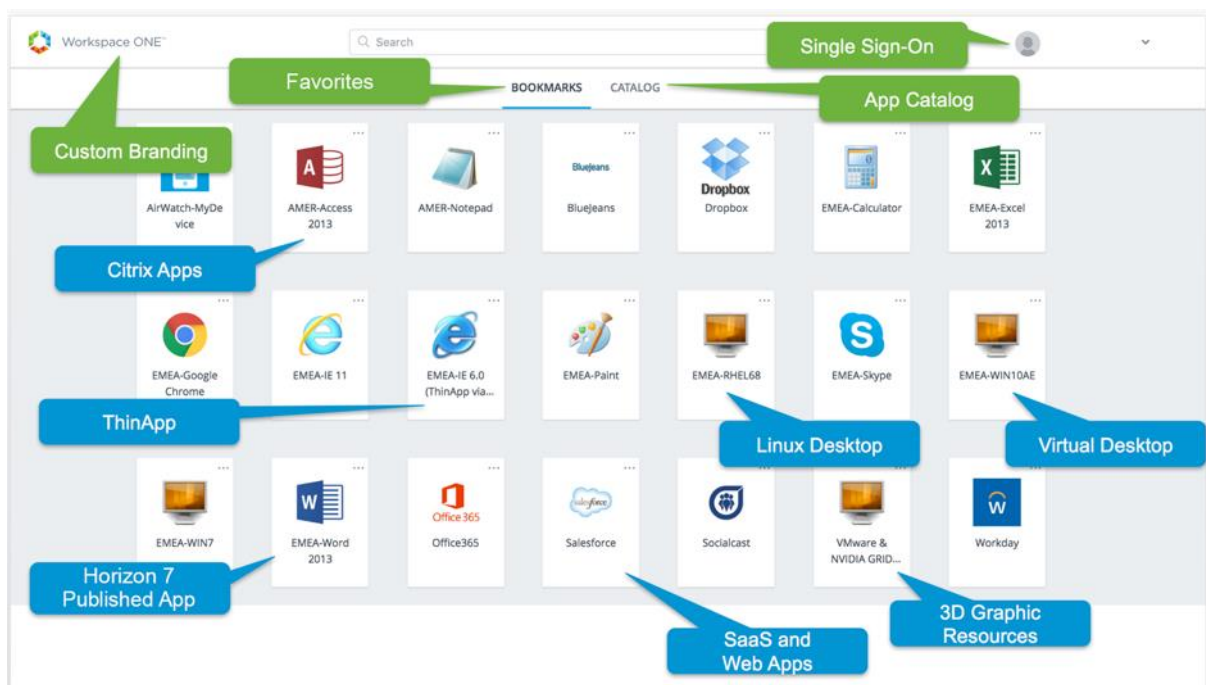


Figure 16 shows the web launcher portal with apps from various resources. This component is optional for this solution. Notably, the Horizon client can show the links to virtual desktops and published apps. VMware Identity Manager is part of the private cloud solution to provide SSO. If you want to implement SSO, see the [Reviewer's Guide for On-Premises VMware Identity Manager](#).

4.8 VMware Horizon

VMware Horizon deploys and manages desktop pools and application pools, manages the connection broker and user entitlements, and so on. For virtual desktops, it can automatically provision full clones, linked clones, or Instant Clones in VMware vSphere environments. It can also broker connections to preexisting physical machines or vCenter VMs, including Remote Desktop Session Hosts (RDSHs), to provide session-based desktops (like terminal services).

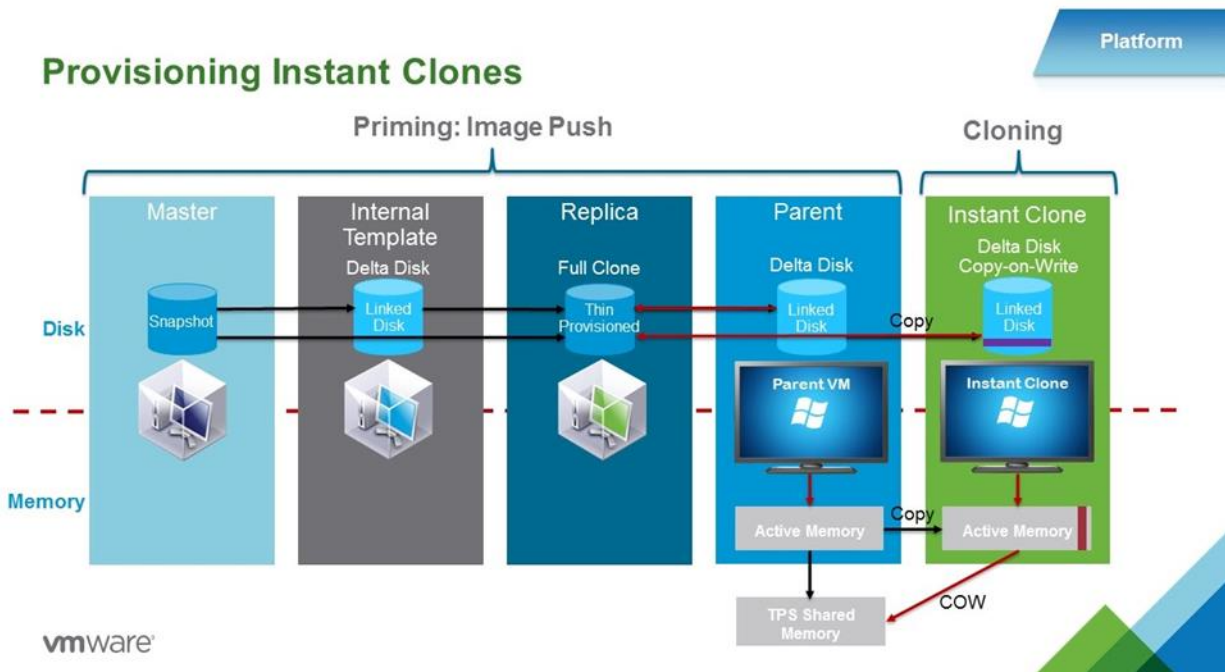
The connection server can provide dedicated or floating assignments to desktop pools. With dedicated assignments, a one-to-one relationship is maintained between users and computers. Therefore, when a user logs in, the user always gets the same desktop. The assignments can be made manually or automatically when users first log in. If you do not need a one-to-one user-to-machine relationship—for example, if users are working multiple shifts, sharing the same computer—you should use a floating pool.

A full clone is cloned from an existing vSphere template, whereas a linked clone or an Instant Clone is provisioned from VMware snapshots.

If you are not using thin provisioning from vCenter or from the storage vendor, the storage requirements for virtual desktops are huge. The storage I/O requirements for virtual desktops with spinning disks were challenging, and many companies have kept a separate silo infrastructure for virtual desktops as a result. With NetApp HCI, you can consolidate these workloads along with other infrastructure workloads on an all-flash storage system with guaranteed service levels.

Linked clones help organizations save on storage space. However, the provisioning time is longer, and linked clones require additional components such as a SQL database and a composer service. An Instant Clone doesn't require any external databases, and it uses fewer vCenter operations than a linked clone.

Figure 17) Provisioning Instant Clones.



Priming or image push operations include the creation of an internal template that is a linked clone of the master image snapshot. The internal template joins the computer to the Windows domain and reboots the computer, and then the replica is created for every datastore that is chosen for the provision of pools. The replica is a thin-provisioned, full clone of the internal template. The parent VM is booted from the replica on every host that is part of the resource pool selected for the desktop pool. The booted-up parent VM is quiesced and “hot-cloned” to rapidly produce derivative (child) VMs, taking advantage of the same disk and memory as the parent VM. The clone starts in a booted-up state.

An Instant Clone, even though the initial priming process takes some time, can expand a pool in seconds. An Instant Clone uses vSphere VMFork technology to provision desktops, and it joins the domain by using cloneprep without any reboots required.

The master image should have Horizon Agent installed so that it can communicate with Horizon connection servers, and virtual desktops can be managed from the connection server. The default option is to install the Horizon View Composer Agent. Make sure to disable the VMware Horizon View Composer Agent and use the VMware Horizon Instant Clone Agent.

Horizon 7 supports both Windows and Linux-based desktops, including the Red Hat Enterprise Linux, Ubuntu, CentOS, and NeoKylin operating systems. For application pools, it needs RDSH servers deployed in a farm. On supported Windows Servers, enable the Remote Desktop Services (RDS) role and select the RDSH role services. If you plan to host graphics applications, install Horizon Agent with the 3D RDSH option enabled.

A connection server can create a farm automatically from the VM snapshot, or it can manage manually deployed physical machines or vCenter VMs. The manual option can also be used for the VMs provisioned from other tools, such as VMware vRealize Automation as part of a private cloud.

While deploying an automated farm, remember to set the maximum sessions per RDS host to 60.

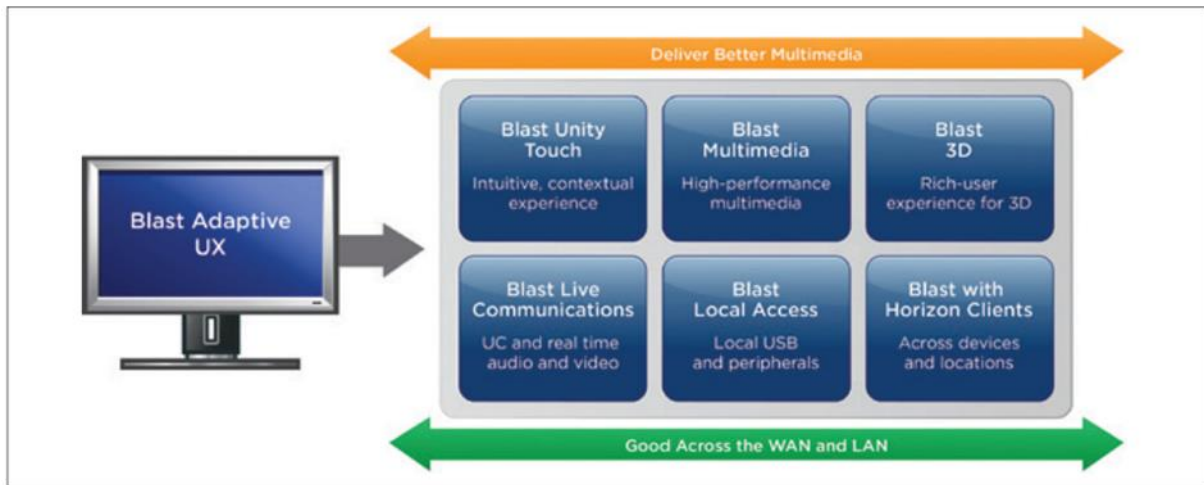
In one farm, up to 200 RDSHs can be added. After the farms are defined, RDS-based desktop pools or application pools can be created from the farm. While the application pool is being created, the pool can

automatically sense the installed applications from the farm and pick the applications to publish. If an app is not listed, you can add it manually by providing its name and the executable path.

Within a connection server, you can entitle desktop pools or application pools to users or groups. After entitlement, users can launch a Horizon client to connect to the desktop or application.

If an HTML access agent is installed and enabled in the pool, users can launch the desktop or application from a web browser, which makes access from any device easy. VMware Horizon 7 provides the Blast communication protocol along with PCoIP. Blast Extreme is a new display technology built on the H.264 protocol. Blast Extreme Adaptive Transport (BEAT) maintains a great user experience across a wide variety of network types, ranging from a corporate LAN to public Wi-Fi and mobile networks.

Figure 18) Blast protocol features.



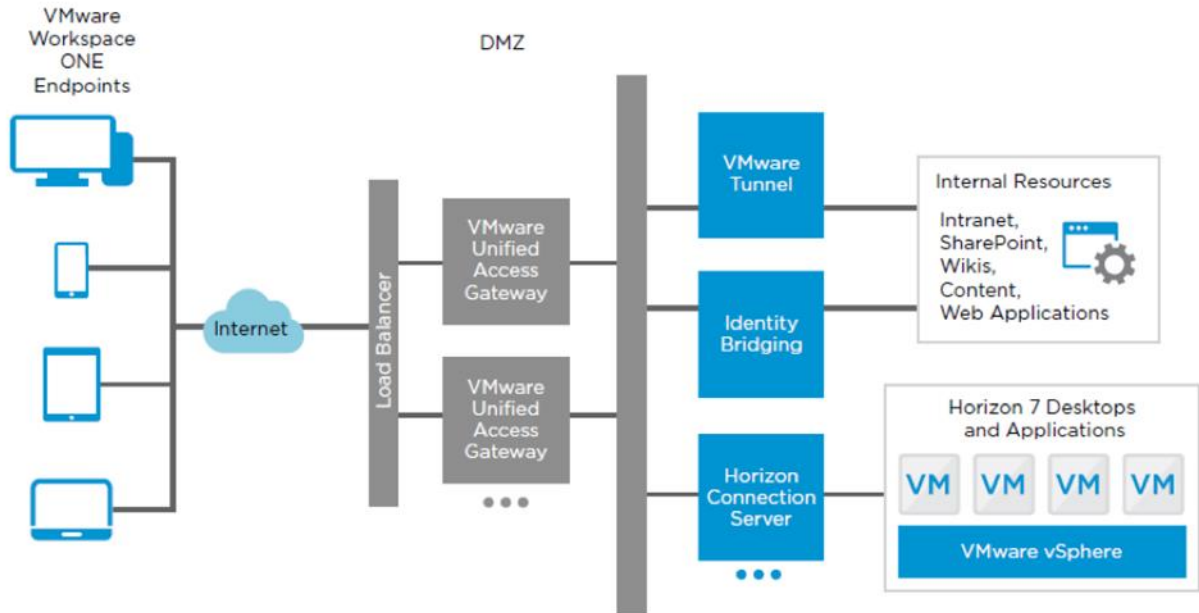
Blast Multimedia provides high-performance multimedia streaming for a rich user experience. Blast 3D provides rich virtualized graphics that deliver workstation-class performance. Blast Local Access provides access to local devices, USB, and device peripherals.

The Session Collaboration feature of Horizon Agent allows multiple users to view and modify the same desktop. This feature can be useful in healthcare, design, engineering, and education organizations for peer reviews, design iterations, and training. The desktop owner can invite users to collaborate in real time on their desktops.

One connection server can handle up to 4,000 sessions (2,000 sessions is recommended), and connection servers (up to a total of seven) can be added for high availability with N+1 and load balancing. The Horizon clients can access the connection servers through a load balancer to distribute the load. If NSX is deployed, the load balancer is available as part of NSX Edge services.

When users try to connect remotely, they can access the desktop pools, hosted applications, intranet resources, or SaaS-based applications through the UAG without VPN access. One UAG can support 2,000 connections.

Figure 19) Unified Access Gateway logical diagram.



NetApp recommends having the Horizon 7 Enterprise components in the management block and keeping the desktop pool and application pool in the resource blocks, each with its own vCenter and NSX server, as shown in Figure 4. A Horizon pod—a combination of management blocks and resource blocks—can support up to 10,000 connections. If more connections are required, you must enable the Cloud Pod Architecture (CPA) feature, which is a federated pod.

Table 2) Horizon sizing.

Item	Maximum
Number of active sessions in the CPA	200,000
Number of pods in the CPA	25
Number of sites in CPA	10
Active connections per pod	20,000 (10,000 recommended)
Number of active connection server instances per pod	7
Active sessions per connection server with direct connection, RDP, tunnel connection, or PCoIP	4,000 (2,000 recommended)
Sessions per RDSH	150 (60 recommended)
RDSHs per farm	200
VMs per vCenter instance	10,000 (5,000 for Instant Clones)
VMs per pool	4,000 (2,000 recommended)

Item	Maximum
VMs per LUN	500
Hosts per vSphere cluster	64 (Instant Clones) 32 (linked clones)
vGPU-enabled VMs per vSphere host	128
Connection servers per pod	7
Connections per UAG	2000

The CPA allows customers to dynamically move and locate Horizon desktop pools and application pools across multiple data centers for efficient management of end users across distributed locations.

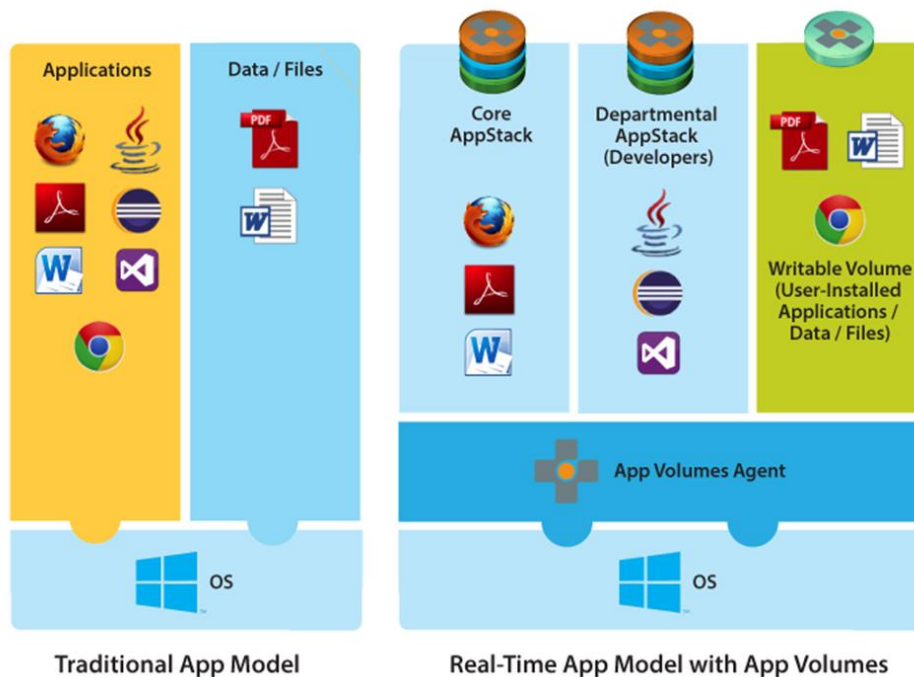
4.9 VMware ThinApp

VMware ThinApp is an agentless architecture for application virtualization and packaging. This product is primarily provided for backward compatibility. If you are already using VMware ThinApp, you can continue to use it in the Horizon 7 environment. For new deployments, VMware App Volumes is preferred.

4.10 VMware App Volumes

VMware App Volumes is a Ruby-based application running on NGINX to provide applications on demand to users and computers. The App Volumes Manager runs this application and requires an external SQL database to store metadata. An AppStack is a collection of applications packaged as a virtual disk. An App Volumes agent is installed on virtual desktops or RDSHs that are used for capturing applications and also on virtual desktops or RDSHs that must consume AppStacks.

Figure 20) App Volumes.



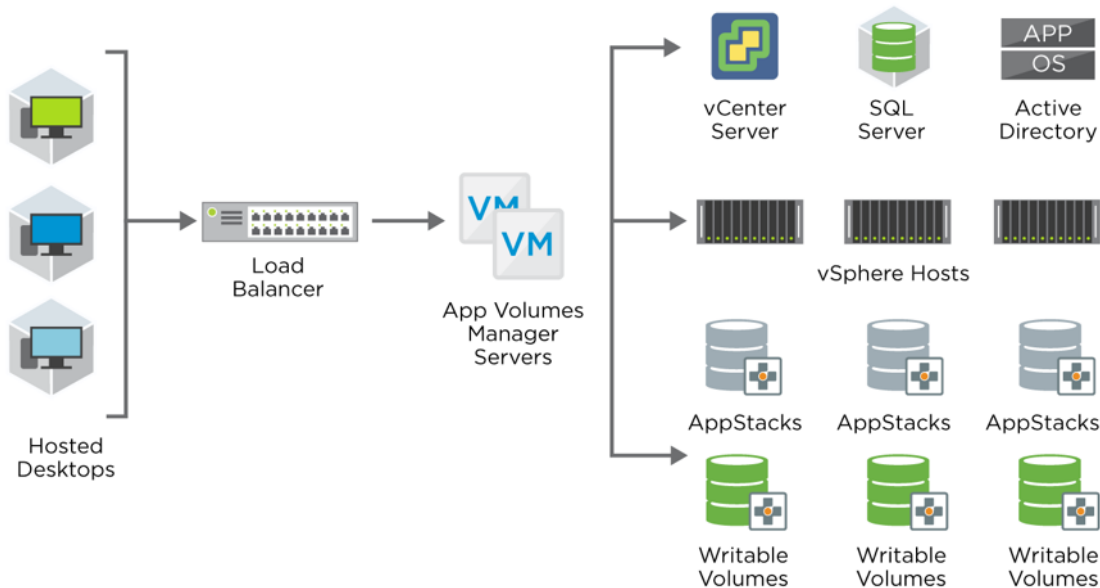
When an AppStack is assigned to users and computers, it only has read-only access to the AppStack volumes. User-writable volumes are used to allow users to install their own applications and to create the profiles.

A major benefit of using App Volumes is to reduce the number of master images in virtual desktop environments by removing the application dependencies on the base image.

Note: An AppStack can only be used with the same OS that is used for capture.

AppStack capture is an easy process. Start AppStack creation by using the App Volumes Manager and pick an AppStack template that was provisioned as part of the deployment process. Also pick one of the machines in your desktop pool or RDSH farm. You must have an App Volumes agent installed on that machine, and the agent must be registered with the App Volumes Manager. Deploy applications manually on that machine, and, after you have installed all the required applications, notify the agent that you have completed the task. The machine reboots and an AppStack is created.

Figure 21) AppStacks and writable volumes.



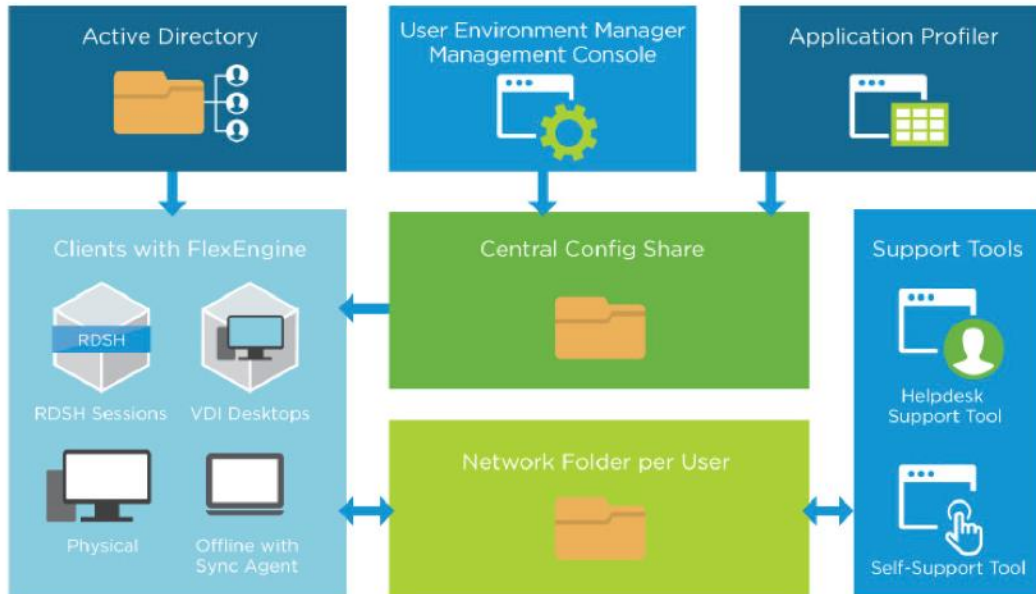
A storage group in an App Volume is a collection of datastores that are used to serve AppStack volumes or user-writable volumes. Storage groups used for AppStack volumes are primarily used for replicating the AppStack volumes across multiple LUNs for availability and load balancing.

VMware recommends that you not assign more than 8 to 10 AppStacks for each user or device. You must mount the writable volume before mounting AppStack volumes to avoid the reboot prompt. If you are using writable volumes, assign AppStack volumes to users rather than to machines. To reduce the performance impact, set the number of AppStack volumes for each user or machine to a low number.

4.11 User Environment Manager

A user profile contains many personalization options for operating systems in addition to application settings. User Environment Manager captures these details to a network share and imports them back according to defined conditions. It doesn't require the user to have a roaming profile to work. User Environment Manager can set smart policies for the Horizon client—for example, enabling printing when accessed in the office and disabling it when accessed from outside.

Figure 22) User Environment Manager file shares.

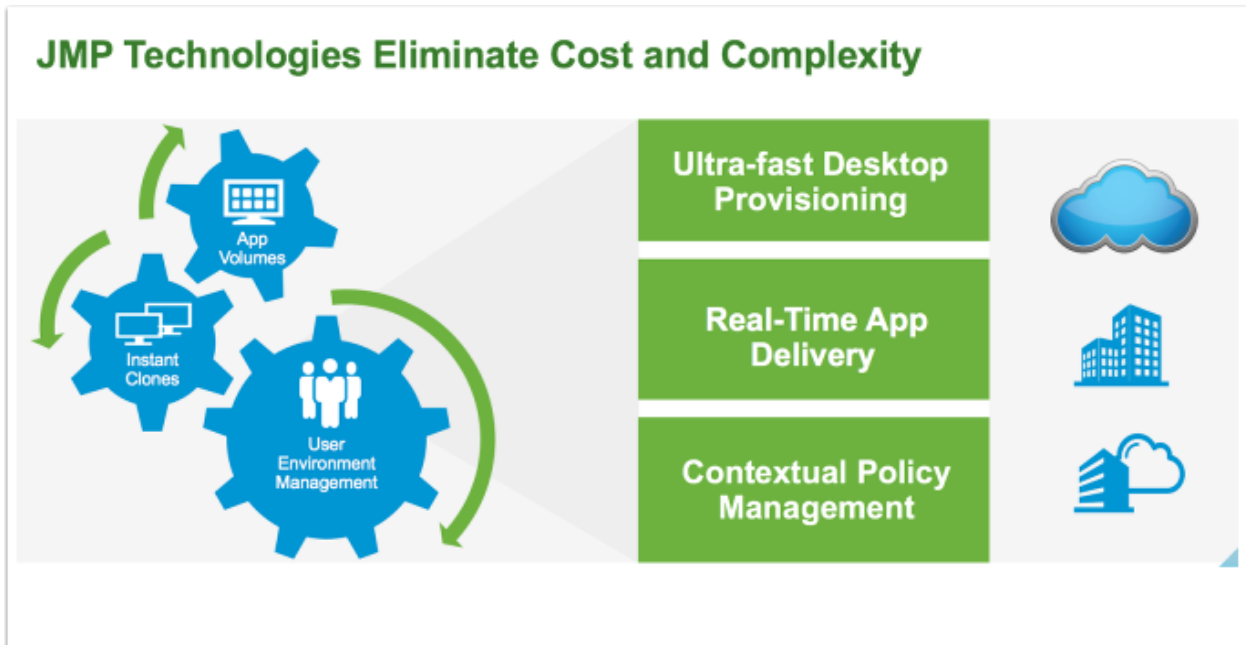


Because this configuration is CIFS-share based, each environment (like testing, production, HR, sales, and engineering) can have its own network configuration shares.

4.12 Just-In-Time Management Platform (JMP)

JMP is a new administrative interface for Horizon connection servers that allows you to pick desktops from the pool and AppStacks from App Volumes. You can also use a single workflow to manage User Environment Manager settings such as ADMX, the display language, drive mappings, environment variables, file and folders, login tasks, application personalization, registry settings, shortcuts, triggered tasks, and Horizon Client smart policies. JMP reduces the number of management interfaces needed for onboarding users.

Figure 23) JMP technologies.



4.13 VMware Mirage

VMware Mirage provides streamlined image management for physical desktops, full-clone virtual machines, and point-of-sale devices. VMware Mirage categorizes a PC into logical layers owned by either IT or the end user. It also sends a complete copy of the system to the Mirage server in the data center and keeps it synchronized. With the layering technology in Mirage, IT has three options for desktop recovery:

- Restore the entire device (OS, applications, user data, and profile)
- Restore just the applications, user data, and profile
- Restore just the user data and profile

VMware Mirage can manage Horizon FLEX (an optional component not included with Horizon 7 Enterprise) virtual images. Horizon FLEX provides virtual desktop access in offline or disconnected mode with VMware Workstation and Fusion products.

If you plan to use VMware Mirage, see [Deployment and Design Considerations for VMware Mirage](#).

4.14 VMware NSX for vSphere

VMware NSX for vSphere brings speed and simplicity to virtual desktop infrastructure (VDI) networking, with policies that dynamically follow virtual desktops. You can create, change, and manage security policies across all virtual desktops with a few clicks. Map these policies to user groups to accelerate virtual desktop onboarding. The security policies can be mapped to users according to role, logical grouping, desktop operating system, and more. VMware NSX Edge services include load balancing (which can be used for connection servers, UAG, and so on), DHCP/DHCP relay (for desktop or application pools), firewall, network address translation (NAT), routing, VPN, and so on.

Figure 24) NSX microsegmentation.



4.15 VMware vRealize Operations Manager

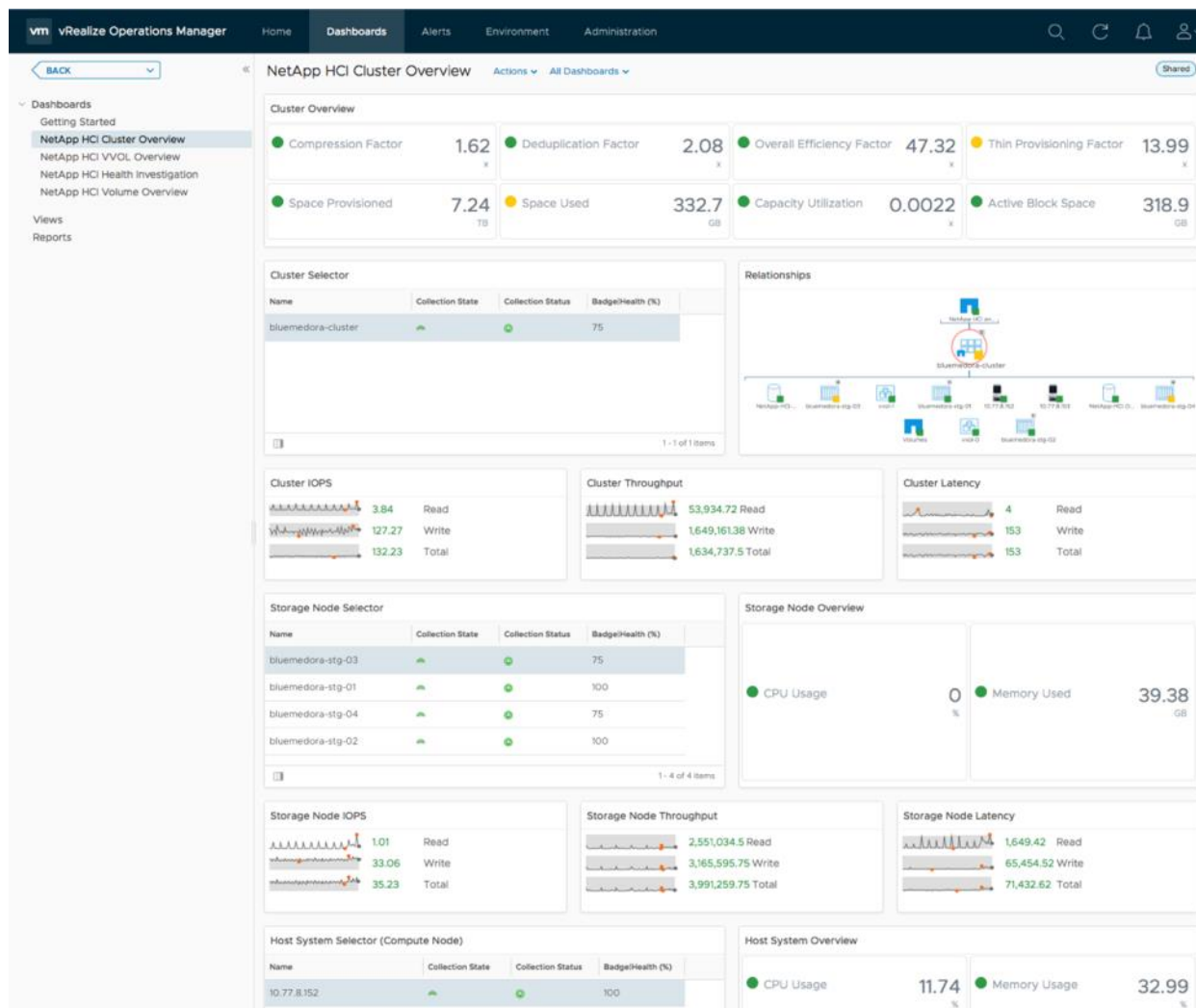
VMware vRealize Operations Manager provides the five key features listed in Figure 25.

Figure 25) Five key features of vRealize Operations Manager.



To monitor NetApp HCI with vRealize Operations Manager, use Blue Medora’s management pack for NetApp HCI.

Figure 26) NetApp HCI integration with vRealize Operations Manager.



NVIDIA provides a GPU management pack for vRealize Operations Manager to monitor metrics from the graphics adapter.

Figure 27) NVIDIA integration with vRealize Operations Manager.

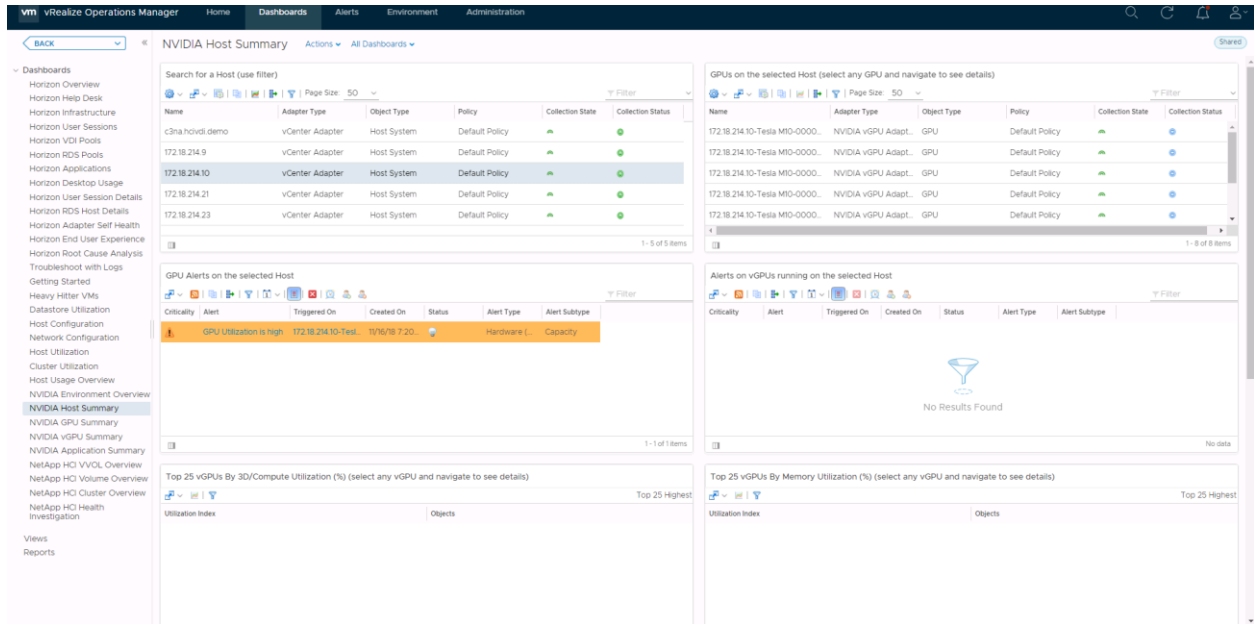
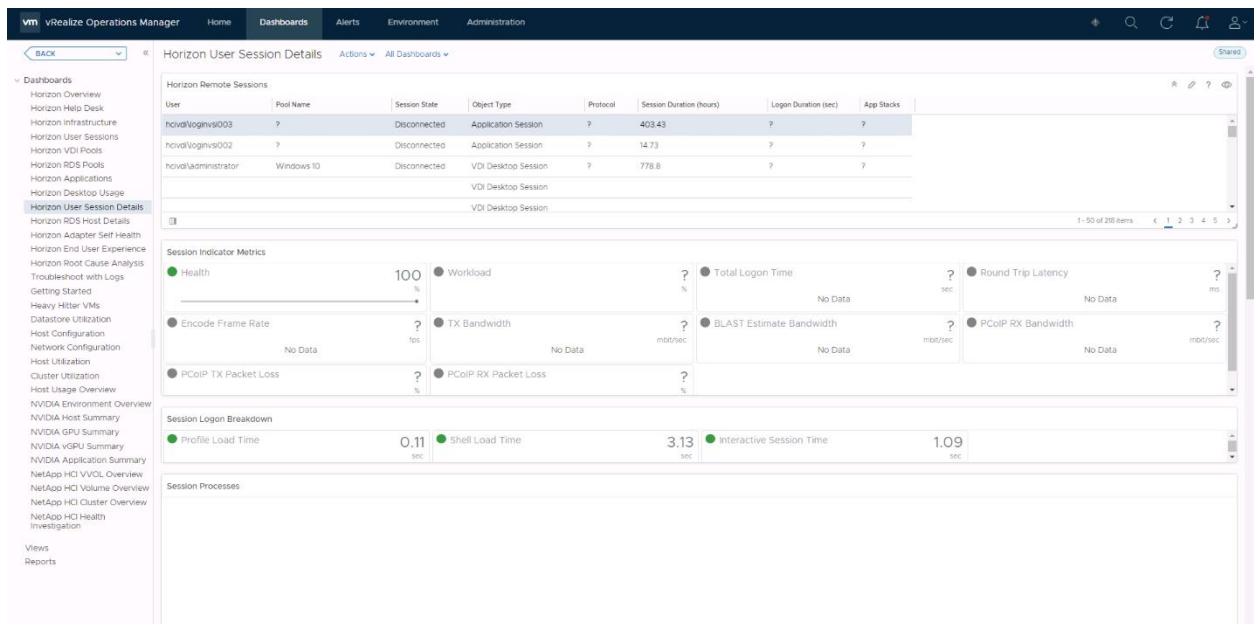


Figure 28) GPU metrics.



The vRealize for Horizon adapter is required to monitor the virtual desktops, application pools, App Volumes, and so on. vRealize Operations Manager Agent must be installed on the VMware Horizon connection servers, and agents must be installed on virtual desktop templates.

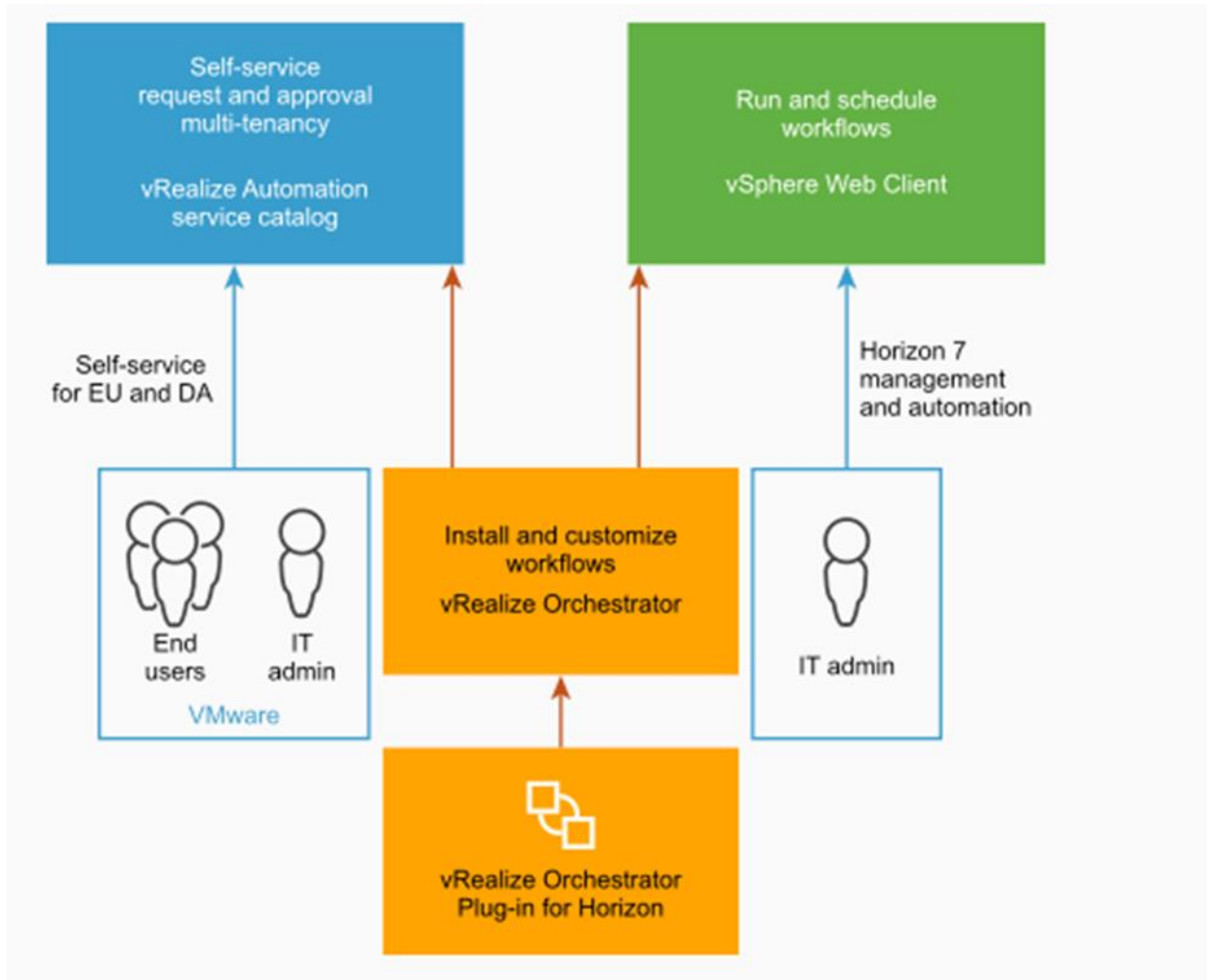
For more information, see the [vRealize Operations Manager sizing guidelines](#).

4.16 VMware vRealize Orchestrator Plug-In for Horizon

VMware Horizon provides vRealize Orchestrator Plug-In to extend vCenter or vRealize Automation with custom workflows. This plug-in allows users to choose which tool to deploy to manage the desktop and

application pools. With Horizon Agent Direct Connect, users can connect to desktops directly without going through the connection broker. VMware Horizon also supports HTML Direct Connection. To identify which features are provided and which features can be created with custom code, see the [plug-in documentation](#).

Figure 29) Architecture of vRealize Orchestrator Plug-In for Horizon.



4.17 VMware vRealize Log Insight

VMware vRealize Log Insight delivers heterogeneous and highly scalable log management with intuitive, actionable dashboards, sophisticated analytics, and broad third-party extensibility, providing deep operational visibility and faster troubleshooting.

To collect logs from desktops, you need to install vRealize Log Insight Agent on the desktops. For hosts such as vSphere Hypervisor or NetApp HCI storage, you can configure it to forward the logs to vRealize Log Insight as a syslog server.

Figure 30) vRealize Log Insight architecture.

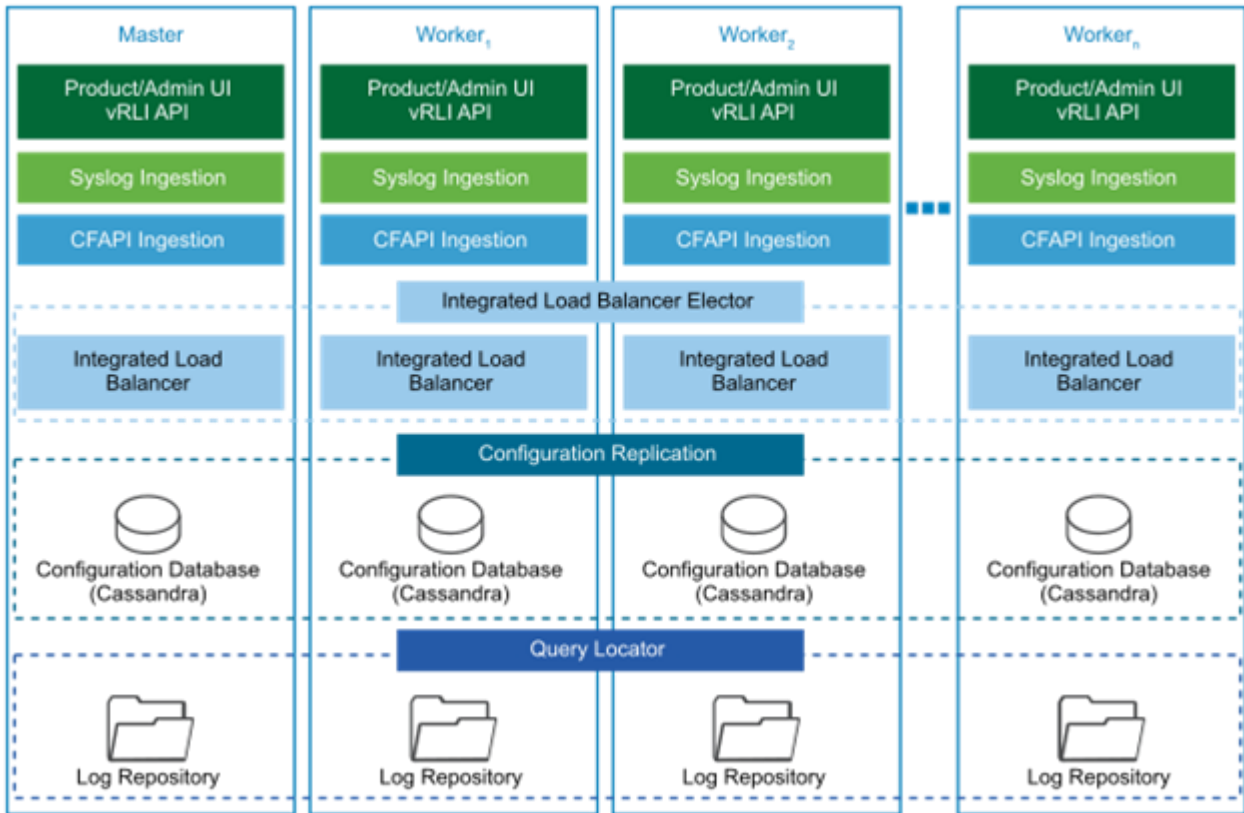
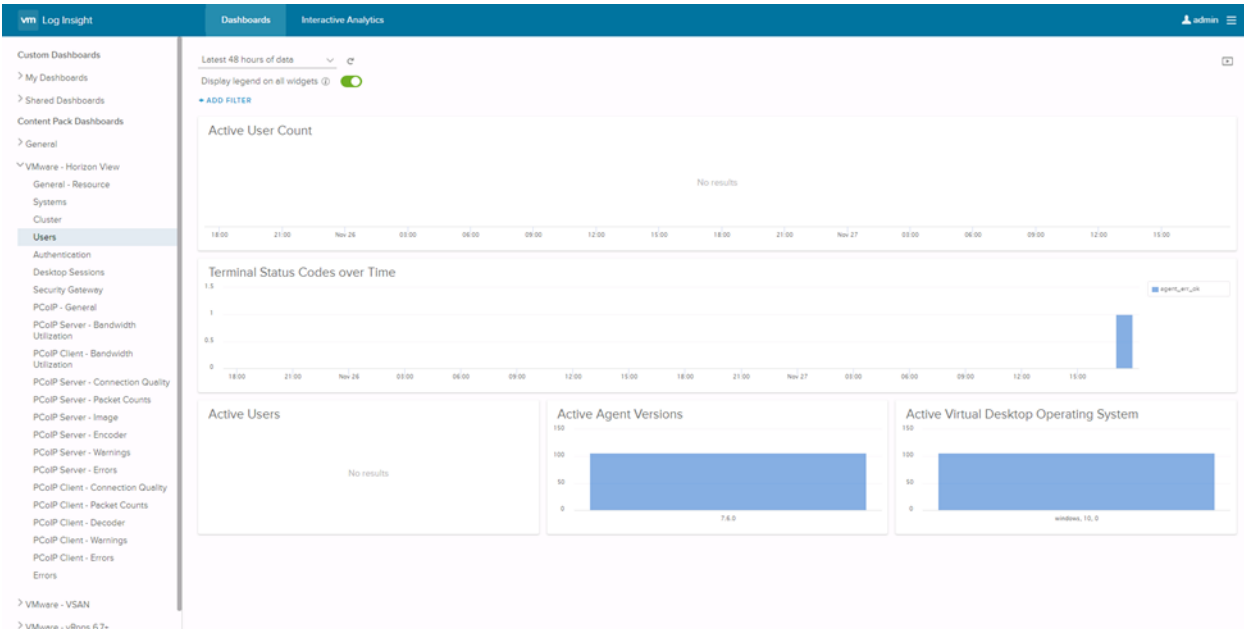


Figure 25) vRealize Log Insight UI.



For sizing information, see [Sizing the vRealize Log Insight Virtual Appliance](#).

4.18 VMware Workstation

VMware Workstation is a desktop hypervisor tool that can run on Windows and Linux desktops or laptops to create VMs running multiple Windows or Linux instances at the same time. When used with the optional product Horizon Flex, a virtual desktop from the desktop pool can be used in offline or disconnected mode.

4.19 VMware Fusion

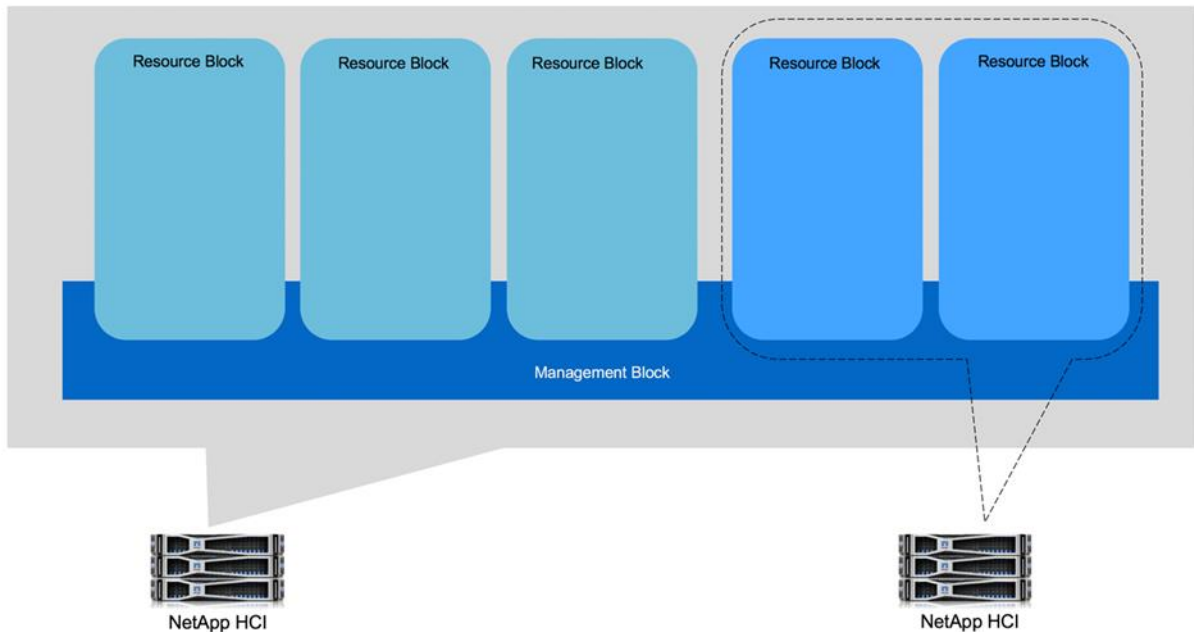
VMware Fusion allows Apple macOS machines to create Windows, Linux, or macOS VMs. When used with the optional product Horizon Flex, a virtual desktop from the desktop pool can be used in offline or disconnected mode.

5 Solution Design

As Figure 4 shows, the End-User Computing on NetApp HCI solution keeps infrastructure VMs separate from workload VMs in the management block and resource block. Having a resource block keeps the workloads securely contained and isolated. With NetApp HCI, each resource block corresponds to an account and a separate VLAN for storage traffic. Every block requires at least four compute nodes when used with NSX. Keep the three controllers on separate hosts and keep an additional one for fault tolerance.

In linked mode, 15 vCenter Servers are allowed. Therefore, 14 resource blocks and one management block are possible for each Horizon pod. The number of compute nodes in vCenter linked mode is 5,000. If you exceed the maximum storage space available in each cluster, you can deploy more clusters for the resource blocks and manage them centrally from the vCenter plug-in.

Figure 31) Resource block scalability.



Based on VMware guidelines, keeping 10,000 sessions for each Horizon pod and 2,000 sessions for each resource block provides five resource blocks for each Horizon pod. If you require more sessions, use the CPA feature to create global entitlement for users across the Horizon pods within the same site, across the sites, or to the cloud.

We have seen user density for NetApp HCI GPU compute nodes in the range of 130 to 150 virtual desktops per node for vSGA mode. To be conservative, we picked 128 users per node. That number is also valid if we use vGPU mode.

The configuration of desktop VMs is described in Table 3.

Table 3) Desktop VM configuration.

Item	Value
vCPU	2
RAM	4GB
NIC (VMXNET3 adapter)	1
VMDK	<ul style="list-style-type: none"> • 70GB (OS) • 10GB (UWV + profile) • 20GB (AppStack)

AppStacks support only Windows OS. Supported Linux VMs with Horizon can use hosted applications on Windows RDSH served by AppStacks.

An Instant Clone provides a delta disk for every virtual desktop. The size of the delta disk depends on the change rate. We assumed a 10% change rate for the sizing calculations, and we also assumed no backup volumes configured on App Volumes and considered 500GB for AppStacks. Table 4 provides a rough estimate for the number of datastores required for 1.5TB volumes.

Table 4) Datastore estimates.

Component	800 Users	1,200 Users	2,000 Users
NetApp HCI compute nodes (resource block)	8	11	17
NetApp HCI compute nodes (management block)	4	4	4
NetApp HCI storage nodes (based on the model used)	> or = 4	> or = 4	> or = 4
1.5TB datastore for hosting virtual desktops	4	6	10
1.5TB datastore for user writable volumes	8	12	20

The datastores that hold AppStack volumes are mostly for read-only access, and the datastores that hold user-writable volumes provide read/write access. With NetApp HCI, there is no need to separate those workloads, and both can coexist on the same datastore.

Hosted applications serve task-oriented users such as hospital workers who require access to specific applications and don't have time to wait for desktops to start up. The application pool is hosted on a farm of RDSHs. If the RDSH reboots during application access, the user gets an error. When the application is relaunched, it is served from the other surviving hosts.

For desktop pools serving Instant Clones, Horizon refreshes the desktop on another host if one of the vSphere hosts crashes. Typically, the desktop is protected by vSphere HA.

The UAG, App Volumes Manager, and connection servers are accessed through a load balancer, and the traffic is redirected to the server in service. AppStack volumes are replicated to multiple datastores by using the storage group, and writable volumes are distributed based on the following two policy sets:

- **Spread.** Distributes files evenly across all the storage locations. When a file is created, the storage with the most available space is selected.
- **Round-robin.** Distributes files by sequentially using the storage locations. When a file is created, the storage with the oldest used time is selected.

The App Volumes datastores that hold the virtual disks for AppStacks and user-writable volumes can be replicated from NetApp HCI to any ONTAP devices on premises or in the cloud. If the User Environment Manager file shares are also hosted on NetApp ONTAP, they too can be replicated to other ONTAP devices on premises or in the cloud. This configuration enables users to burst into the cloud for a short duration and then to bring data back on premises.

6 Technology Components

This section covers the technology components that we used for validating the End-User Computing on NetApp HCI solution.

6.1 Hardware Components

Table 5 lists the hardware components that we used for the solution. These components might vary according to customer requirements.

Table 5) Hardware components.

Component	Count	Description
NetApp HCI compute nodes (resource block)	2	H610C
NetApp HCI compute nodes (management block)	2	H500E
NetApp HCI storage nodes	4	H500S
Cisco Nexus 9000 switch	2	iSCSI/VM/vMotion
1Gb switch	1	IPMI/Management

6.2 Software Components

Table 6) lists the software components that are required to implement the solution. These components might vary according to customer requirements.

Table 6) Software requirements.

Software	Version
NetApp HCI NDE	1.4
VMware Horizon 7 Enterprise	
VMware Identity Manager	3.3.0
VMware Horizon	7.6
VMware App Volumes	2.14.2

Software	Version
VMware NSX for vSphere	6.4.3
VMware vCenter	6.7 Update 1
VMware vSphere	6.7 Update 1
VMware vRealize Operations Manager	6.6.1
VMware vRealize Operations for Published Applications	6.5.1
VMware vRealize Orchestrator Plug-in for Horizon	1.4.0
VMware User Environment Manager	9.5.0
VMware vRealize Log Insight for vCenter	4.6.1
NVIDIA GRID	7.0

7 Solution Verification

This solution architecture is based on VMware vSphere 6.7 Update 1, VMware Horizon 7.6, and NVIDIA GRID 7.0 so that it can support vMotion for the VM running graphics workloads, including vGPU mode. In our testing, our management block was hosted on NetApp HCI 500E/410C nodes, and our resource block was hosted on NetApp HCI H610C nodes. We used Cisco Nexus 9K switches for the 10Gb network, and we used a 1Gb switch for the management network.

We verified that we were able to deploy desktop pools and application pools utilizing the GPU with VMware Horizon. We performed on-demand desktop assignment and application provisioning for a test user, and we confirmed that their settings were preserved with logout and login.

We determined that we could put a vSphere host in maintenance mode, and we migrated a VM running a graphics workload with a vGPU profile to the other host. We were able to perform session sharing while running multiple workloads with Horizon clients and by using HTML with a web browser.

We validated the solution by simulating user-like workloads using the Login Virtual Session Indexer (Login VSI). Login VSI is the industry-standard load-testing tool for testing the performance and scalability of centralized Windows desktop environments, such as server-based computing and VDI.

The testing included the following scenarios:

- Get VSIMax for a single server with no GPU used for a knowledge worker.
- Get VSIMax for a single server with vSGA mode for a knowledge worker.
- Get VSIMax for a single server with vGPU mode for a knowledge worker.
- Get VSIMax for a single server with vSGA mode for a multimedia workload.
- Get VSIMax for a single server with vGPU mode for a multimedia workload.

All the above testing was performed using JMP to entitle the desktop pool, an AppStack volume for Login VSI workloads, and the User Environment Manager policies.

8 Conclusion

NetApp HCI can scale or shrink based on your business needs so that you can provide flexible options for your customers. The QoS feature makes NetApp HCI simple to integrate with existing workloads. NetApp HCI GPU compute nodes provide an enhanced user experience for knowledge-based workers who can

benefit from hardware 3D acceleration. With Horizon 7, users can securely connect to desktop pools or application pools from any device, anywhere.

Where to Find Additional Information

To learn more about the information described in this document, refer to the following documents and websites:

NetApp

- NetApp Product Documentation
<https://www.netapp.com/us/documentation/index.aspx>
- NetApp HCI Documentation Center
<http://docs.netapp.com/hci/index.jsp>
- NetApp Data Fabric
<https://www.netapp.com/us/info/what-is-data-fabric.aspx>
- NetApp HCI Datasheet
<https://www.netapp.com/us/media/ds-3881.pdf>
- NetApp HCI Technical docs
<http://docs.netapp.com/hci/index.jsp>
- NetApp HCI Deployment Guide
https://library.netapp.com/ecm/ecm_download_file/ECMLP2844053
- NetApp HCI Network Setup Guide
<https://www.netapp.com/us/media/tr-4679.pdf>
- VMware vRealize Operations Management Pack for NetApp HCI and SolidFire
<https://bluemedora.com/resource/vmware-vrealize-operations-management-pack-for-netapp-hci-solidfire/>
- NetApp SolidFire® vRealize Orchestrator Plug-in
<https://github.com/solidfire/vrealize-orchestrator-plugin>
- NetApp HCI Theory of Operations
<https://www.netapp.com/us/media/wp-7261.pdf>
- NetApp Element software
<https://www.netapp.com/us/products/data-management-software/element-os.aspx>
- NetApp ONTAP Select
<https://www.netapp.com/us/products/data-management-software/ontap-select-sds.aspx>
- NetApp Interoperability Matrix Tool
<https://mysupport.netapp.com/matrix/#welcome>

NVIDIA

- NVIDIA Virtual GPU Software Documentation
<https://docs.nvidia.com/grid/>

VMware

- VMware Tech Zone
<https://techzone.vmware.com/>
- VMware Workspace ONE and VMware Horizon 7 Enterprise Edition On-Premises Reference Architecture
<https://techzone.vmware.com/resource/vmware-workspace-one-and-vmware-horizon-7-enterprise-edition-premises-reference>

- Deploying Hardware-Accelerated Graphics with VMware Horizon 7
<https://techzone.vmware.com/resource/deploying-hardware-accelerated-graphics-vmware-horizon-7>
- VMware Horizon 7 Sizing Limits and Recommendations
<https://kb.vmware.com/s/article/2150348>
- VMware Configuration Maximums
<https://configmax.vmware.com/>
- Reviewer's Guide for On-Premises VMware Identity Manager
<https://techzone.vmware.com/resource/reviewers-guide-premises-vmware-identity-manager>
- Deployment and Design Considerations for VMware Mirage
<https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/vmware-horizon-mirage-deployment-design-considerations.pdf>
- Network Ports in VMware Horizon 7
<https://techzone.vmware.com/resource/network-ports-vmware-horizon-7>
- Horizon for Linux FAQ
<https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/products/horizon/vmware-horizon-for-linux-faq.pdf>
- vRealize Operations Manager 6.6.1 Sizing Guidelines
<https://kb.vmware.com/s/article/2150421>
- Sizing the vRealize Log Insight Virtual Appliance
<https://docs.vmware.com/en/vRealize-Log-Insight/4.6/com.vmware.log-insight.getting-started.doc/GUID-284FC5F4-B832-47A7-912E-D407A760CAE4.html>

Version History

Version	Date	Document Version History
Version 1.0	January 2019	Initial version

Refer to the [Interoperability Matrix Tool \(IMT\)](#) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

Copyright Information

Copyright © 2018–2019 NetApp, Inc. All rights reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP "AS IS" AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

RESTRICTED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (c)(1)(ii) of the Rights in Technical Data and Computer Software clause at DFARS 252.277-7103 (October 1988) and FAR 52-227-19 (June 1987).

Trademark Information

NETAPP, the NETAPP logo, and the marks listed at <http://www.netapp.com/TM> are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.