Dell Precision Appliance for Wyse - VMware Horizon

Dell Engineering
April 2017

A Dell Appliance Architecture

Revisions

Date	Description
July 2016	NVIDIA® GRID vGPU™ with Dell PowerEdge R730 platform and NVIDIA® Tesla® M60 GPU cards.
Sep 2016	Minor updates to Endpoints section.
Dec 2016	Removed NVIDIA® GRID K2 and added NVIDIA® Tesla® M10 GPU information.
April 2017	Added test results to document

Contents

1 Introduction

1.1 Purpose

This document addresses the architecture design, configuration and implementation considerations for the key components required to deliver graphics-enabled virtual desktops using the Dell Precision Appliance for Wyse. The underlying technology is VMware Horizon® on VMware vSphere® 6 with graphics hardware acceleration provided by NVIDIA® graphics processing units (GPUs) using NVIDIA GRID vGPU™ software.

NOTE: The appliance is a compute-only resource intended to be used with an existing or new deployment of Horizon View. For complete details on designing and deploying your entire Horizon View environment, please refer to the **Dell Wyse Datacenter for VMware Horizon** reference architecture located here: LINK

1.2 Scope

Relative to delivering graphics-enabled virtual desktops, the objectives of this document are to:

- Define the detailed technical design for the appliance.
- Define the hardware requirements to support the design.
- Define the constraints which are relevant to the design.
- Define relevant risks, issues, assumptions and concessions – referencing existing ones where possible.
- Provide a breakdown of the design into key elements such that the reader receives an incremental or modular explanation of the design.
- Provide component selection guidance where applicable.

1.3 What's New

- Incorporated test results into document.

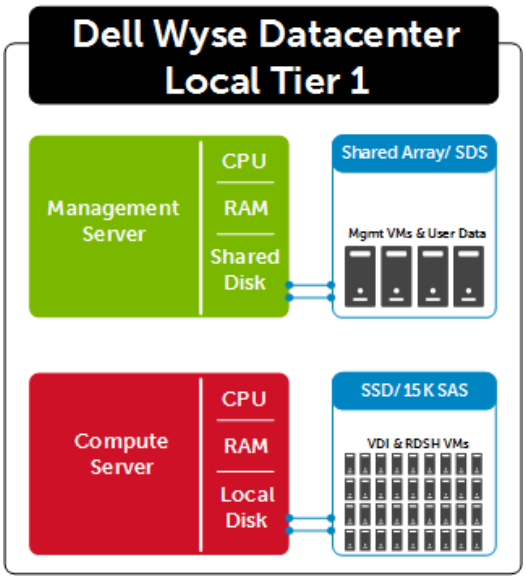
2 Solution Architecture Overview

2.1 Introduction

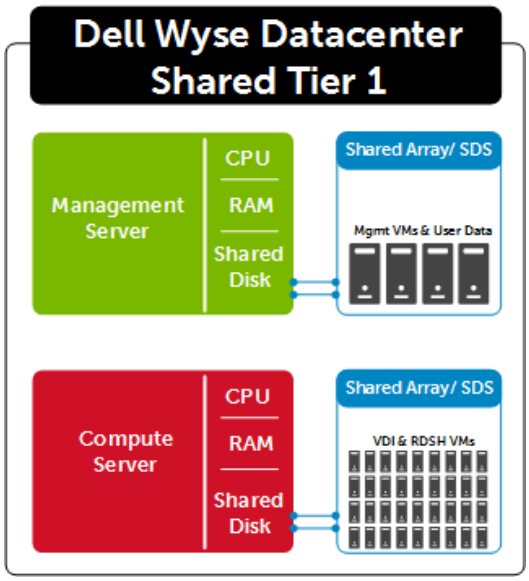
The Dell Precision Appliance for Wyse delivers graphics-enabled virtual desktops with GPU hardware acceleration for unparalleled graphics performance. The appliance is a “building block” for the Dell Wyse Datacenter Solution portfolio and is a PowerEdge R730 server that can be used as a graphics compute resource in a number of VDI deployments. With the combination of powerful processors, large memory, versatile storage options, and NVIDIA® GPU cards, the R730 performs exceptionally well in demanding environments and delivers outstanding functionality in just 2U of rack space.

2.2 Physical Architecture Overview

The core Dell Wyse Datacenter architecture consists of two models: Local Tier1 and Shared Tier1. “Tier 1” in the Dell Wyse Datacenter context defines from which disk source the VDI sessions execute. Local Tier1 includes rack servers or blades with SSDs/HDDs while Shared Tier 1 can include rack or blade servers due to the usage of shared Tier 1 storage. Tier 2 storage is present in both solution architectures and, while having a reduced performance requirement, is utilized for user data and Management VM execution. Management VM execution occurs using Tier 2 storage for all solution models. Dell Wyse Datacenter is a 100% virtualized solution architecture.



In the Shared Tier 1 solution model, an additional high-performance shared storage array is added to handle the execution of the VDI sessions. Neither management VMs nor compute VMs are stored to local disks in this model.



NOTE: The appliance can be configured as a compute host using the Local or Shared Tier 1 model.

2.3 Solution Layers

The Dell Wyse Datacenter Solution leverages a core set of hardware and software components consisting of five primary layers:

- Compute Server Layer
- Management Server Layer
- Networking Layer
- Storage Layer
- Thin Client Layer (please refer to the Dell Wyse Thin Clients section)

These components have been integrated and tested to provide the optimal balance of high performance and lowest cost per user. The Dell Wyse Datacenter stack is designed to be cost effective allowing IT departments to implement high-performance fully virtualized desktop environments.

NOTE: The appliance functions at the Compute Server layer. The following section describes how it is configured and integrated with the other components of a Dell Wyse Datacenter solution. For details on the other solution layers, please refer to the **Dell Wyse Datacenter for VMware Horizon** reference architecture located here: LINK

2.4 Appliance Configurations

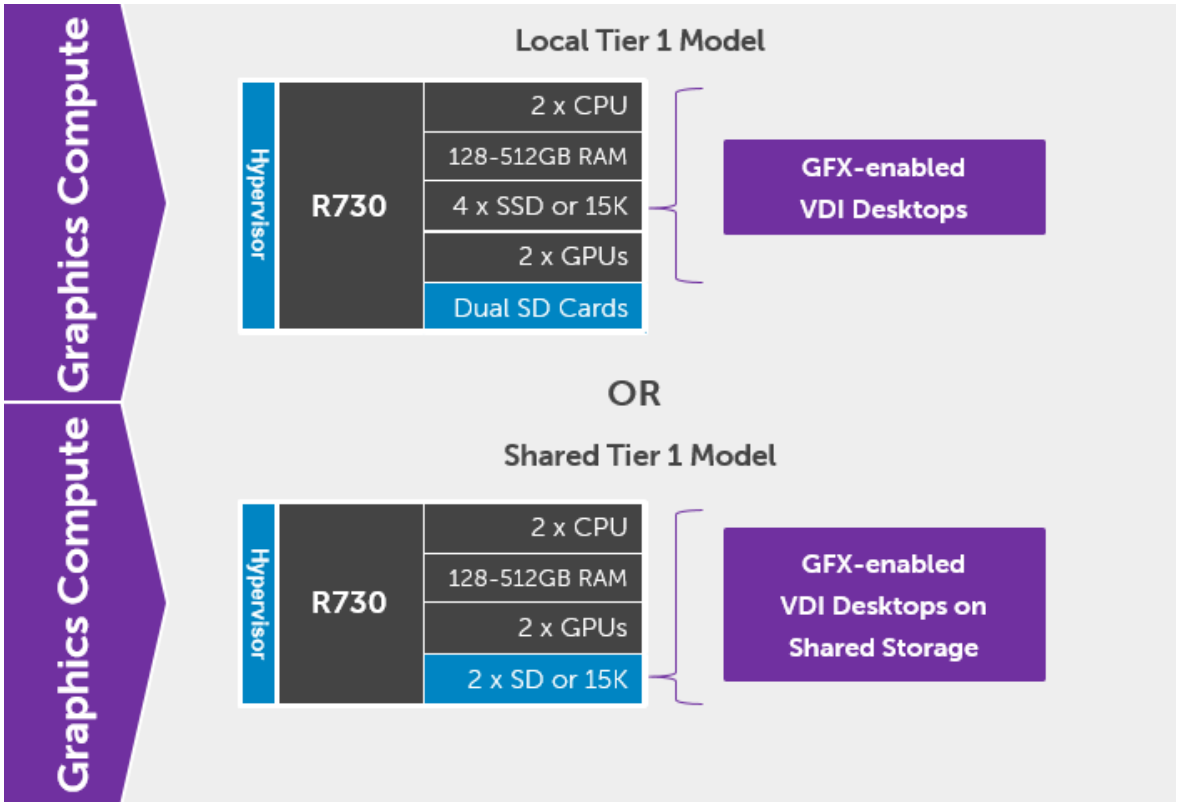
The Dell Precision Appliance for Wyse is a compute-only resource using vSphere hypervisor and providing graphics acceleration for VDI desktops. The appliance does not host management VMs and must be added to the compute layer of an existing or new deployment of Horizon View. Appliances are sold with preconfigured processors, memory, disks, and GPU cards to achieve either higher graphics VM density or performance as validated by Dell engineering (see Hardware Components for details). These components can also be modified to meet specific customer needs. Appliances can be configured to follow the Local Tier 1 or Shared Tier 1 disk model as described in the Physical Architecture Overview section:

Local Tier 1

- Hypervisor is installed to dual SD flash media (mirrored) only.
- VDI sessions reside on and are executed from datastore(s) created from local SSD or spinning disks.

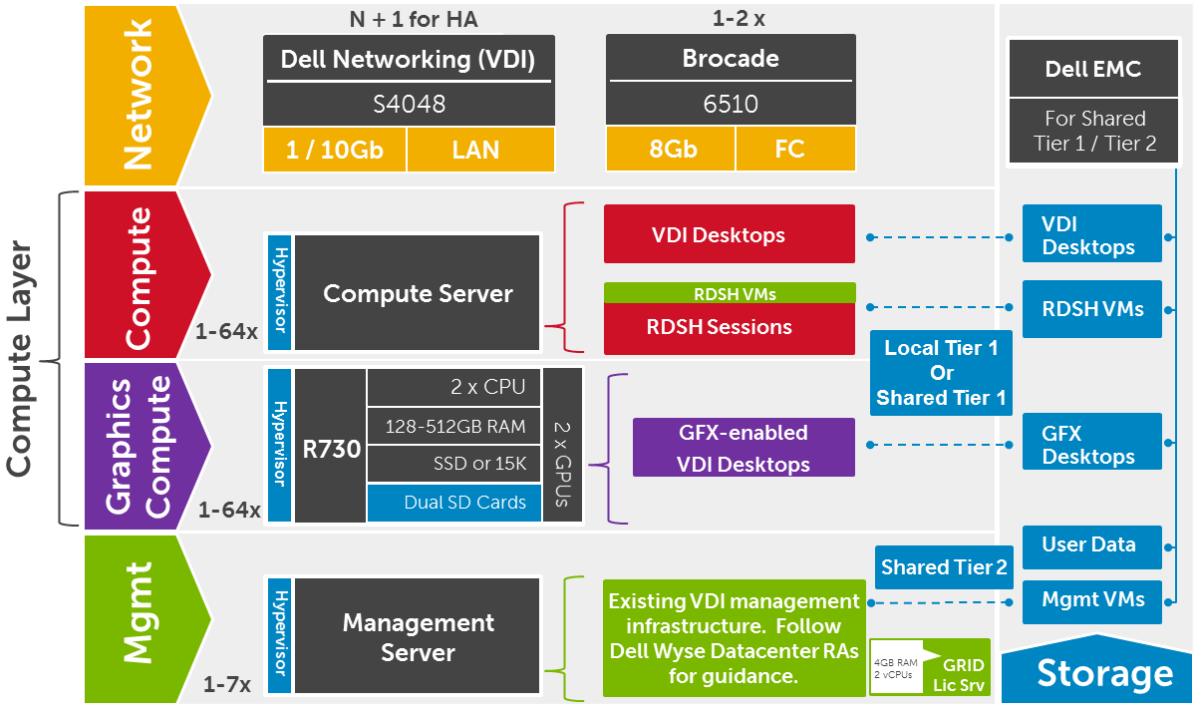
Shared Tier 1

- Hypervisor is installed to dual SD flash media (mirrored) or to local disk if preferred; however, Dell only factory installs vSphere to internal SD media so customers will have to perform installation to local disk if they choose that option.
- VDI sessions reside on and are executed from datastore(s) on shared storage.



Designed to be compute nodes with graphics acceleration, the appliances fit in perfectly with the Local Tier 1 for Rack and Shared Tier 1 for Rack models of the Dell Wyse Datacenter solutions offerings. In the image

below, **Compute** indicates a non-graphics compute resource capable of executing non-graphics desktops or RDSH VMs while **Graphics Compute** indicates a graphics compute resource (Dell Precision Appliance for Wyse) capable of executing graphics-enabled desktops. Although they both function at the Compute layer, they are distinct and not interchangeable. If using clustering, appliances should only be added to clusters with other like-configured appliances. As noted below, up to 64 appliances can be added to each cluster with vSphere 6.



NOTE: For complete details on designing and deploying your entire Horizon View environment including cabling, network architecture, and shared storage scaling guidance, please refer to the **Dell Wyse Datacenter for VMware Horizon** reference architecture located here: LINK

3 Hardware Components

3.1 Dell Precision Appliance for Wyse

The Dell Precision Appliance for Wyse is offered in the PowerEdge R730 (rack) server platform. Appliances can be ordered in the default **High Density** or **High Performance** configurations shown below or customized with different hardware components as needed. In general, the appliances are equipped to deliver high-performance virtualized graphics; however, differences between the default configurations are summarized below.

High Density

- Utilizes processors that have a relatively lower clock speed with a higher core count
- Best suited for higher number of users (up to 128 per appliance depending on GPU card) that don't require large amounts of graphics memory per user (up to 4GB frame buffer)
- In general, multi-threaded applications will benefit from a higher core count

High Performance

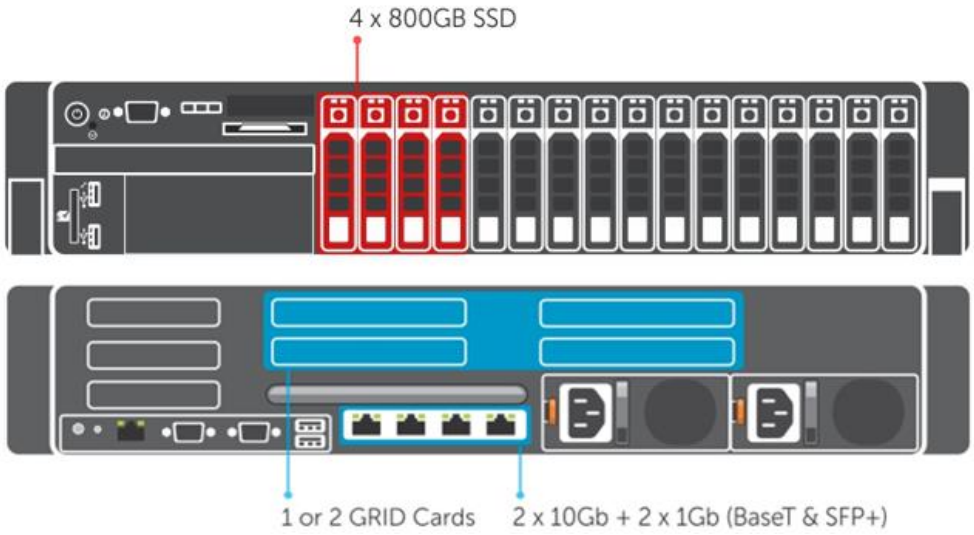
- Utilizes processors that have a relatively faster clock speed with a lower core count
- For users that require largest amount of graphics memory available (8GB frame buffer) along with full graphics feature set including CUDA and OpenCL support (up to 4 users per appliance with M60-8Q profile)
- In general, single-threaded applications will benefit from a higher clock speed

R730	High Density	High Performance
CPU	2 x E5-2698v4 (20C, 2.2GHz)	2 x E5-2667v4 (8C, 3.2GHz)
Memory	**CHOICE** (128GB – 512GB)	**CHOICE** (128GB – 512GB)
GPU Boards	NVIDIA Tesla M10 or M60	NVIDIA Tesla M60
Graphics	vGPU	vGPU
Storage Ctrls	PERC H730P (2GB cache)	PERC H730P (2GB cache)
Storage	2 x 8GB SD Cards **(ESXi Hypervisor)** SSD or 15K **Choice** w/ Default: 4 x 800GB SSD	2 x 8GB SD Cards **(ESXi Hypervisor)** SSD or 15K **Choice** w/ Default: 4 x 800GB SSD
Network	**Choice** w/ Default: 2 x 10Gb (SFP+/ BT) & 2 x 1Gb NDS	**Choice** w/ Default: 2 x 10Gb (SFP+/ BT) & 2 x 1Gb NDS
iDRAC	iDRAC8 Ent w/ integrated Dell Remote Access Controller	iDRAC8 Ent w/ integrated Dell Remote Access Controller
Power	2 x 1100W PSUs	2 x 1100W PSUs

3.1.1 PowerEdge R730

The best-in-class Dell PowerEdge R730 is not only the foundation of the Dell Wyse Datacenter solution portfolio, but also is the 2U platform configured with NVIDIA GPU cards that creates the Dell Precision Appliance. This dual socket CPU platform runs the fastest Intel Xeon E5-2600 v4 family of processors, can host up to 24 DIMMs of DDR4 RAM, supports up to 16 x 2.5" SAS disks, SFP+ or BaseT NICs, and can be outfitted with up to 2 double-wide GPU cards delivering up to 64 graphics accelerated users per node depending on configuration.

For more information on the R730, please visit: Link



NOTE: The default disk configuration is shown in the diagram here; however, this can be modified to match the customer's local storage needs.

3.2 GPUs

3.2.1 NVIDIA Tesla GPUs

Accelerate your most demanding enterprise data center workloads with NVIDIA® Tesla® GPU accelerators. Scientists can now crunch through petabytes of data up to 10x faster than with CPUs in applications ranging from energy exploration to deep learning. Plus, Tesla accelerators deliver the horsepower needed to run bigger simulations faster than ever before. For enterprises deploying VDI, Tesla accelerators are perfect for accelerating virtual desktops.

3.2.1.1 NVIDIA Tesla M10

The NVIDIA® Tesla® M10 is a dual-slot 10.5 inch PCI Express Gen3 graphics card featuring four mid-range NVIDIA Maxwell™ GPUs and a total of 32GB GDDR5 memory per card (8GB per GPU). The Tesla® M10 doubles the number of H.264 encoders over the NVIDIA® Kepler™ GPUs and improves encoding quality, which enables richer colors, preserves more details after video encoding, and results in a high-quality user experience.

The NVIDIA® Tesla® M10 GPU accelerator works with NVIDIA GRID™ software to deliver the industry's highest user density for virtualized desktops and applications. It supports up to 64 desktops per GPU card (up to 128 desktops per server) and gives businesses the power to deliver great graphics experiences to all of their employees at an affordable cost.

Specs	Tesla M10
Number of GPUs	4 x NVIDIA Maxwell™ GPUs
Total CUDA cores	2560 (640 per GPU)
GPU Clock	Idle: 405MHz / Base: 1033MHz
Total memory size	32GB GDDR5 (8GB per GPU)
Max power	225W
Form Factors	Dual slot (4.4" x 10.5")
Aux power	8-pin connector
PCIe	x16 (Gen3)
Cooling solution	Passive

3.2.1.2 NVIDIA Tesla M60

The NVIDIA® Tesla® M60 is a dual-slot 10.5 inch PCI Express Gen3 graphics card featuring two high-end NVIDIA Maxwell™ GPUs and a total of 16GB GDDR5 memory per card. This card utilizes NVIDIA GPU Boost™ technology which dynamically adjusts the GPU clock to achieve maximum performance. Additionally, the Tesla® M60 doubles the number of H.264 encoders over the NVIDIA® Kepler™ GPUs.

The NVIDIA® Tesla® M60 GPU accelerator works with NVIDIA GRID™ software to provide the industry's highest user performance for virtualized workstations, desktops,

and applications. It allows enterprises to virtualize almost any application (including professional graphics applications) and deliver them to any device, anywhere.

Specs	Tesla M60
Number of GPUs	2 x NVIDIA Maxwell™ GPUs
Total CUDA cores	4096 (2048 per GPU)
Base Clock	899 MHz (Max: 1178 MHz)
Total memory size	16GB GDDR5 (8GB per GPU)
Max power	300W
Form Factors	Dual slot (4.4” x 10.5”)
Aux power	8-pin connector
PCIe	x16 (Gen3)
Cooling solution	Passive/ Active

3.3 Network

The Dell Precision Appliance for Wyse requires a minimum 1Gb connectivity for graphics-enabled VDI network traffic; however, Dell recommends 10Gb networking for best performance. For network switch recommendations when designing and deploying your Horizon View environment, please refer to the **Dell Wyse Datacenter for VMware Horizon** reference architecture located here: LINK

3.4 Dell Wyse Endpoints

The following Dell Wyse clients will deliver outstanding graphics performance and are the recommended choices for high end graphics use cases.

For a complete list of recommended Dell Wyse clients for a superior VMware Horizon View experience, please refer to the **Dell Wyse Datacenter for VMware Horizon** reference architecture located here: LINK

3.4.1 Wyse 5030 PCoIP Zero Client

Uncompromising computing with the benefits of secure, centralized management. The Wyse 5030 PCoIP zero client for VMware Horizon and Amazon WorkSpaces is a secure, easily managed zero client that provides outstanding graphics performance for advanced applications such as CAD, 3D solids modeling, video editing and advanced worker-level office productivity applications. Smaller than a typical notebook, this dedicated zero client is designed specifically for VMware Horizon and Amazon WorkSpaces. It features the latest processor technology from Teradici to process the PCoIP protocol in silicon and includes client-side content caching to deliver the highest level of performance available over 2 HD displays in an extremely compact, energy-efficient form factor. The Wyse 5030 delivers a rich user experience while resolving the challenges of provisioning, managing, maintaining and securing enterprise desktops. For more information, please visit: Link

3.4.2 Wyse 5050 AIO Zero Client

The Wyse 5050 All-in-One (AIO) PCoIP zero client combines the security and performance of the Wyse 5030 PCoIP zero client for VMware with the elegant design of Dell's best-selling P24 LED monitor. The Wyse 5050 AIO provides a best-in-class virtual experience with superior manageability – at a better value than purchasing a zero client and high resolution monitor separately. A dedicated hardware PCoIP engine delivers the highest level of display performance available for advanced applications, including CAD, 3D solids modeling, video editing and more. Elegant in appearance and energy efficient, the Wyse 5050 AIO is a fully functional VMware Horizon endpoint that delivers a true PC-like experience. It offers the full benefits of an efficient and secure centralized computing environment, like rich multimedia, high-resolution 3D graphics, HD media, and full USB peripheral interoperability locally (LAN) or remotely (WAN). For more information, please visit: Link

3.4.3 Wyse 7030 PCoIP Zero Client

Uncompromising computing with the benefits of secure, centralized management. The Wyse 7030 PCoIP zero client for VMware Horizon and Amazon WorkSpaces is a secure, easily managed zero client that provides outstanding graphics performance for advanced applications such as CAD, 3D solids modeling, video editing and advanced worker-level office productivity applications. About the size of a notebook, this dedicated zero client designed specifically for VMware Horizon and Amazon WorkSpaces. It features the latest processor technology from Teradici to process the PCoIP protocol in silicon and includes client-side content caching to deliver the highest level of display performance available over 4 HD displays in a compact, energy-efficient form factor. The Dell Wyse 7030 delivers a rich user experience while resolving the challenges of provisioning, managing, maintaining and securing enterprise desktops. For more information, please visit: Link

3.4.4 Wyse 7040 Thin Client with WES7P

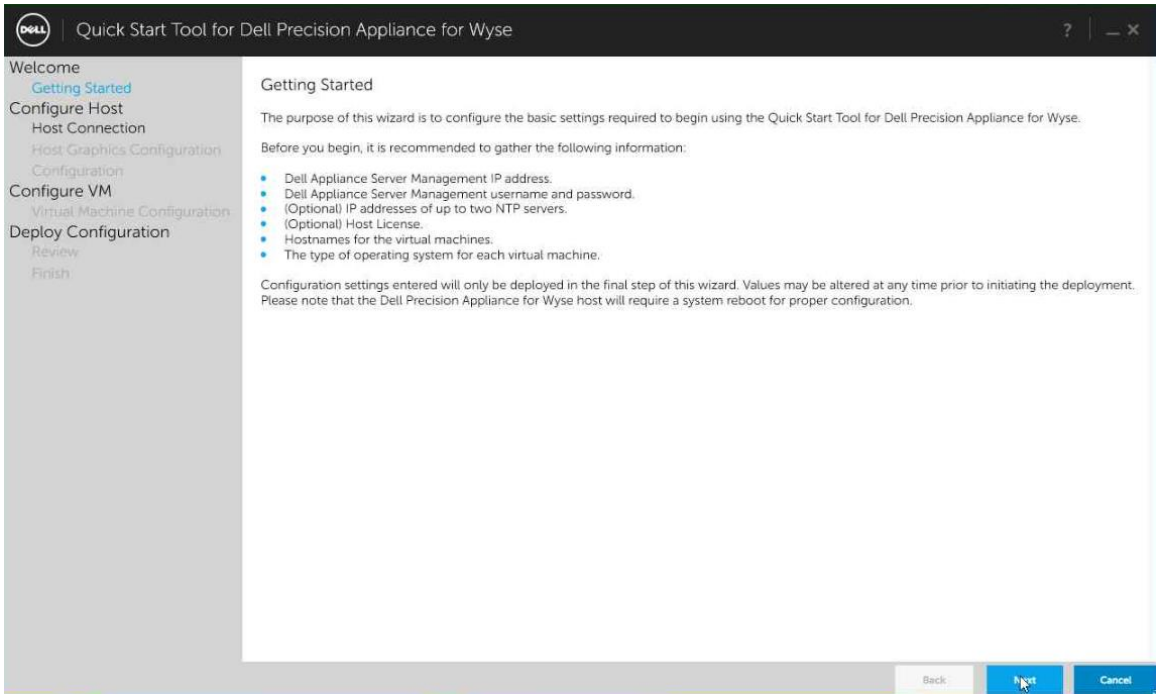
The Wyse 7040 is a high-powered, ultra-secure thin client. Equipped with 6th generation Intel i5/i7 processors, it delivers extremely high graphical display performance (up to three displays via display-port daisy-chaining, with 4K resolution available on a single monitor) for seamless access to the most demanding applications. The Wyse 7040 is compatible with both data center hosted and client-side virtual desktop environments and is compliant with all relevant U.S. Federal security certifications including OPAL compliant hard-drive options, VPAT/Section 508, NIST BIOS, Energy-Star and EPEAT. Wyse enhanced Windows Embedded Standard 7P OS provides additional security features such as BitLocker. The Wyse 7040 offers a high level of connectivity including dual NIC, 6 x USB3.0 ports and an optional second network port, with either copper or fiber SFP interface. Wyse 7040 devices are highly manageable through Intel vPRO, Wyse Device Manager (WDM), Microsoft System Center Configuration Manager (SCCM) and Dell Command Configure (DCC). For more information, please visit: Link

4 Software Components

4.1 Dell Quick Start Tool (QST)

The Quick Start Tool for the Dell Precision Appliance for Wyse is designed to be a “use once” utility that will install the graphics driver to the appliance and prepare it for graphics-enabled virtual machines. Additionally, the tool will walk the user through preparing a specified number of virtual machines (or master images) for graphics acceleration by configuring the VM parameters and graphics profiles to be used. At this point, the user can take over and install the desktop OS and applications to be used in the virtual machine(s)/master image(s). In summary, the QST provides the following benefits:

- Ensures that the host and resulting virtual machines are configured correctly and consistently
- Automates a difficult and complex setup procedure by turning it in to just a few button clicks



NOTE: The QST for the Dell Precision Appliance for Wyse is a freely available, **unsupported** tool provided for convenience with no guarantees. It is not required to configure, setup, or use your Dell Precision Appliance for Wyse and is to be used at your own discretion and risk. The QST can be downloaded from here: LINK

NOTE: The QST will only work with ESXi 6.0 – 6.0U2 and only with NVIDIA Tesla M10 or M60 GPU cards.

4.2 VMware

NOTE: The appliance must have the VMware vSphere 6 (or higher) hypervisor installed on it but it will not host the VMware Horizon View components mentioned below. The appliance is intended to be used with an existing or new deployment of Horizon View. For complete details on designing and deploying your entire Horizon View environment, please refer to the **Dell Wyse Datacenter for VMware Horizon** reference architecture located here: LINK

4.2.1 VMware Horizon 7

The solution is based on VMware Horizon which provides a complete end-to-end solution delivering Microsoft Windows virtual desktops to users on a wide variety of endpoint devices. Virtual desktops are dynamically assembled on demand, providing users with pristine, yet personalized, desktops each time they log on.

VMware Horizon provides a complete virtual desktop delivery system by integrating several distributed components with advanced configuration tools that simplify the creation and real-time management of the virtual desktop infrastructure. For the complete set of details, please see the Horizon View resources page at http://www.vmware.com/products/horizon-view/resources.html

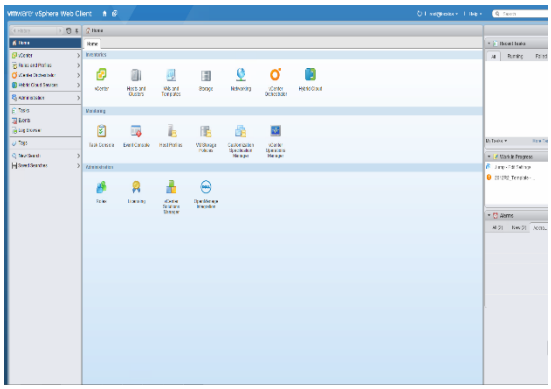
The core Horizon View components include:

- **View Connection Server (VCS) –** Installed on servers in the data center and brokers client connections, The VCS authenticates users, entitles users by mapping them to desktops and/or pools, establishes secure connections from clients to desktops, support single sign-on, sets and applies policies, acts as a DMZ security server for outside corporate firewall connections and more.
- **View Client –** Installed on endpoints. Is software for creating connections to View desktops that can be run from tablets, Windows, Linux, or Mac PCs or laptops, thin clients and other devices.
- **View Portal –** A web portal to access links for downloading full View clients. With HTML Access Feature enabled enablement for running a View desktop inside a supported browser is enabled.
- **View Agent –** Installed on all VMs, physical machines and Terminal Service servers that are used as a source for View desktops. On VMs the agent is used to communicate with the View client to provide services such as USB redirection, printer support and more.
- **View Administrator –** A web portal that provides admin functions such as deploy and management of View desktops and pools, set and control user authentication and more.
- **View Composer –** This software service can be installed standalone or on the vCenter server and provides enablement to deploy and create linked clone desktop pools (also called non-persistent desktops).
- **vCenter Server –** This is a server that provides centralized management and configuration to entire virtual desktop and host infrastructure. It facilitates configuration, provision, management services. It is installed on a Windows Server host (can be a VM).
- **View Transfer Server –** Manages data transfers between the data center and the View desktops that are checked out on the end users' desktops in offline mode. This Server is required to support

desktops that run the View client with Local Mode options. Replications and syncing are the functions it will perform with offline images.

4.3 Hypervisor Platforms

4.3.1 VMware vSphere 6



The vSphere hypervisor also known as ESXi is a bare-metal hypervisor that installs directly on top of your physical server and partitions it into multiple virtual machines. Each virtual machine shares the same physical resources as the other virtual machines and they can all run at the same time. Unlike other hypervisors, all management functionality of vSphere is done through remote management tools. There is no underlying operating system, reducing the install footprint to less than 150MB.

VMware vSphere 6 includes three major layers: Virtualization, Management and Interface. The Virtualization layer includes infrastructure and application services. The Management layer is central for configuring, provisioning and managing virtualized environments. The Interface layer includes the vSphere web client.

Throughout the Dell Wyse Datacenter solution, all VMware and Microsoft best practices and prerequisites for core services are adhered to (NTP, DNS, Active Directory, etc.). The vCenter 6 VM used in the solution is a single Windows Server 2012 R2 VM or vCenter 6 virtual appliance, residing on a host in the management layer. SQL server is a core component of the Windows version of vCenter and is hosted on another VM also residing in the management layer. It is recommended that all additional Horizon components be installed in a distributed architecture, one role per server VM.

For more information on VMware vSphere, visit http://www.vmware.com/products/vsphere.

4.4 NVIDIA GRID vGPU

NVIDIA GRID vGPU™ brings the full benefit of NVIDIA hardware-accelerated graphics to virtualized solutions. This technology provides exceptional graphics performance for virtual desktops equivalent to local PCs when sharing a GPU among multiple users.

GRID vGPU™ is the industry's most advanced technology for sharing true GPU hardware acceleration between multiple virtual desktops—without compromising the graphics experience. Application features and compatibility are exactly the same as they would be at the user's desk.

With GRID vGPU™ technology, the graphics commands of each virtual machine are passed directly to the GPU, without translation by the hypervisor. This allows the GPU hardware to be time-sliced to deliver the ultimate in shared virtualized graphics performance.

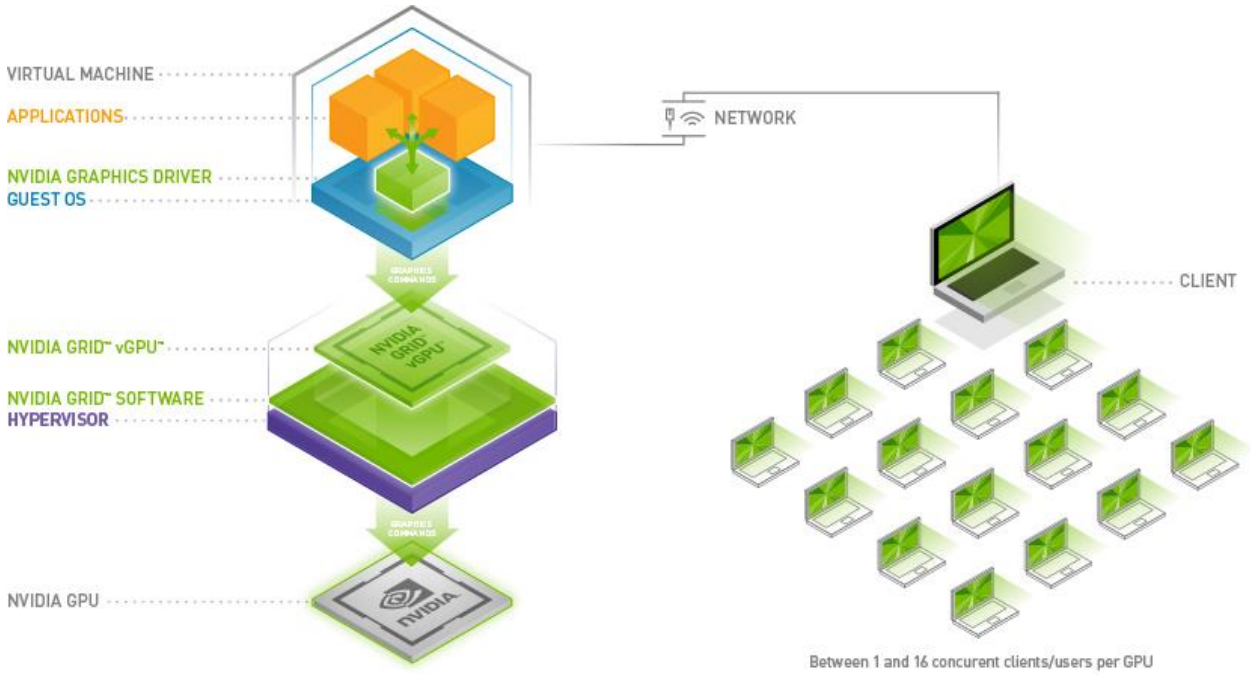


Image provided curtesy of NVIDIA Corporation, Copyright NVIDIA Corporation

4.4.1 vGPU Profiles

Virtual Graphics Processing Unit, or GRID vGPU™, is technology developed by NVIDIA® that enables hardware sharing of graphics processing for virtual desktops. This solution provides a hybrid shared mode allowing the GPU to be virtualized while the virtual machines run the native NVIDIA video drivers for better performance. Thanks to OpenGL support, VMs have access to more graphics applications. When utilizing vGPU, the graphics commands from virtual machines are passed directly to the GPU without any hypervisor translation. Every virtual desktop has dedicated graphics memory so they always have the resources they need to launch and run their applications at full performance. All this is done without sacrificing server performance and so is truly cutting edge.

The combination of Dell servers, NVIDIA GRID vGPU™ technology and NVIDIA GRID™ cards enable high-end graphics users to experience high fidelity graphics quality and performance, for their favorite applications at a reasonable cost.

For more information about NVIDIA GRID vGPU, please visit: LINK

The number of users per appliance is determined by the number of GPU cards in the system (max 2), vGPU profiles used for each GPU in a card (2 GPUs per card), and GRID license type. The same profile must be used on a single GPU but profiles can differ across GPUs in a single card.

NVIDIA® Tesla® M10 GRID vGPU Profiles:

Card	vGPU Profile	Graphics Memory (Frame Buffer)	Virtual Display Heads	Maximum Resolution	Maximum Graphics-Enabled VMs		
					Per GPU	Per Card	Per Server (2 cards)
Tesla M10	M10-8Q	8GB	4	4096x2160	1	4	8
	M10-4Q	4GB	4	4096x2160	2	8	16
	M10-2Q	2GB	4	4096x2160	4	16	32
	M10-1Q	1GB	2	4096x2160	8	32	64
	M10-0Q	512MB	2	2560x1600	16	64	128
	M10-1B	1GB	4	2560x1600	8	32	64
	M10-0B	512MB	2	2560x1600	16	64	128
	M10-8A	8GB	1	1280x1024	1	4	8
	M10-4A	4GB			2	8	16
	M10-2A	2GB			4	16	32
	M10-1A	1GB			8	32	64

Card	vGPU Profile	Guest VM OS Supported*		GRID License Required
		Win	64bit Linux	
Tesla M10	M10-8Q	●	●	GRID Virtual Workstation
	M10-4Q	●	●	
	M10-2Q	●	●	
	M10-1Q	●	●	
	M10-0Q	●	●	
	M10-1B	●		GRID Virtual PC
	M10-0B	●		
	M10-8A	●		GRID Virtual Application
	M10-4A	●		
	M10-2A	●		
	M10-1A	●		

Supported Guest VM Operating Systems*	
Windows	**Linux**
Windows 7 (32/64-bit)	RHEL 6.6 & 7
Windows 8.x (32/64-bit)	CentOS 6.6 & 7
Windows 10 (32/64-bit)	Ubuntu 12.04 & 14.04 LTS
Windows Server 2008 R2	
Windows Server 2012 R2	
Windows Server 2016	

***NOTE:** Supported guest operating systems listed as of the time of this writing. Please refer to NVIDIA's documentation for latest supported operating systems.

NVIDIA® Tesla® M60 GRID vGPU Profiles:

Card	vGPU Profile	Graphics Memory (Frame Buffer)	Virtual Display Heads	Maximum Resolution	Maximum Graphics-Enabled VMs		
					Per GPU	Per Card	Per Server (2 cards)
Tesla M60	M60-8Q	8GB	4	4096x2160	1	2	4
	M60-4Q	4GB	4	4096x2160	2	4	8
	M60-2Q	2GB	4	4096x2160	4	8	16
	M60-1Q	1GB	2	4096x2160	8	16	32
	M60-0Q	512MB	2	2560x1600	16	32	64
	M60-1B	1GB	4	2560x1600	8	16	32
	M60-0B	512MB	2	2560x1600	16	32	64
	M60-8A	8GB	1	1280x1024	1	2	4
	M60-4A	4GB			2	4	8
	M60-2A	2GB			4	8	16
	M60-1A	1GB			8	16	32

Card	vGPU Profile	Guest VM OS Supported*		GRID License Required
		Win	64bit Linux	
Tesla M60	M60-8Q	●	●	GRID Virtual Workstation
	M60-4Q	●	●	
	M60-2Q	●	●	
	M60-1Q	●	●	
	M60-0Q	●	●	
	M60-1B	●		GRID Virtual PC
	M60-0B	●		
	M60-8A	●		GRID Virtual Application
	M60-4A	●		
	M60-2A	●		
	M60-1A	●		

Supported Guest VM Operating Systems*	
Windows	**Linux**
Windows 7 (32/64-bit)	RHEL 6.6 & 7
Windows 8.x (32/64-bit)	CentOS 6.6 & 7
Windows 10 (32/64-bit)	Ubuntu 12.04 & 14.04 LTS
Windows Server 2008 R2	
Windows Server 2012 R2	
Windows Server 2016	

***NOTE:** Supported guest operating systems listed as of the time of this writing. Please refer to NVIDIA's documentation for latest supported operating systems.

4.4.1.1 GRID vGPU Licensing and Architecture

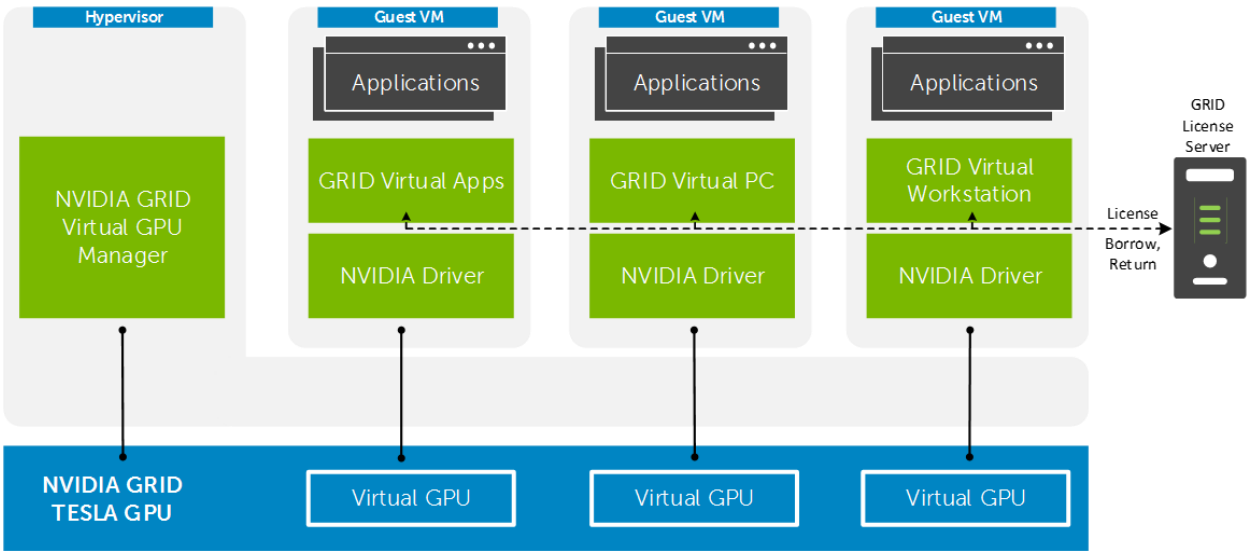
NVIDIA® GRID vGPU™ is offered as a licensable feature on Tesla® GPUs. vGPU can be licensed and entitled using one of the three following software editions.

NVIDIA GRID Virtual Applications	NVIDIA GRID Virtual PC	NVIDIA GRID Virtual Workstation
For organizations deploying XenApp or other RDSH solutions. Designed to deliver Windows applications at full performance.	For users who want a virtual desktop, but also need a great user experience leveraging PC applications, browsers, and high-definition video.	For users who need to use professional graphics applications with full performance on any device, anywhere.
Up to 2 displays @ 1280x1024 resolution supporting virtualized Windows applications	Up to 4 displays @ 2560x1600 resolution supporting Windows desktops, and NVIDIA Quadro features	Up to 4 displays @ 4096x2160* resolution supporting Windows or Linux desktops, NVIDIA Quadro, CUDA**, OpenCL** & GPU pass-through

*0Q profiles only support up to 2560x1600 resolution

**CUDA and OpenCL only supported with M10-8Q, M10-8A, M60-8Q, or M60-8A profiles

The GRID vGPU Manager, running on the hypervisor installed via the VIB, controls the vGPUs that can be assigned to guest VMs. A properly configured VM obtains a license from the GRID license server during the boot operation for a specified license level. The NVIDIA graphics driver running on the guest VM provides direct access to the assigned GPU. When the VM is shut down, it releases the license back to the server. If a vGPU enabled VM is unable to obtain a license, it will run at full capability without the license but users will be warned each time it tries and fails to obtain a license.



(Image provided courtesy of NVIDIA Corporation, Copyright NVIDIA Corporation)

5 Solution Architecture for Horizon View Appliance

5.1 Management Role Configuration

The appliance is designed to be a compute-only resource within a VMware Horizon View environment. For complete details on designing and deploying your entire Horizon View environment including Horizon management role requirements, SQL databases, and DNS, please refer to the **Dell Wyse Datacenter for VMware Horizon** reference architecture located here: LINK

5.1.1 NVIDIA GRID License Server Requirements

When using NVIDIA Tesla cards, graphics enabled VMs must obtain a license from a GRID License server on your network to be entitled for vGPU. To configure, a virtual machine with the following specifications must be added to a management host in addition to the management role VMs.

Role	vCPU	RAM (GB)	NIC	OS + Data vDisk (GB)	Tier 2 Volume (GB)
NVIDIA GRID License Srv	2	4	1	40 + 5	-

GRID License server software can be installed on a system running the following operating systems:

- Windows 7 (x32/x64)
- Windows 8.x (x32/x64)
- Windows 10 x64
- Windows Server 2008 R2
- Windows Server 2012 R2
- Red Hat Enterprise 7.1 x64
- CentOS 7.1 x64

Additional license server requirements:

- A fixed (unchanging) IP address. The IP address may be assigned dynamically via DHCP or statically configured, but must be constant.
- At least one unchanging Ethernet MAC address, to be used as a unique identifier when registering the server and generating licenses in NVIDIA's licensing portal.
- The date/time must be set accurately (all hosts on the same network should be time synchronized).

5.2 Storage Architecture Overview

The Dell Precision Appliance for Wyse is available with local Tier 1 storage but can be configured to utilize shared Tier 1 storage if desired. Both options are explained below. In general, the Dell Wyse Datacenter solution architecture has various Tier 1 and Tier 2 storage options to provide maximum flexibility to suit any use case. Customers have the choice to leverage best-of-breed Fiber Channel solutions from Dell EMC. For detailed explanations of these storage options, please refer to the **Dell Wyse Datacenter for VMware Horizon** reference architecture located here: LINK

5.2.1 Local Tier 1 Storage

Selecting the local Tier 1 storage model means that the compute host servers use locally installed hard drives to house the user desktop VMs. In this model, Tier 1 storage exists as local hard disks or SSDs on the Compute hosts themselves. To achieve the required performance level, RAID 10 is recommended for use across all local disks. A single volume per local Tier 1 Compute host is sufficient to host the provisioned desktop VMs along with their respective write caches.

5.2.2 Shared Tier 1 Storage

Selecting the Shared Tier 1 model means that the virtualization compute hosts are deployed without Tier 1 local storage and leverage shared storage hosted on a high performance Dell EMC Storage array. In this model, shared storage is leveraged for Tier 1 and used for VDI execution and write cache. Based on the heavy performance requirements of Tier 1 for VDI, it is recommended to use separate arrays for Tier 1 and Tier 2 when possible. Dell recommends using 500GB LUNs for VDI and running no more than 125 VMs per volume which will yields 4GB usable space per VM. If more usable space is required per VM, reduce the number of VMs per volume accordingly so that **Number of VMs per volume** = **500GB** / **Per VM disk space**. A VMware Horizon View replica to support a 1 to 500 desktop VM ratio should be located in a dedicated Replicas volume.

Volumes	Size (GB)	Storage Array	Purpose	File System
VDI-BaseImages	100 GB	Tier 1	Storage for Base VDI images	VMFS
VDI-Replicas	100 GB	Tier 1	Storage for Replica images	VMFS
VDI-1	500	Tier 1	Max 125 desktop VMs	VMFS
VDI-2	500	Tier 1	Max 125 desktop VMs	VMFS
VDI-3	500	Tier 1	Max 125 desktop VMs	VMFS
VDI-4	500	Tier 1	Max 125 desktop VMs	VMFS

*Volume names are examples

5.3 Virtual Networking

The network configurations presented in this section only represent virtual networking for the appliance (graphics compute host). For networking details for your management and non-graphics compute hosts, please refer to the **Dell Wyse Datacenter for VMware Horizon** reference architecture located here: LINK

NOTE: vMotion cannot be used with vGPU virtual machines.

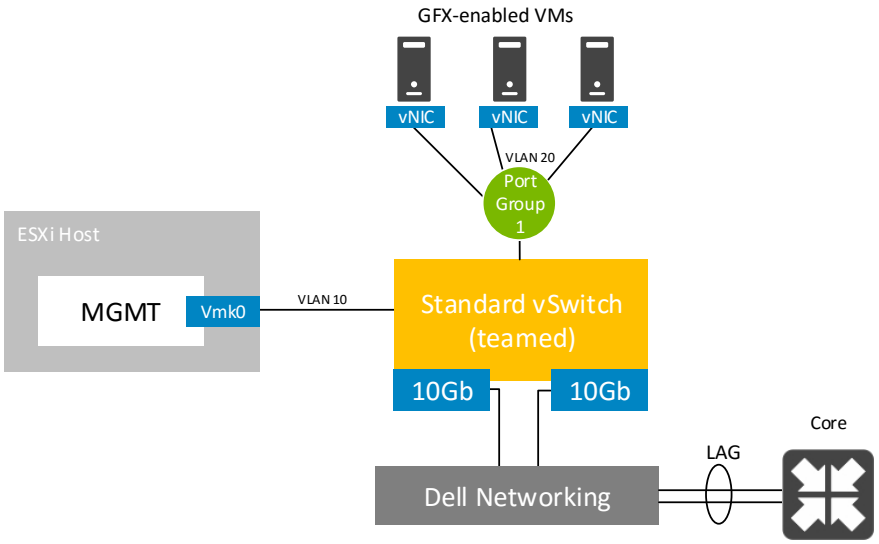
5.3.1 Local Tier 1

In this model, appliances do not need access to shared storage since they are hosting VDI VMs on local disk. The following outlines the VLAN requirements for the appliances in this solution model:

- Appliances (Graphics compute hosts - Local Tier 1)
    - Management VLAN: Configured for hypervisor infrastructure traffic – L3 routed via core switch
    - VDI VLAN: Configured for VDI session traffic – L3 routed via core switch
- A VLAN for iDRAC is configured for all hardware management traffic – L3 routed via core switch

Following best practices, LAN and block storage traffic is separated in solutions >500 users. This traffic is combined within a single switch in smaller stacks to minimize the initial investment, however, VLANs are required for each traffic type to enable traffic separation. By default, each Local Tier 1 Compute host will have a quad port NDC which includes both 10Gb and 1Gb interfaces. Configure the LAN traffic from the server to the ToR switch as a LAG.

dvSwitches should be used as desired for VM traffic especially in larger deployments to ease the management burden across numerous hosts. Desktop VMs will connect to the primary port group on the external vSwitch. Network share values should be configured equally among the VMKernel port groups that share a physical set of network adapters.

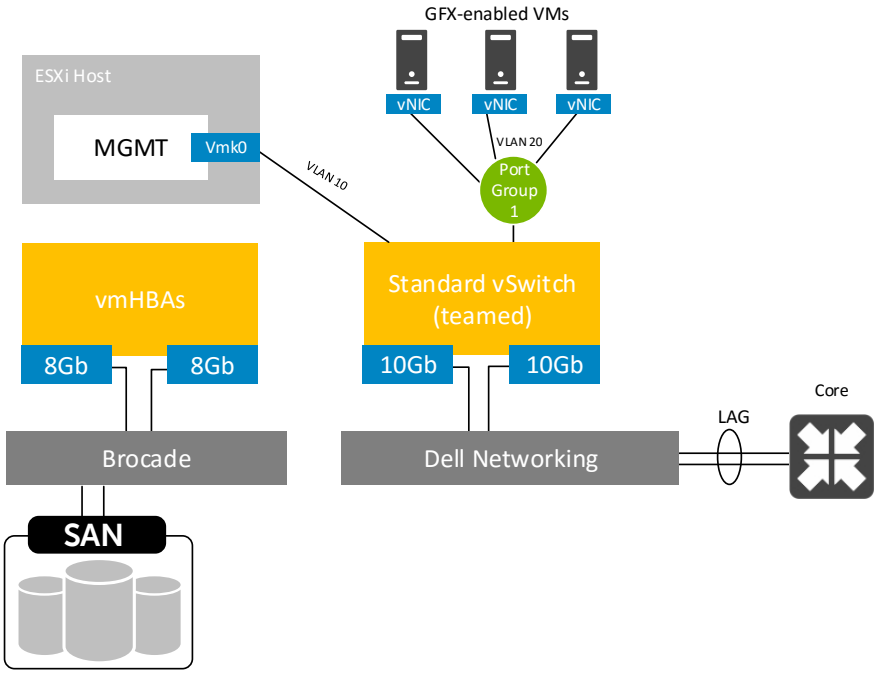


5.3.2 Shared Tier 1

Using Fiber Channel based storage eliminates the need to build iSCSI into the network stack but requires additional storage fabrics to be built out. The following outlines the VLAN requirements for the appliances in this solution model:

- Appliances (Graphics compute hosts - Shared Tier 1)
    - Management VLAN: Configured for hypervisor Management traffic – L3 routed via core switch
    - VDI VLAN: Configured for VDI session traffic – L3 routed via core switch
- A VLAN for iDRAC is configured for all hardware management traffic – L3 routed via core switch

FC and LAN traffic are physically separated into discrete switching Fabrics. By default, each Shared Tier 1 Compute host will have a quad port NDC, which includes both 10Gb and 1Gb interfaces, as well as dual port (2 x 8Gb) FC HBAs. LAN traffic from the server to the ToR switch is configured as a LAG. Network share values should be configured equally among the VMKernel port groups that share a physical set of network adapters.



5.4 Scaling Guidance

The use of GPU cards dictates a maximum number of graphics accelerated VMs per host based on a 1:1 mapping of vGPU to VM with the number of vGPUs available determined by the GPU card, the number of GPU cards, and vGPU profiles chosen. For example, with two Tesla M60 GPU cards, an appliance can have a maximum of 32 VMs using the M60-1Q profile. Therefore, vertical scalability per appliance can be adjusted

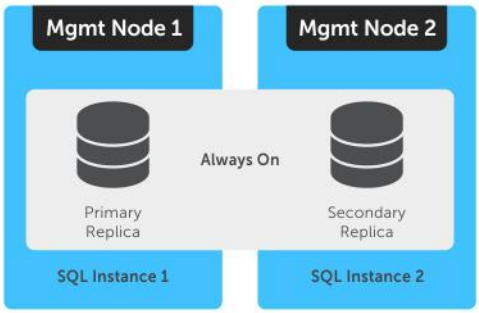
with additional host RAM and CPU choices but will be primarily limited by the GPUs and vGPU configuration. The graphics compute layer can be scaled horizontally by adding additional appliances and clusters as needed bearing in mind a maximum of 64 nodes per vSphere cluster.

Please refer to tables in the vGPU Profiles section for the maximum vGPU allocations for each profile.

For scaling guidance on all Dell Wyse Datacenter solution components, please refer to the **Dell Wyse Datacenter for VMware Horizon** reference architecture located here: LINK.

5.5 Solution High Availability

High availability (HA) is offered to protect each architecture solution layer, individually if desired. Following the N+1 model, additional ToR switches are added to the Network layer and stacked to provide redundancy as required, additional compute and management hosts are added to their respective layers, vSphere clustering is introduced in both the management and compute layers, SQL is configured for Always On or clustered and F5 is leveraged for load balancing.



The HA options provide redundancy for all critical components in the stack while improving the performance and efficiency of the solution as a whole.

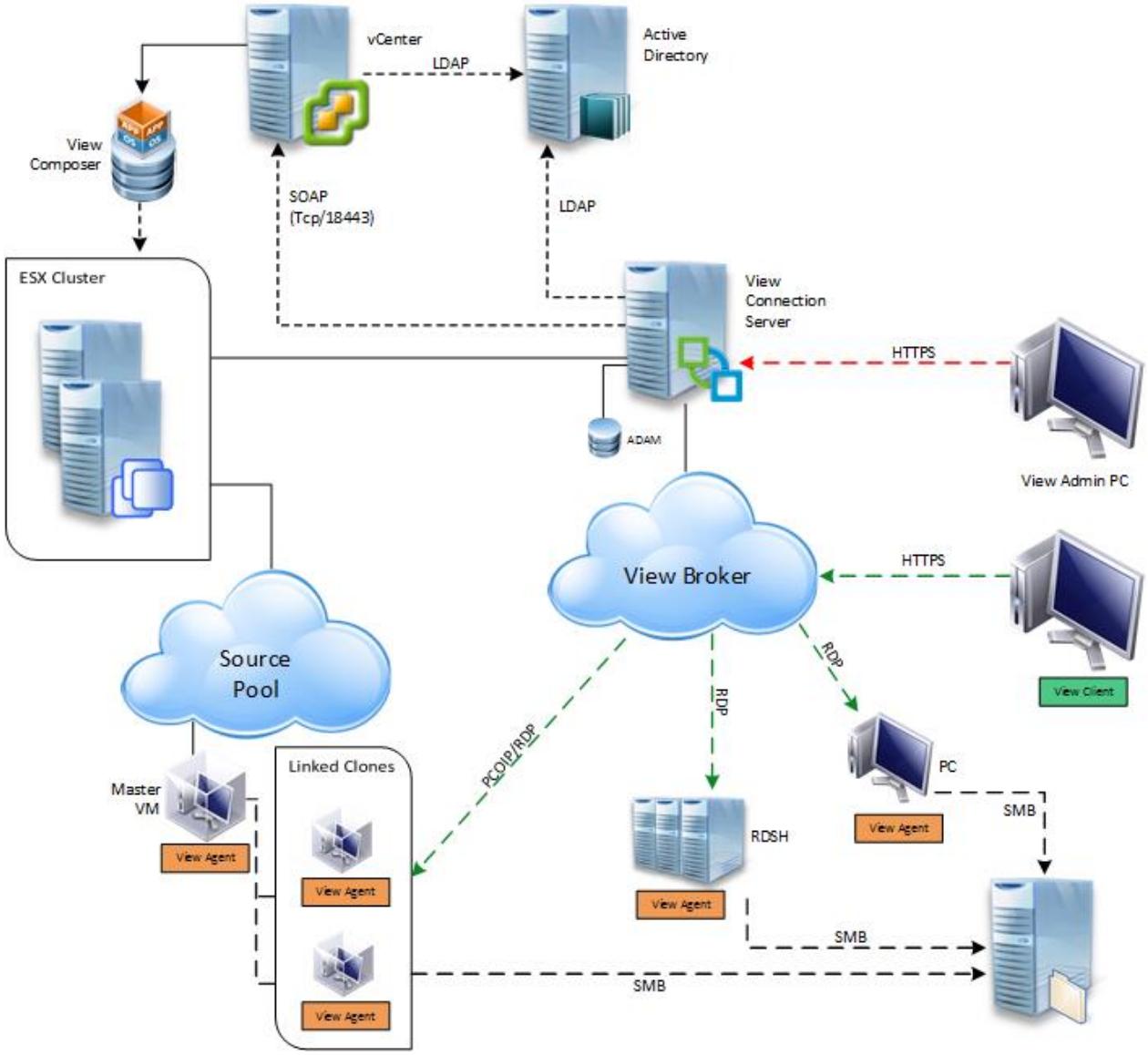
- Additional switches added to the existing thereby equally spreading each host's network connections across multiple switches.
- Additional ESXi hosts added in the compute or mgmt. layers to provide N+1 protection.
- Applicable VMware Horizon infrastructure server roles are duplicated and spread amongst mgmt. host instances where connections to each are load balanced via the addition of F5 appliances.
- SQL Server databases also are protected through the addition and configuration of an "Always On" Failover Cluster Instance or Availability Group.

For complete HA details, please refer to the **Dell Wyse Datacenter for VMware Horizon** reference architecture located here: LINK

NOTE: vMotion cannot be used with vGPU virtual machines.

5.6 Dell Wyse Datacenter for Horizon Communication Flow

The following diagram provides an overview of the communication flow within a Horizon View environment.



6 Solution Performance and Testing

The maximum appliance density is determined by the number of GPU cards present in the appliance and the vGPU profiles used for the virtual machines. Please refer to tables in the vGPU Profiles section for the maximum vGPU allocations for each profile.

User density summary

Host Config	Hypervisor	Broker & Provisioning	Workload	Template	GPU Profile	User Density
High Density w/M60 GPUs	ESXi 6.0 U2	Horizon 7	Graphics LVSI Custom – Density	Windows 8.1 x64 NVIDIA Driver: 362.56	M60-1Q	32
High Density w/M60 GPUs	ESXi 6.5	Horizon 7	Graphics LVSI Custom – Density	Windows 8.1 x64 NVIDIA Driver: 369.71	M60-1Q	32
High Density w/M10 GPUs	ESXi 6.0 U2	Horizon 7	Graphics LVSI Power + ProLibrary	Windows 8.1 x64 NVIDIA Driver: 368.82	M10-1Q	64

The detailed validation results and analysis of these reference designs are in the following sections.

6.1 Test and performance analysis methodology

6.1.1 Testing process

In order to ensure the optimal combination of end-user experience (EUE) and cost-per-user, performance analysis and characterization (PAAC) on Dell Wyse Datacenter solutions is carried out using a carefully designed, holistic methodology that monitors both hardware resource utilization parameters and EUE during load-testing.

Login VSI is currently the load-generation tool used during PAAC of Dell Wyse Datacenter solutions. Each user load is tested against multiple runs. First, a pilot run to validate that the infrastructure is functioning and valid data can be captured, and then, subsequent runs allowing correlation of data.

At different times during testing, the testing team will complete some manual “User Experience” Testing while the environment is under load. This will involve a team member logging into a session during the run and completing tasks similar to the User Workload description. While this experience will be subjective, it will help provide a better understanding of the end user experience of the desktop sessions, particularly under high load, and ensure that the data gathered is reliable.

6.1.1.1 Load generation

Login VSI by Login Consultants is the de-facto industry standard tool for testing VDI environments and server-based computing (RDSH environments). It installs a standard collection of desktop application software (e.g. Microsoft Office, Adobe Acrobat Reader) on each VDI desktop; it then uses launcher systems to connect a specified number of users to available desktops within the environment. Once the user is connected, the workload is started via a logon script which starts the test script once the user environment is configured by the login script. Each launcher system can launch connections to a number of 'target' machines (i.e. VDI desktops). The launchers and Login VSI environment are configured and managed by a centralized management console.

Additionally, the following login and boot paradigm is used:

- Users are logged in within a login timeframe of 1 hour. Exception to this login timeframe occurs when testing low density solutions such as GPU/graphics based configurations. With those configurations, users are logged on every 10-15 seconds.
- All desktops are pre-booted in advance of logins commencing.
- All desktops run an industry-standard anti-virus solution. Windows Defender is used for Windows 10 due to issues implementing McAfee.

6.1.1.2 Profiles and workloads

It's important to understand user workloads and profiles when designing a desktop virtualization solution in order to understand the density numbers that the solution can support. At Dell, we use five workload / profile levels, each of which is bound by specific metrics and capabilities with two targeted at graphics-intensive use cases. We will present more detailed information in relation to these workloads and profiles below but first it is useful to define the terms "profile" and "workload" as they are used in this document.

- **Profile:** This is the configuration of the virtual desktop - number of vCPUs and amount of RAM configured on the desktop (i.e. available to the user).
- **Workload:** This is the set of applications used by performance analysis and characterization (PAAC) of Dell Wyse Datacenter solutions (e.g. Microsoft Office applications, PDF Reader, Internet Explorer etc.)

Load-testing on each profile is carried out using an appropriate workload that is representative of the relevant use case and summarized in the table below:

Profile to workload mapping

Profile Name	Workload
Task Worker	Login VSI Task worker
Knowledge Worker	Login VSI Knowledge worker
Power Worker	Login VSI Power worker

Graphics LVSI Power + ProLibrary	Graphics - Login VSI Power worker with ProLibrary
Graphics LVSI Custom	Graphics – LVSI Custom

Login VSI workloads are summarized in the sections below. Further information for each workload can be found on Login VSI's website.

Login VSI Task Worker Workload

The Task Worker workload runs fewer applications than the other workloads (mainly Excel and Internet Explorer with some minimal Word activity, Outlook, Adobe, copy and zip actions) and starts/stops the applications less frequently. This results in lower CPU, memory and disk IO usage.

Login VSI Knowledge Worker Workload

The Knowledge Worker workload is designed for virtual machines with 2vCPUs. This workload and contains the following activities:

- Outlook, browse messages.
- Internet Explorer, browse different webpages and a YouTube style video (480p movie trailer) is opened three times in every loop.
- Word, one instance to measure response time, one instance to review and edit a document.
- Doro PDF Printer & Acrobat Reader, the Word document is printed and exported to PDF.
- Excel, a very large randomized sheet is opened.
- PowerPoint, a presentation is reviewed and edited.
- FreeMind, a Java based Mind Mapping application.
- Various copy and zip actions.

Login VSI Power Worker Workload

The Power Worker workload is the most intensive of the standard workloads. The following activities are performed with this workload:

- Begins by opening four instances of Internet Explorer which remain open throughout the workload.
- Begins by opening two instances of Adobe Reader which remain open throughout the workload.
- There are more PDF printer actions in the workload as compared to the other workloads.
- Instead of 480p videos a 720p and a 1080p video are watched.
- The idle time is reduced to two minutes.
- Various copy and zip actions.

Graphics - Login VSI Power Worker with ProLibrary workload

For lower performance graphics testing where lower amounts of graphics memory are allocated to each VM, the Power worker + Pro Library workload is used. The Login VSI Pro Library is an add-on for the Power worker workload which contains extra content and data files. The extra videos and web content of the Pro Library utilizes the GPU capabilities without overwhelming the lower frame buffer assigned to the desktops. This type of workload is typically used with high density vGPU and sVGA or other shared graphics configurations.

Graphics – LVSI Custom workload

This is a custom Login VSI workload specifically for higher performance, intensive graphics testing. For this workload, SPECwpc benchmark application is installed to the client VMs. During testing, a script is started that launches SPECwpc which executes the Maya and sw-03 modules for high performance tests and module sw-03 only for high density tests. The usual activities such as Office application execution are not performed with this workload. This type of workload is typically used for lower density/high performance pass-through, vGPU, and other dedicated, multi-user GPU configurations.

6.1.2 Resource monitoring

The following sections explain respective component monitoring used across all Dell Wyse Datacenter solutions where applicable.

6.1.2.1 GPU resources

ESXi hosts

For gathering of GPU related resource usage, a script is executed on the ESXi host before starting the test run and stopped when the test is completed. The script contains NVIDIA System Management Interface commands to query each GPU and log GPU utilization and GPU memory utilization into a .csv file.

ESXi 6.5 and above includes the collection of this data in the vSphere Client/Monitor section. GPU processor utilization, GPU temperature, and GPU memory utilization can be collected the same was as host CPU, host memory, host Network, etc.

6.1.2.2 VMware vCenter

VMware vCenter is used for VMware vSphere-based solutions to gather key data (CPU, Memory, Disk and Network usage) from each of the compute hosts during each test run. This data is exported to .csv files for single hosts and then consolidated to show data from all hosts (when multiple are tested). While the report does not include specific performance metrics for the Management host servers, these servers are monitored during testing to ensure they are performing at an expected performance level with no bottlenecks.

6.1.3 Resource utilization

Poor end-user experience is one of the main risk factors when implementing desktop virtualization but a root cause for poor end-user experience is resource contention: hardware resources at some point in the solution have been exhausted, thus causing the poor end-user experience. In order to ensure that this does not happen, PAAC on Dell Wyse Datacenter solutions monitors the relevant resource utilization parameters and

applies relatively conservative thresholds as shown in the table below. Thresholds are carefully selected to deliver an optimal combination of good end-user experience and cost-per-user, while also providing burst capacity for seasonal / intermittent spikes in usage. Utilization within these thresholds is used to determine the number of virtual applications or desktops (density) that are hosted by a specific hardware environment (i.e. combination of server, storage and networking) that forms the basis for a Dell Wyse Datacenter RA.

Resource utilization thresholds

Parameter	Pass/Fail Threshold
Physical Host CPU Utilization (AHV & ESXi hypervisors)*	100%
Physical Host CPU Utilization (Hyper-V)	85%
Physical Host Memory Utilization	85%
Network Throughput	85%
Storage IO Latency	20ms

*Turbo mode is enabled; therefore, the CPU threshold is increased as it will be reported as over 100% utilization when running with turbo.

6.2 Test configuration details

The following components were used to complete the validation testing for the solution:

Hardware and software test components

Component	Description/Version
Hardware platform(s)	PowerEdge R730
Hypervisor(s)	ESXi 6.0 U2, Build 3620759 and ESXi 6.5, Build 4564106 (M60 w/ Win 10 tests)
Broker technology	Horizon 7 and Horizon 7.0.3
NVIDIA GRID Software	Versions 3.1 (M60 tests), 4.0 (M10 tests), and 4.1 (M60 w/Win 10 tests)
Broker database	Microsoft SQL 2014
Virtual desktop OS	Windows 8.1 Enterprise 64-bit and Windows 10 Enterprise 64-bit

Office application suite	Office 2010 w/Win 8.1 (M60 tests), Office 2016 w/Win 10 (M60 tests), and Office 2013 (M10 tests)
Login VSI test suite	Version 4.1
VMware Horizon Agent	7.0.0.3618085 and 7.0.2.4368292
VMware Tools	10.0.6.3560309 and 10.272 (M60 w/Win 10 tests)
NVIDIA GRID 3.1 Driver for ESXi	361.45.09 (M60 tests)
NVIDIA GRID 4.0 Driver for ESXi	367.34 (M10 tests)
NVIDIA GRID 4.1 Driver for ESXi	367.64 / OEM 650.0.0.3240417.vib (M60 w/Win 10 tests)

6.2.1 Compute VM Configurations

The following table summarizes the compute VM configurations for the various profiles/workloads tested.

Desktop VM specifications

User Profile	vCPUs	ESXi Memory Configured	ESXi Memory Reservation	Hyper-V Startup Memory	Hyper-V Min | Max Dynamic	Operating System
Task Worker	1	2GB	1GB	1GB	1GB | 2GB	Windows 10 Enterprise 64-bit
Knowledge Worker	2	3GB	1.5GB	1.5GB	1GB | 3GB	Windows 10 Enterprise 64-bit
Power Worker	2	4GB	2GB	2GB	1GB | 4GB	Windows 10 Enterprise 64-bit
Graphics LVSI Power + ProLibrary	2	4 GB	4GB			Windows 8.1/10 Enterprise 64-bit
Graphics LVSI Custom – Density	2	4 GB	4GB			Windows 8.1/10 Enterprise 64-bit
Graphics LVSI Custom - Performance	4	8GB	8GB			Windows 8.1/10 Enterprise 64-bit

Screen resolutions

User Profile	Screen Resolution
Task Worker	1280 X 720
Knowledge Worker	1920 X 1080
Power Worker	1920 X 1080
Graphics LVSI Power + ProLibrary	1920 X 1080*
Graphics LVSI Custom – Density	1920 X 1080*
Graphics LVSI Custom - Performance	1920 X 1080*

*Although SPECwpc 2.0 now supports native resolution of 2560x1600 as can be confirmed from the SPECwpc log file, it is not officially supported with this version on SPECwpc 2.0. Some parts of the viewsets do not scale to the higher resolution without showing artifacts. Although very few, these artifacts have been observed during the test runs with 2560x1600 resolution. Recommended resolution for running SPECwpc is still 1920x1080.

6.2.2 Platform Configurations

The hardware configurations that were tested are summarized in the table(s) below.

High Density - M60 GPUs hardware configuration

Enterprise Platform	Platform Config	CPU	Memory	RAID Ctlr	HD Config	Network	GPUs
R730	High Density	E5-2698v4 (20 Core, 2.2GHz)	256GB @2400 MT/s	PERC H730, 2GB Cache	4 X 800GB, Intel S3610 SSD's	Intel 10Gbps 2P X540 Intel 2P X520 + 2P I350 NDC	2 x NVIDIA Tesla M60

High Density – M10 GPUs hardware configuration

Enterprise Platform	Platform Config	CPU	Memory	RAID Ctlr	HD Config	Network	GPUs
R730	High Density	E5-2698v4 (20 Core, 2.2GHz)	512GB @2400 MT/s	PERC H730, 2GB Cache	4 X 800GB, Intel S3610 SSD's	Intel 10Gbps 2P X540 Intel 2P X520 + 2P I350 NDC	2 x NVIDIA Tesla M10

6.3 Test results and analysis

The following table summarizes the test results for the compute hosts using the various workloads and configurations. Refer to the prior section for platform configuration details.

Host metrics test result summary

Platform Config	Hypervisor	Broker & Provisioning	Login VSI Workload	Density Per Host	Avg CPU	Avg Mem Consumed	Avg Mem Active	Avg IOPS / User	Avg Net Mbps / User
High Density - M60 GPUs	ESXi 6.0 U2	Horizon 7	Graphics LVSI Custom – Density	32	78%	144GB	62GB	18	16.15
High Density - M60 GPUs	ESXi 6.5	Horizon 7	Graphics LVSI Custom – Density	32	70%	143GB	135GB	14	15.1
High Density – M10 GPUs	ESXi 6.0 U2	Horizon 7	Graphics LVSI Power + ProLibrary	64	73%	290GB	70GB	15	3.2

GPU metrics test result summary

Platform Config	Hypervisor	Broker & Provisioning	Login VSI Workload	Density Per Host	GPU Profile	Avg GPU Processor	Avg GPU Memory
High Density - M60 GPUs	ESXi 6.0 U2	Horizon 7	Graphics LVSI Custom – Density	32	M60-1Q	87%	39%
High Density - M60 GPUs	ESXi 6.5	Horizon 7	Graphics LVSI Custom – Density	32	M60-1Q	85%	43%
High Density – M10 GPUs	ESXi 6.0 U2	Horizon 7	Graphics LVSI Power + ProLibrary	64	M10-1Q	70%	28%

Density Per Host: Density reflects number of users per compute host that successfully completed the workload test within the acceptable resource limits for the host. For clusters, this reflects the average of the density achieved for all compute hosts in the cluster.

Avg CPU: This is the average CPU usage over the steady state period. For clusters, this represents the combined average CPU usage of all compute hosts. On the latest Intel series processors, the ESXi host CPU metrics will exceed the rated 100% for the host if Turbo Boost is enabled (by default). An additional 35% of CPU is available from the Turbo Boost feature but this additional CPU headroom is not reflected in the

VMware vSphere metrics where the performance data is gathered. Therefore, CPU usage for ESXi hosts is adjusted and a line indicating the potential performance headroom provided by Turbo boost is included in each CPU graph.

Avg Consumed Memory: Consumed memory is the amount of host physical memory consumed by a virtual machine, host, or cluster. For clusters, this is the average consumed memory across all compute hosts over the steady state period.

Avg Mem Active: For ESXi hosts, active memory is the amount of memory that is actively used, as estimated by VMkernel based on recently touched memory pages. For clusters, this is the average amount of guest "physical" memory actively used across all compute hosts over the steady state period.

Avg IOPS/User: IOPS calculated from the average Disk IOPS figure over the steady state period divided by the number of users.

Avg Net Mbps/User: Amount of network usage over the steady state period divided by the number of users. For clusters, this is the combined average of all compute hosts over the steady state period divided by the number of users on a host.

Avg GPU utilization: Values included for GPU processor utilization and GPU memory utilization as a percentage during steady state.

For the ESXi 6.5 tests, GPU memory utilization in the vSphere Client/Monitor section does not include the the same memory utilization collected from the host via command line with nvidia-smi command:

vSphere Client/Monitor/GPU offers :

- Memory Usage [%] = amount of memory used in % = 92% - stays close to same during test
- Memory Use [KB] = amount of memory used in KB = 7789193 – stays close to same during test

This looks more like memory reserved, than actually active. It looks like a static value, similar to memory.total, memory.used, memory.free from the nvidia-smi collection tool.

nvidia-smi collection offers :

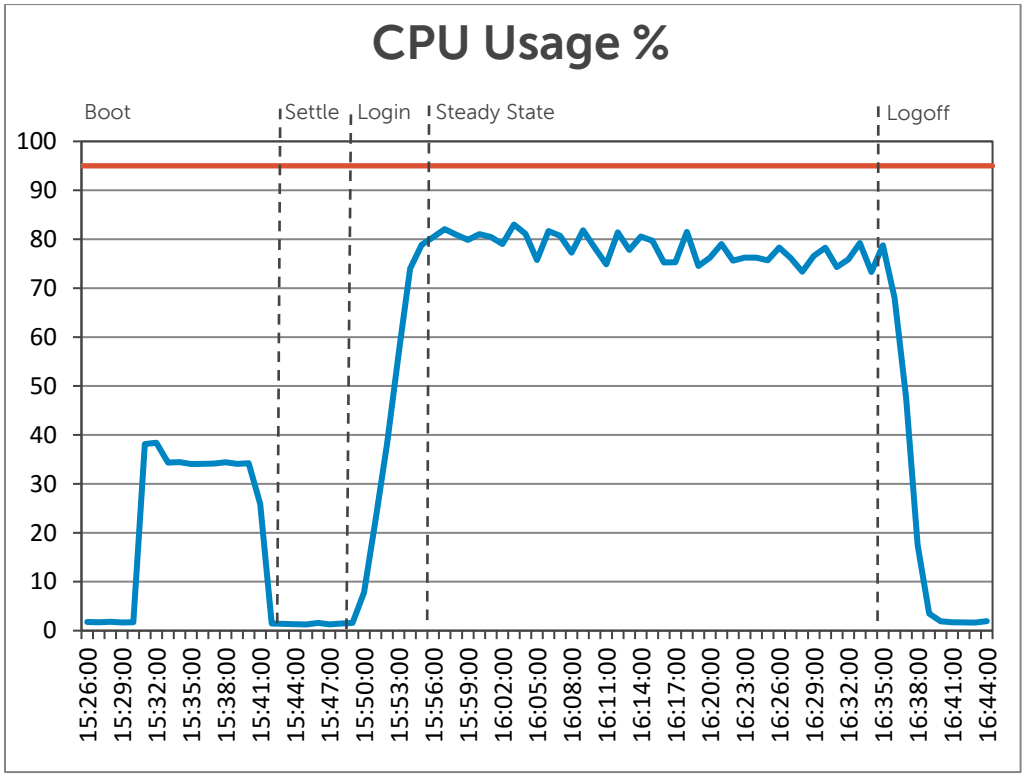
- Utilization.memory [%] = average 45% - this is not a static value
- memory.total [MiB] = 8191 MiB – stays close to the same during test
- memory.used [MiB] = 7603 MiB – stays close to the same during test
- memory.free MiB] = 588 MiB – stays close to the same during test

Results shown for the ESXi 6.5 test do not include the nvidia-smi script collection as an increase of +20% CPU utilization was observed due to its execution.

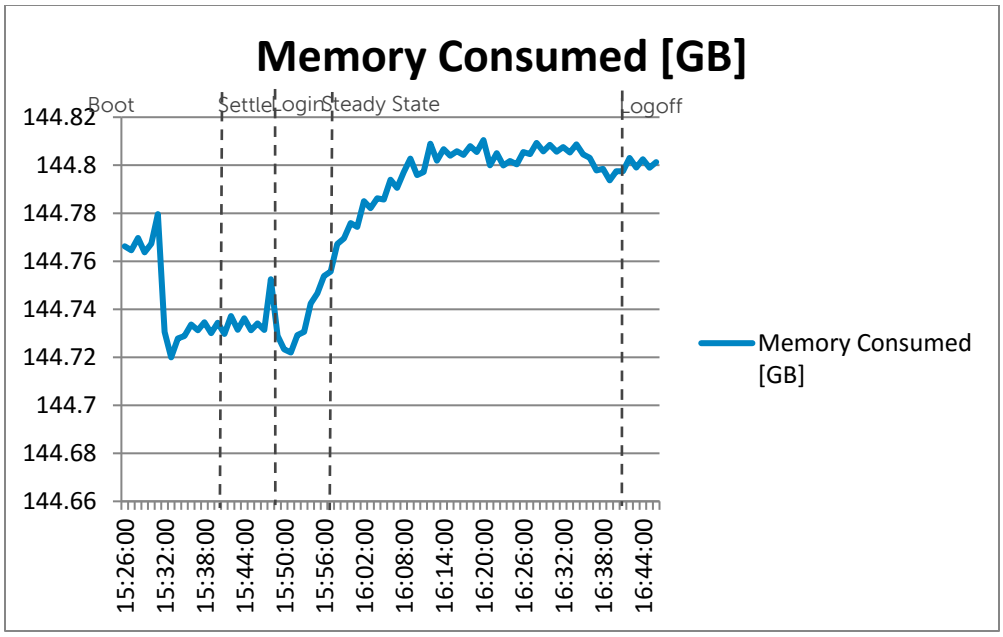
6.3.1 R730 High Density – M60 GPUs

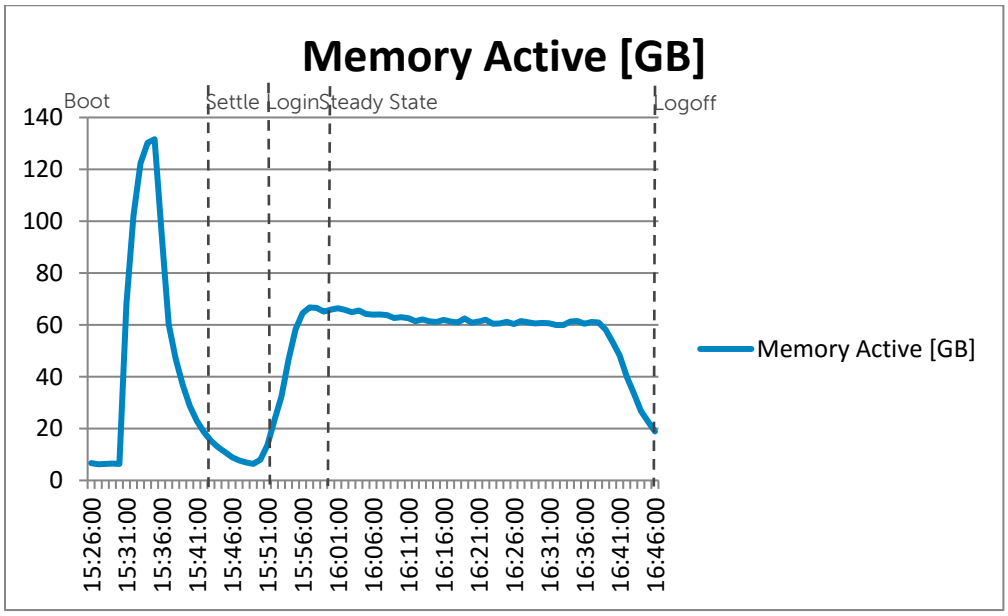
Refer to the Platform Configuration section for hardware configuration details.

6.3.1.1 Graphics LVSI Custom – Density, 32 Users, ESXi 6.0 U2, Horizon 7

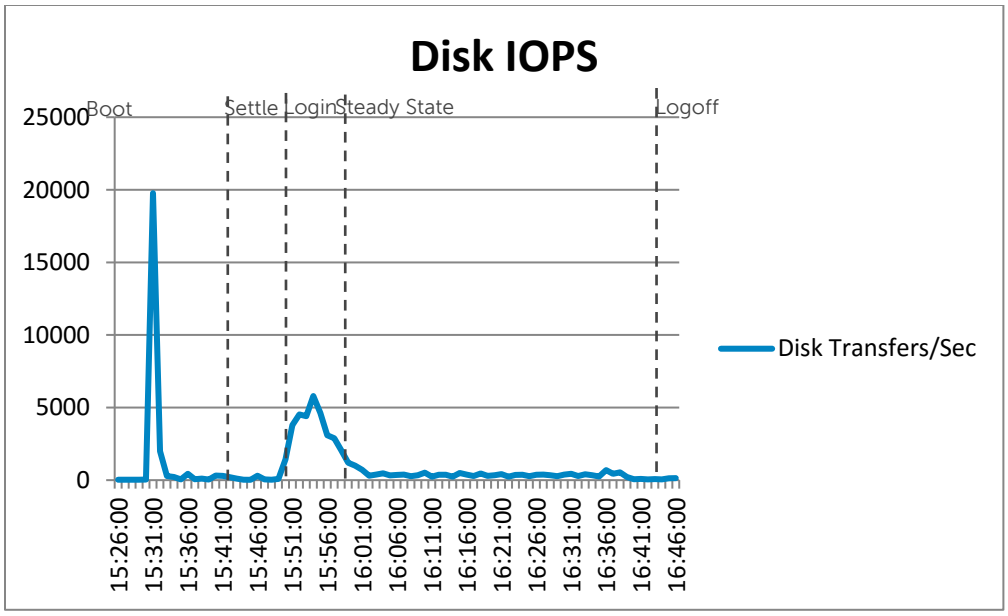


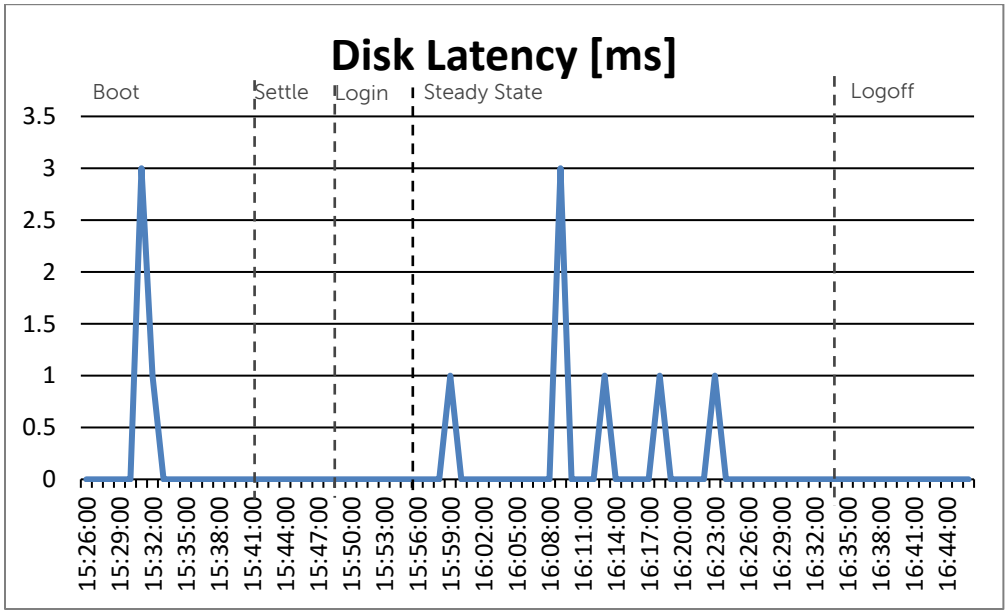
CPU usage for 32 clients is averaging below the 95% acceptance mark for this particular CPU [with Turbo enabled]. Maximum possible clients for the profile M60-1Q with a 2 GPU board/server configuration is 32 clients with a maximum resolution of 4096x2160. Maximum resolution was not used in this test since the test environment does presently not support this resolution. The resolution used was 1920x1080.



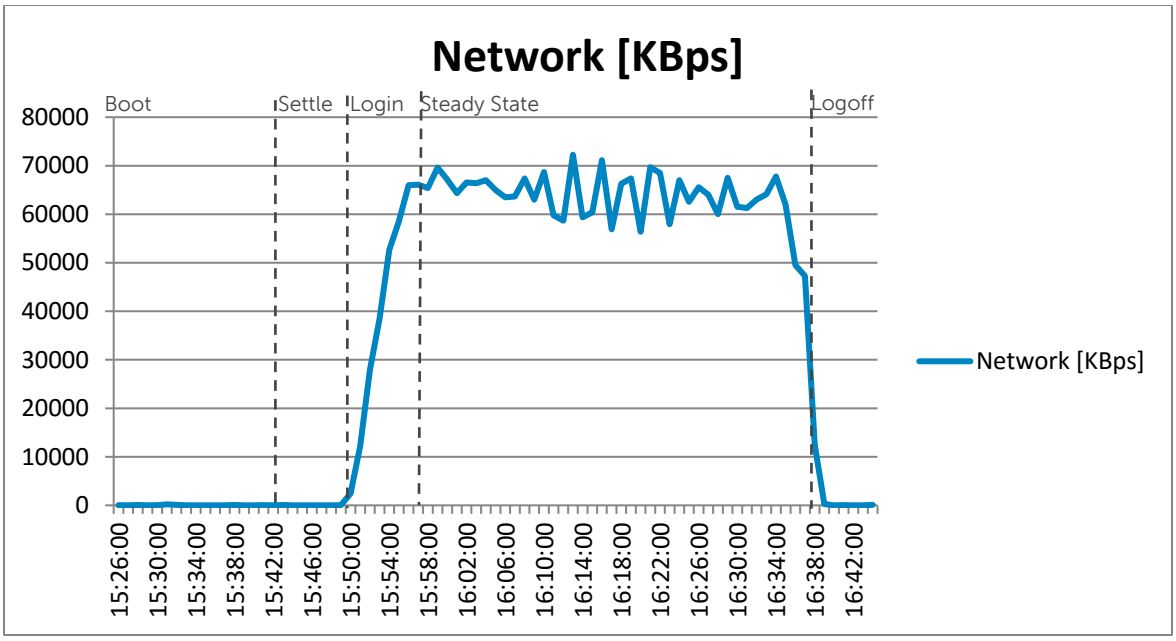


Host memory shows no bottle neck. For 32 clients with 4GB reserved RAM consumed memory should be 128GB + an overhead for vSphere services. On average the memory actively used by the clients is around half of the available memory. Swap Used [GB] and Balloon [GB] show 0 GB throughout the test.

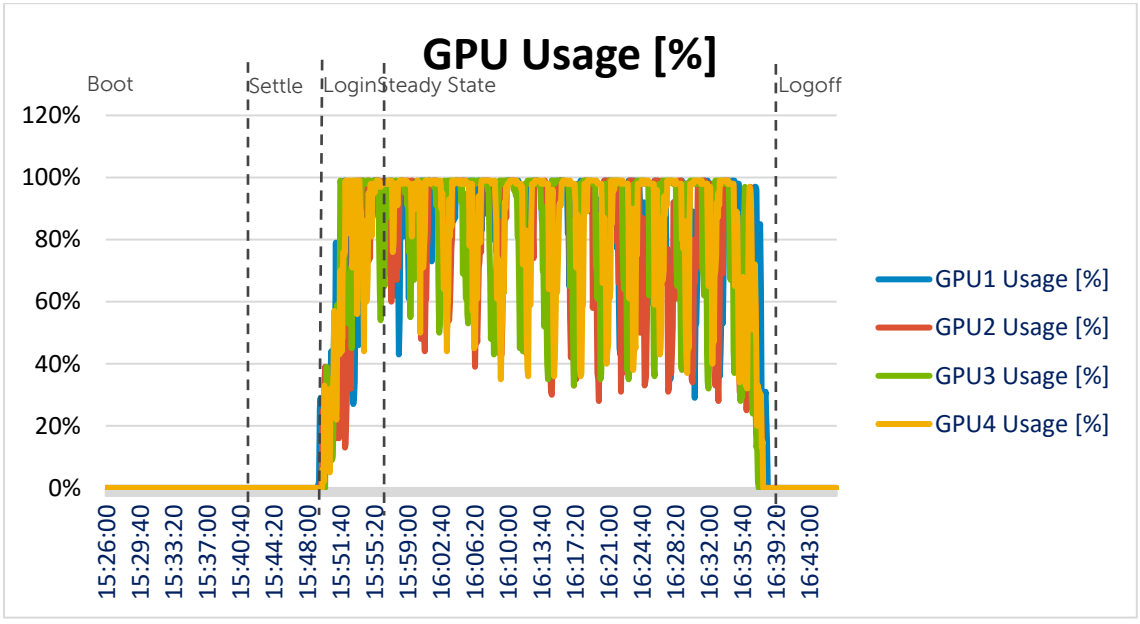




As expected, highest disk activity is during the log-in period.



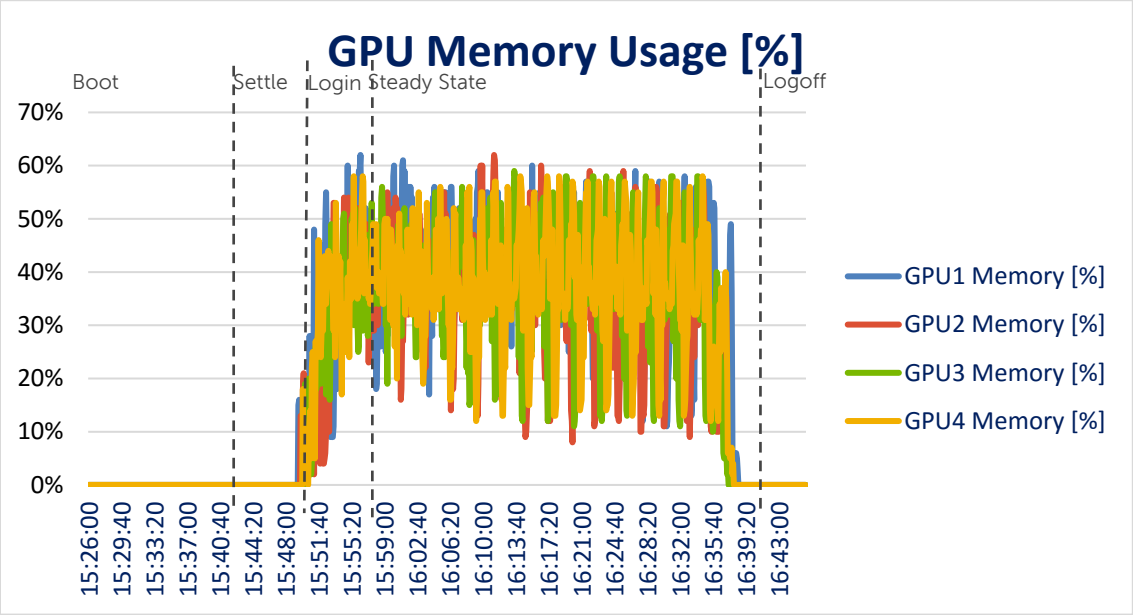
Network usage peak values are around the 0.58 Mbps [72256 Kbps] mark.



During steady-state average GPU processor utilization shows 87%. GPUs are expected to run at full capacity. A low GPU load could point to issues with server CPUs not being able to handle the graphics load or other issues holding back the GPU. This is not the case here.

GPU resources are not used continuously while running the graphics application. A graphics application clip runs for a few minutes. During that time the GPU is used. After this clip finishes, a new one is loaded during

which the GPU resources are not used. This explains the dips in the GPU usage and also the GPU memory usage.



On average 39% GPU memory was used during steady state testing. GPU memory utilization will depend on the graphics applications itself. SPECview module sw-03 could be considered a medium graphics load. High GPU memory usage should not be expected with this module.

SPECwpc Results

The table below shows the completion times for each sw-03 module viewset category.

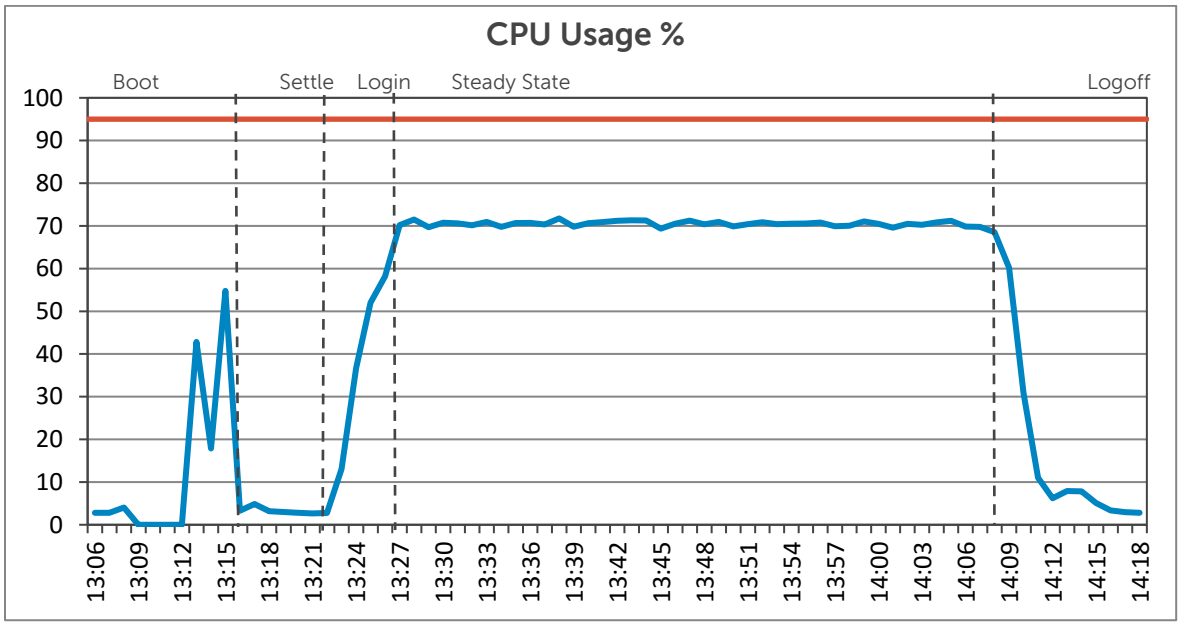
Summary | RawScores | M&E | ProdDev | LifeSci | FSI | Energy | GeneralOps | Configuration

Raw Categories

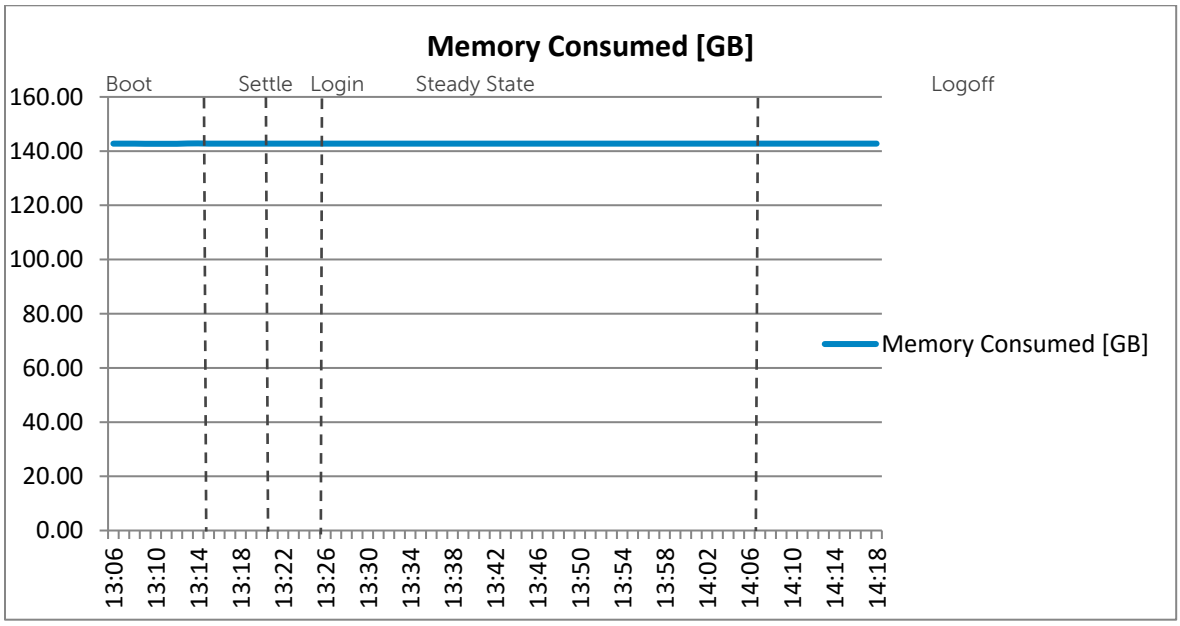
SPECwpcResults

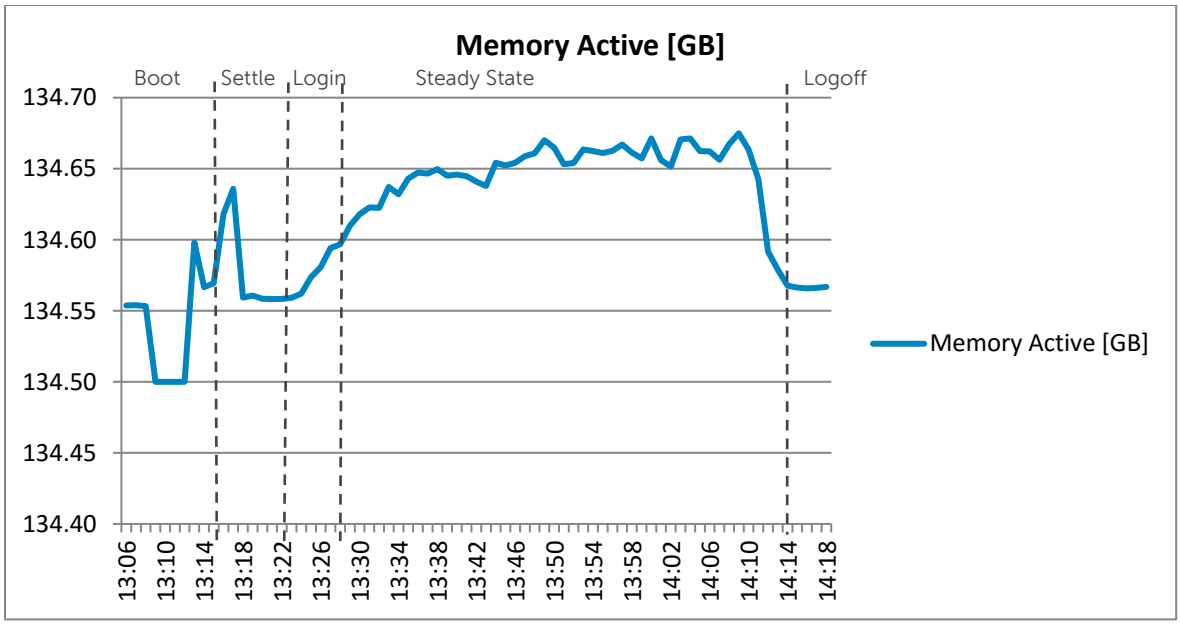
Test	Subtest	Ctgry	Run	Time	Score
sw-03	sw-03	1	1	7/6/2016 4:09:15 PM	14.01
sw-03	sw-03	2	1	7/6/2016 4:09:55 PM	9.95
sw-03	sw-03	3	1	7/6/2016 4:10:18 PM	18.99
sw-03	sw-03	4	1	7/6/2016 4:10:38 PM	19.68
sw-03	sw-03	5	1	7/6/2016 4:11:05 PM	15.09
sw-03	sw-03	6	1	7/6/2016 4:11:22 PM	53.07
sw-03	sw-03	7	1	7/6/2016 4:11:38 PM	59.8
sw-03	sw-03	8	1	7/6/2016 4:11:54 PM	33.33
sw-03	sw-03	9	1	7/6/2016 4:12:51 PM	6.49
sw-03	sw-03	10	1	7/6/2016 4:13:09 PM	21.31
sw-03	sw-03	11	1	7/6/2016 4:13:25 PM	52.13

6.3.1.2 Graphics LVSI Custom – Density, 32 Users, ESXi 6.5, Horizon 7.0.3

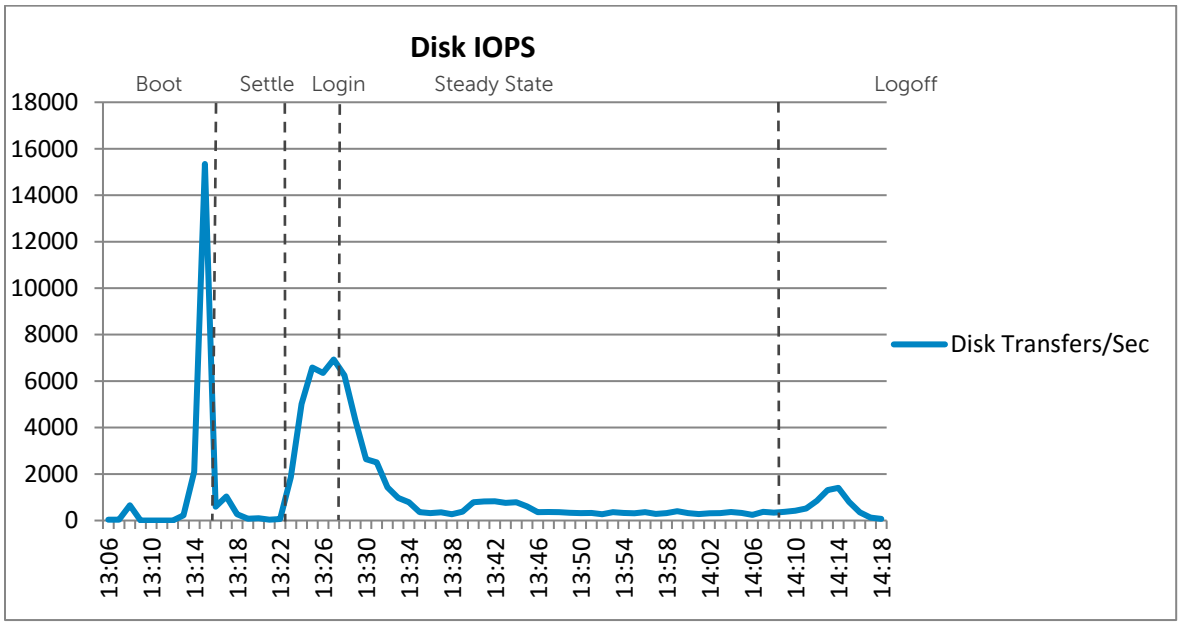


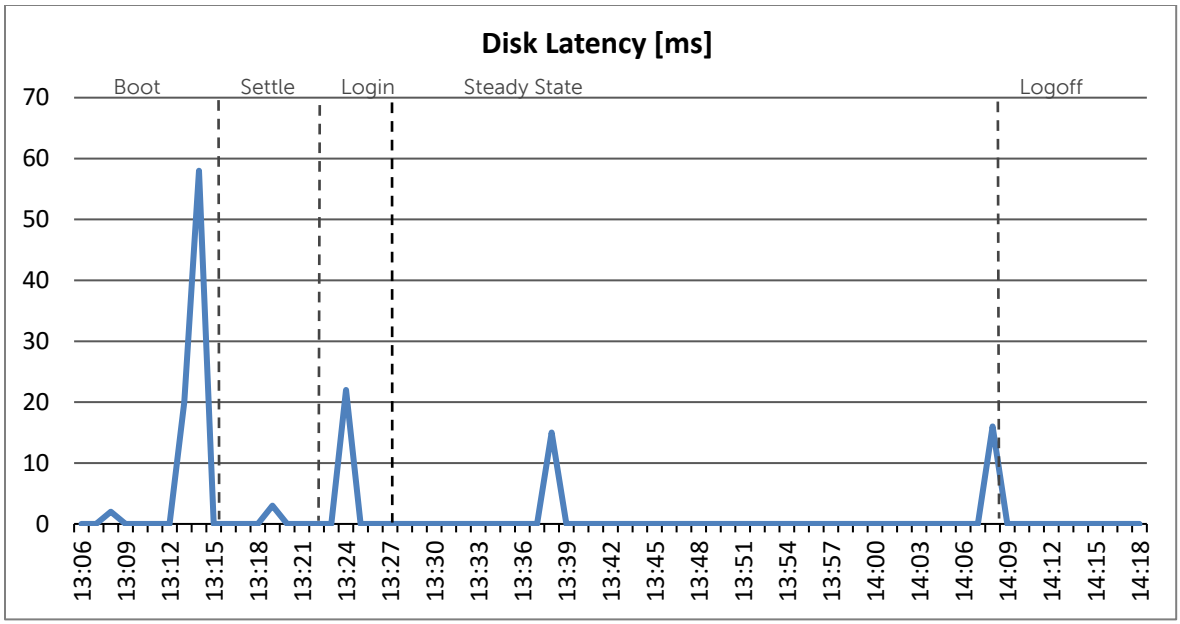
CPU usage for 32 clients is averaging 70 % for 32 clients for the profile M60-1Q. This is with no nvidia-smi GPU utilization collection running on the host.



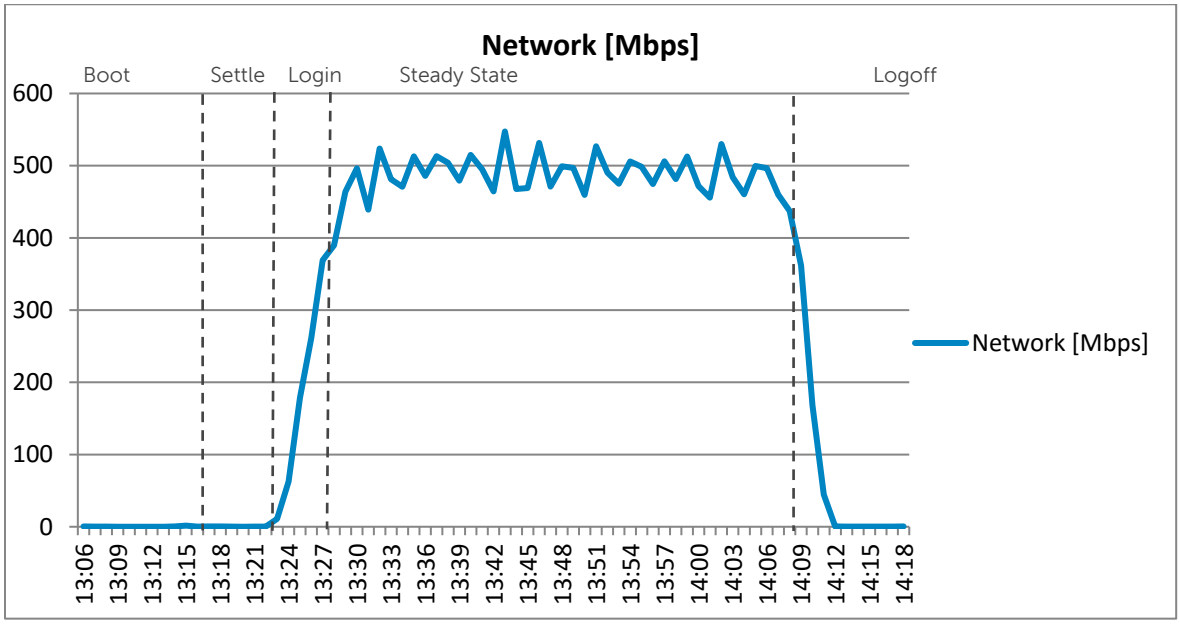


Host memory shows no bottle neck. For 32 clients with 4GB reserved RAM consumed memory should be 128GB + an overhead for vSphere services. On average the memory actively used by the clients is around half of the available memory. Swap Used [GB] and Balloon [GB] show 0 GB throughout the test.

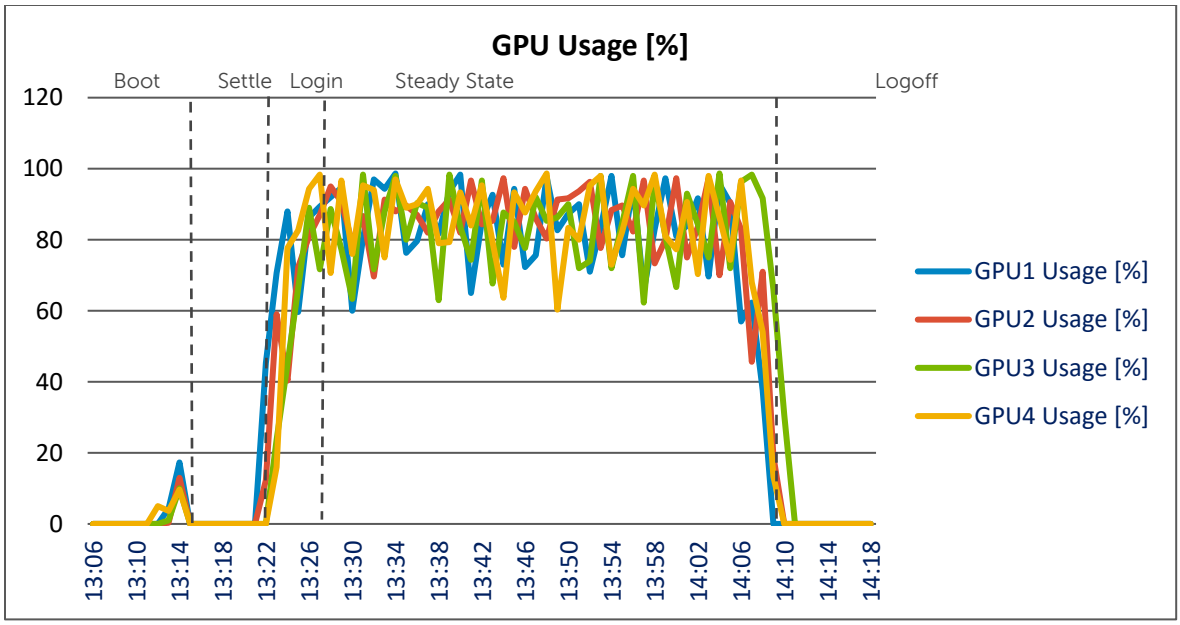




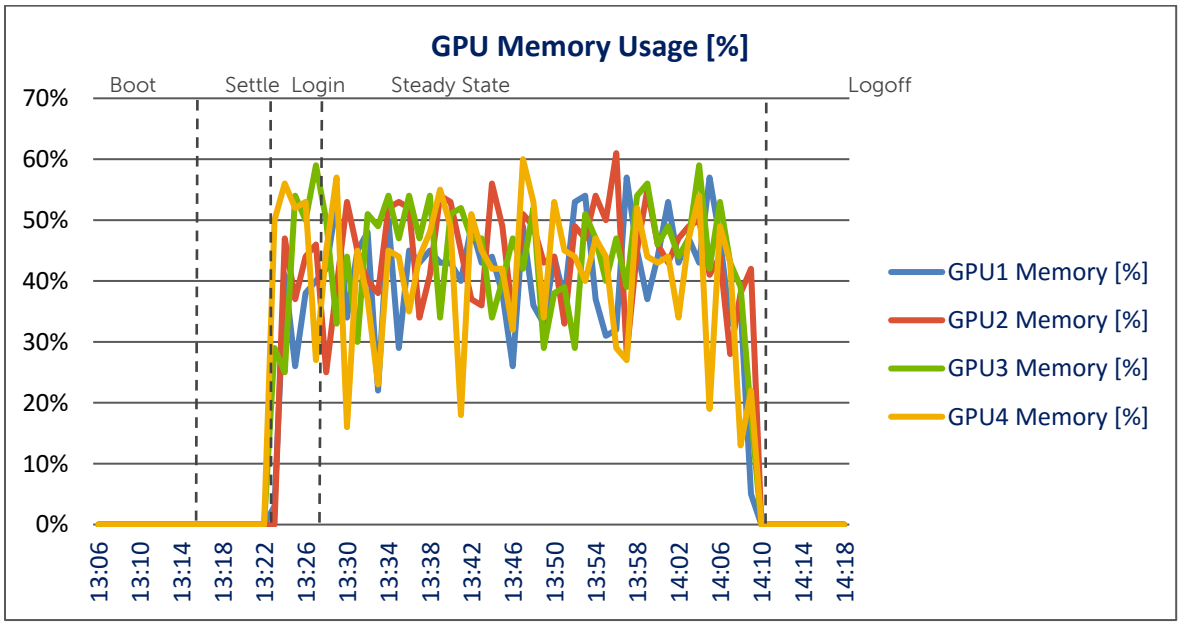
Highest disk activity was observed during the boot period. Tests were configured in such a way as to launch a session on a Client VM in 10 second intervals. After all clients have launched, processing values settle for the steady state values.



Network usage peak values are around the 550 Mbps mark.



During steady-state average GPU processor utilization shows 84%. GPUs are expected to run at full capacity. A low GPU load could point to issues with server CPUs not being able to handle the graphics load or other issues holding back the GPU. This is not the case here.



On average, 43% GPU memory was used during steady state testing. GPU memory utilization will depend on the graphics applications itself. SPECview module sw-03 could be considered a medium graphics load. High GPU memory usage should not be expected with this module.

SPECwpc Results

The tables below show the scores for the sw-03 module viewset category captured from VMs running on the indicated GPU.

GPU 1 [ID #6]			Total Score 0.96		
Test	**Ctgry**	**Run**	**Time**	**Baseline Score**	**Score**
sw-03	1	1	2/16/2017 2:03:59 PM	11.94	18.51
sw-03	2	1	2/16/2017 2:04:23 PM	9.56	16.92
sw-03	3	1	2/16/2017 2:04:45 PM	9.57	29.81
sw-03	4	1	2/16/2017 2:05:08 PM	10.06	18.55
sw-03	5	1	2/16/2017 2:06:03 PM	40.71	7.2
sw-03	6	1	2/16/2017 2:06:21 PM	34.78	25.37
sw-03	7	1	2/16/2017 2:06:38 PM	133.14	29.07
sw-03	8	1	2/16/2017 2:06:57 PM	88.49	20.89
sw-03	9	1	2/16/2017 2:07:14 PM	12.84	38
sw-03	10	1	2/16/2017 2:07:30 PM	53.59	49.2
sw-03	11	1	2/16/2017 2:07:46 PM	38.34	54

GPU 2 [ID #7]			Total Score 0.79		
Test	**Ctgry**	**Run**	**Time**	**Baseline Score**	**Score**
sw-03	1	1	2/16/2017 2:02:31 PM	11.94	18.48
sw-03	2	1	2/16/2017 2:02:54 PM	9.56	18.48
sw-03	3	1	2/16/2017 2:03:15 PM	9.57	36.87
sw-03	4	1	2/16/2017 2:03:38 PM	10.06	18.36
sw-03	5	1	2/16/2017 2:04:35 PM	40.71	7.11
sw-03	6	1	2/16/2017 2:04:55 PM	34.78	21.37
sw-03	7	1	2/16/2017 2:05:12 PM	133.14	24.4
sw-03	8	1	2/16/2017 2:05:37 PM	88.49	15.93
sw-03	9	1	2/16/2017 2:06:01 PM	12.84	16.95
sw-03	10	1	2/16/2017 2:06:17 PM	53.59	32.07
sw-03	11	1	2/16/2017 2:06:34 PM	38.34	35.07

GPU 3 [ID #84]	Total Score 1.0				
Test	**Ctrgy**	**Run**	**Time**	**Baseline Score**	**Score**
sw-03	1	1	2/16/2017 2:06:02 PM	11.94	15.85
sw-03	2	1	2/16/2017 2:06:30 PM	9.56	14.69
sw-03	3	1	2/16/2017 2:06:52 PM	9.57	20.44
sw-03	4	1	2/16/2017 2:07:19 PM	10.06	14.95
sw-03	5	1	2/16/2017 2:07:59 PM	40.71	9.68
sw-03	6	1	2/16/2017 2:08:17 PM	34.78	39.4
sw-03	7	1	2/16/2017 2:08:33 PM	133.14	41.49
sw-03	8	1	2/16/2017 2:08:50 PM	88.49	26.21
sw-03	9	1	2/16/2017 2:09:06 PM	12.84	43.4
sw-03	10	1	2/16/2017 2:09:22 PM	53.59	47.4
sw-03	11	1	2/16/2017 2:09:38 PM	38.34	52.6

GPU 4 [ID #85]	Total Score : 0.69				
Test	**Ctgry**	**Run**	**Time**	**Baseline Score**	**Score**
sw-03	1	1	2/16/2017 2:02:21 PM	11.94	19.71
sw-03	2	1	2/16/2017 2:02:49 PM	9.56	14.75
sw-03	3	1	2/16/2017 2:03:16 PM	9.57	16.78
sw-03	4	1	2/16/2017 2:03:40 PM	10.06	17.72
sw-03	5	1	2/16/2017 2:04:34 PM	40.71	7.32
sw-03	6	1	2/16/2017 2:04:54 PM	34.78	22.39
sw-03	7	1	2/16/2017 2:05:12 PM	133.14	26.18
sw-03	8	1	2/16/2017 2:05:36 PM	88.49	16.08
sw-03	9	1	2/16/2017 2:06:03 PM	12.84	14.56
sw-03	10	1	2/16/2017 2:06:20 PM	53.59	26.67
sw-03	11	1	2/16/2017 2:06:37 PM	38.34	28.31

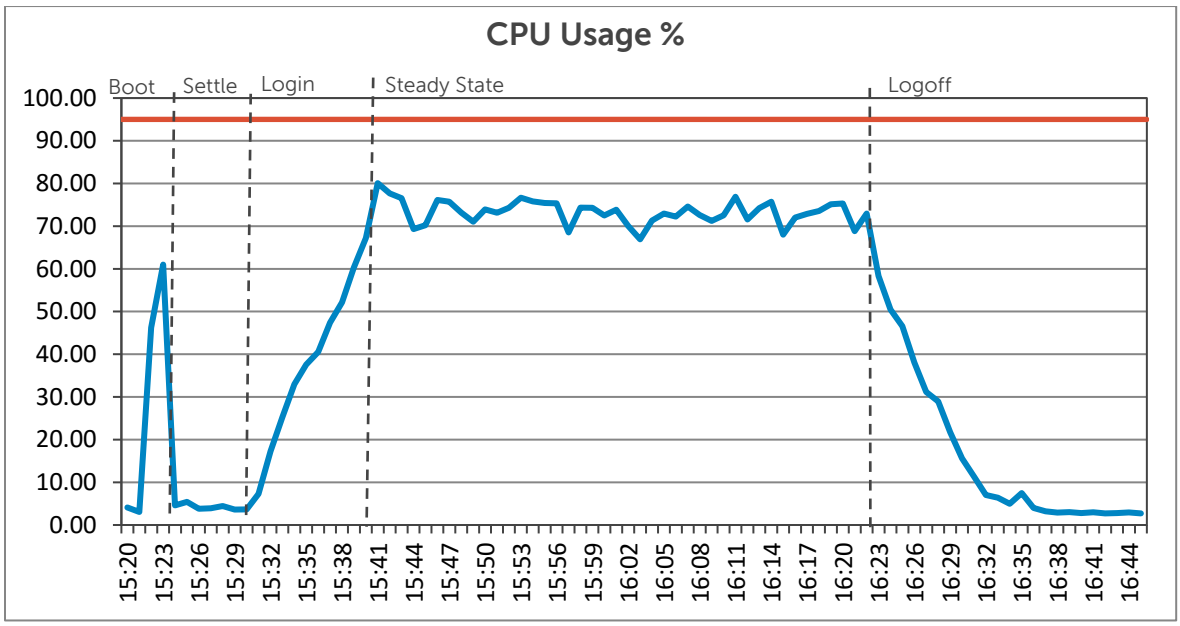
Score results and completion times show expected range for this configuration.

In general, completion times of all categories for each client VM lie within 1 sec range.
Total score usually ranges with 0.2 – 0.3 points. NVIDIA Shared Graphics can not guarantee a fixed performance value, this is because GPU resources are allocated on demand and might vary slighty during usage.

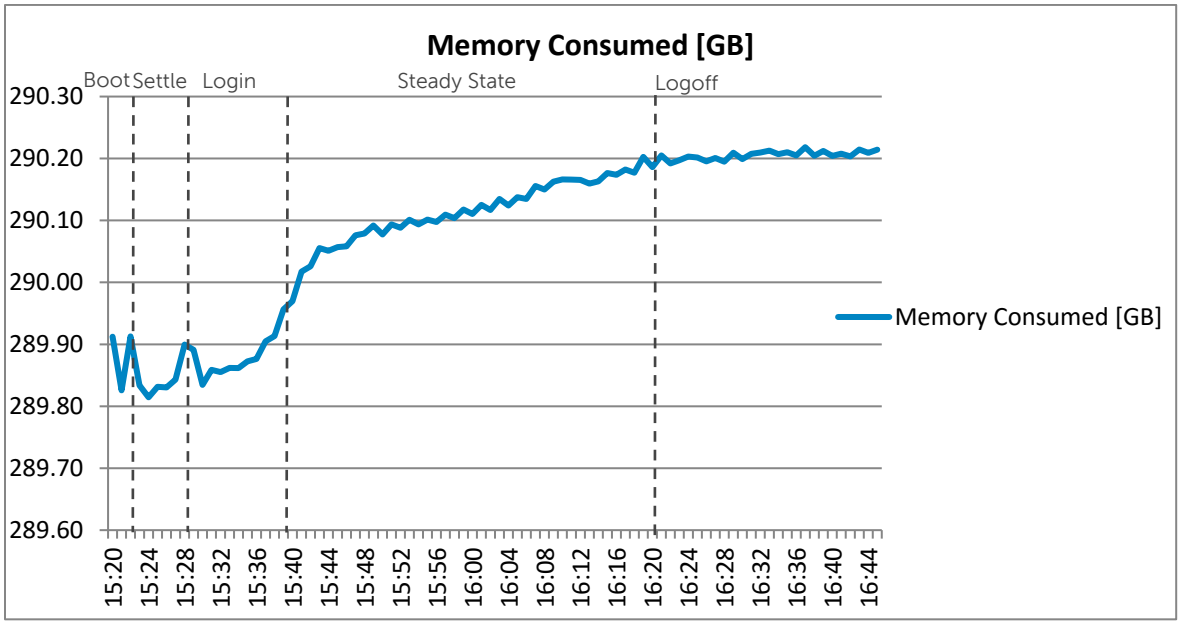
6.3.2 R730 High Density – M10 GPUs

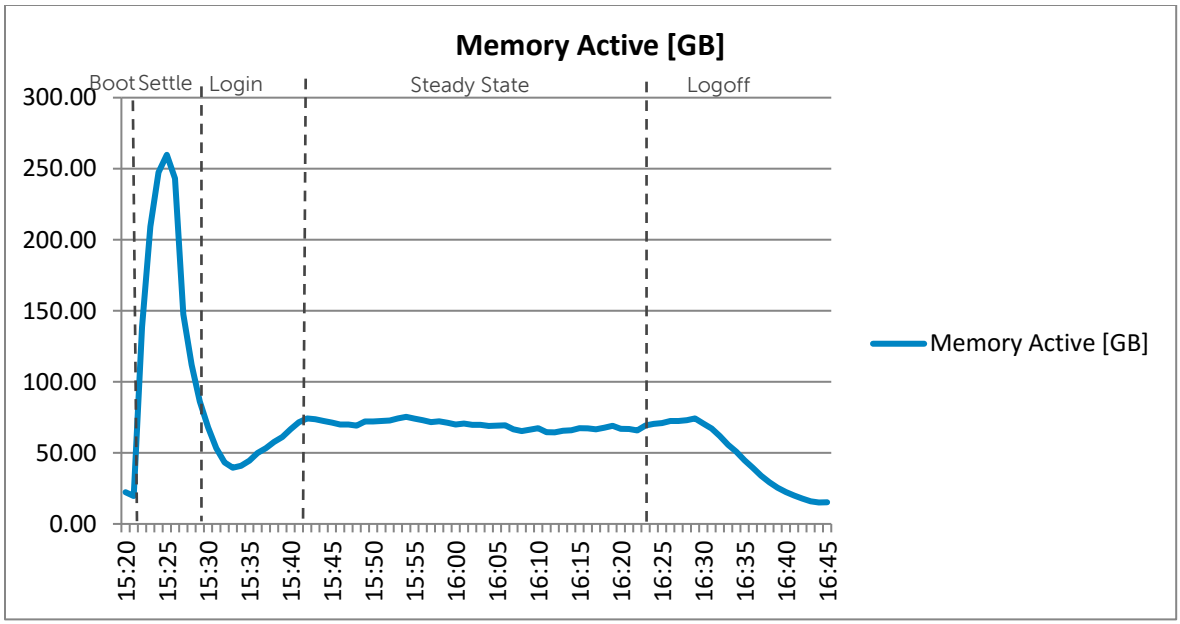
Refer to the Platform Configuration section for hardware configuration details.

6.3.2.1 Graphics LVSI Power + ProLibrary, 64 Users, ESXi 6.0 U2, Horizon 7

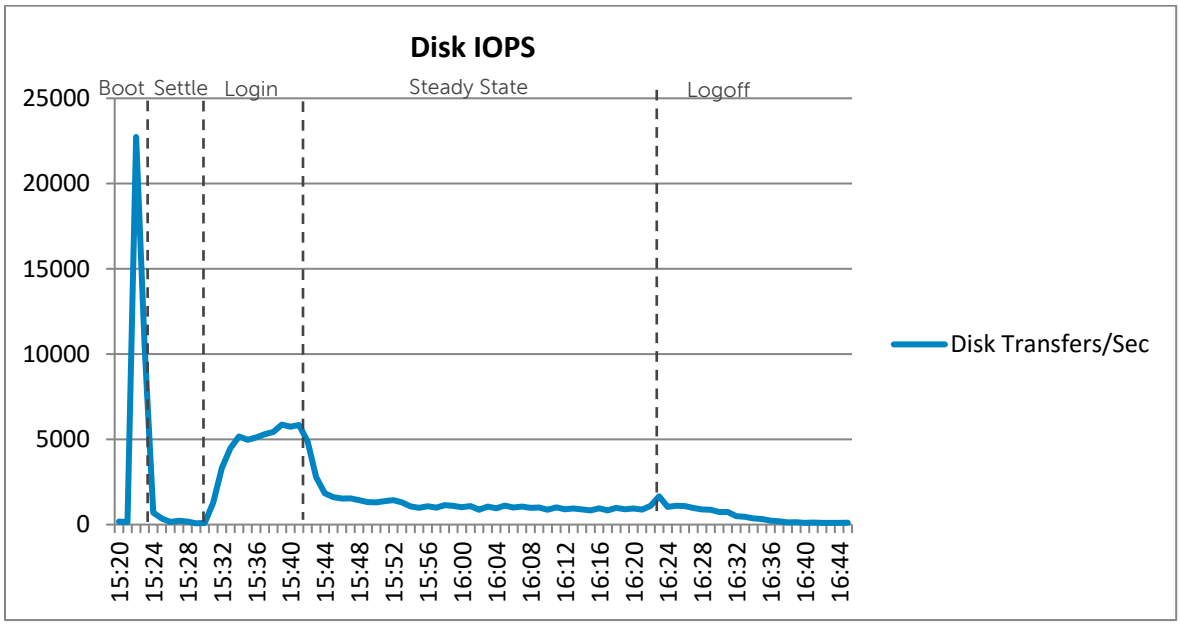


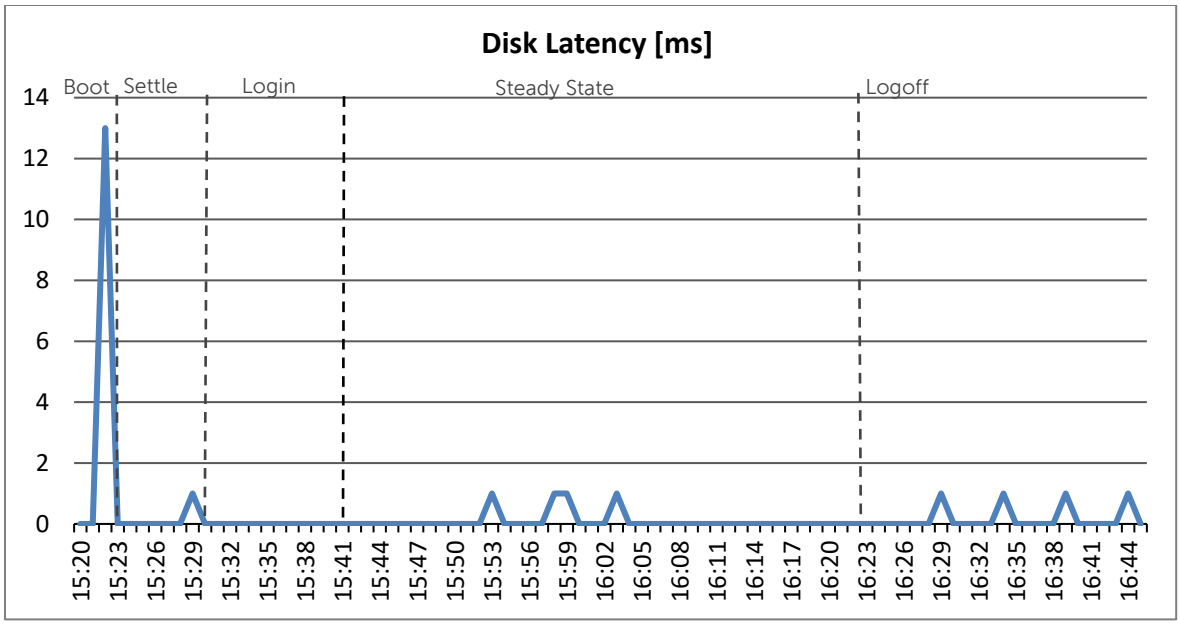
CPU usage for 64 clients is averaging 73% which is below the 95% threshold.



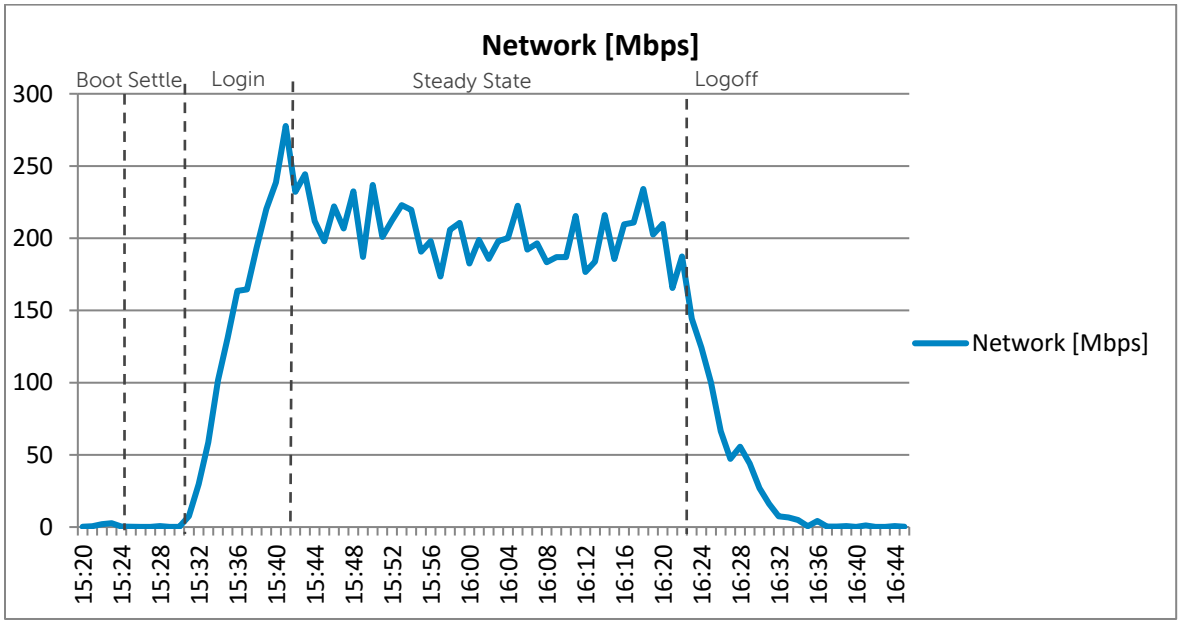


Host memory shows no bottle neck. For 64 clients with 4GB reserved RAM consumed memory should be 256GB + an overhead for vSphere services. On average the memory actively used by the clients is around half of the available memory. Swap Used [GB] and Balloon [GB] show 0 GB throughout the test.

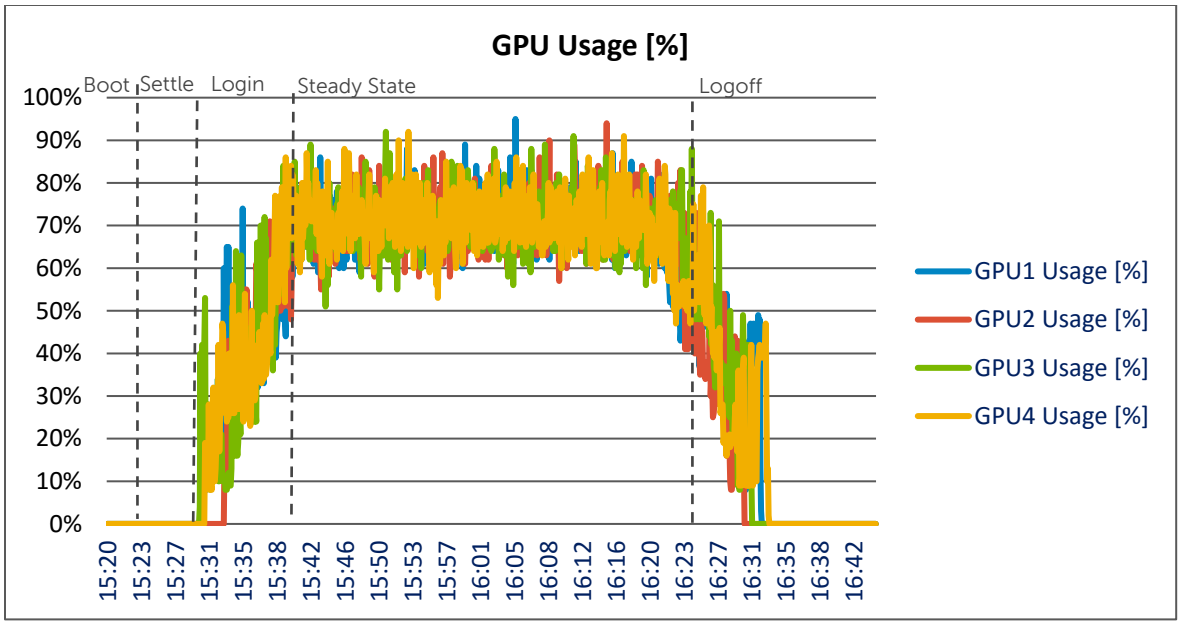




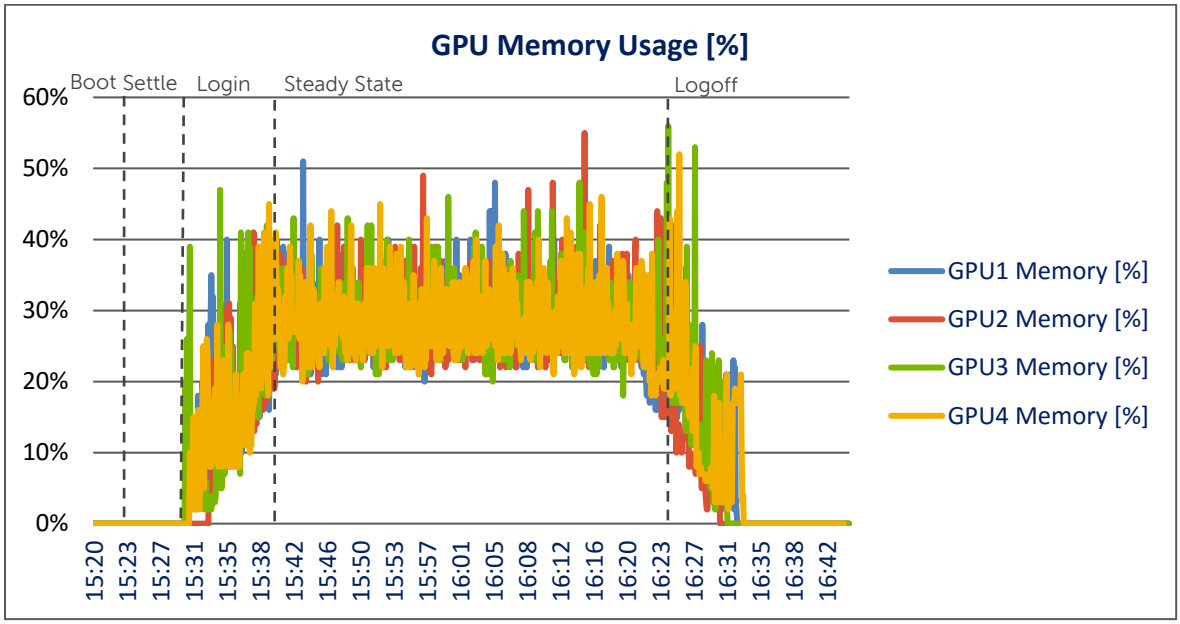
Highest disk activity was observed during the boot period. Tests were configured in such a way as to launch a session on a Client VM in 10 second intervals. After all clients have launched, processing values settle for the steady state values. Disk IOPS were observed to be still higher than the settled value at this point, completing the login process. During the indicated Steady State period, IOPS per user average 21. IOPS will eventually settle to a value of 15 IOPS per user.



Network usage peak values are around the 220 Mbps mark.



During steady-state average GPU processor utilization shows 70%. GPUs are expected to run at full capacity. A low GPU load could point to issues with server CPUs not being able to handle the graphics load or other issues holding back the GPU. This is not the case here.

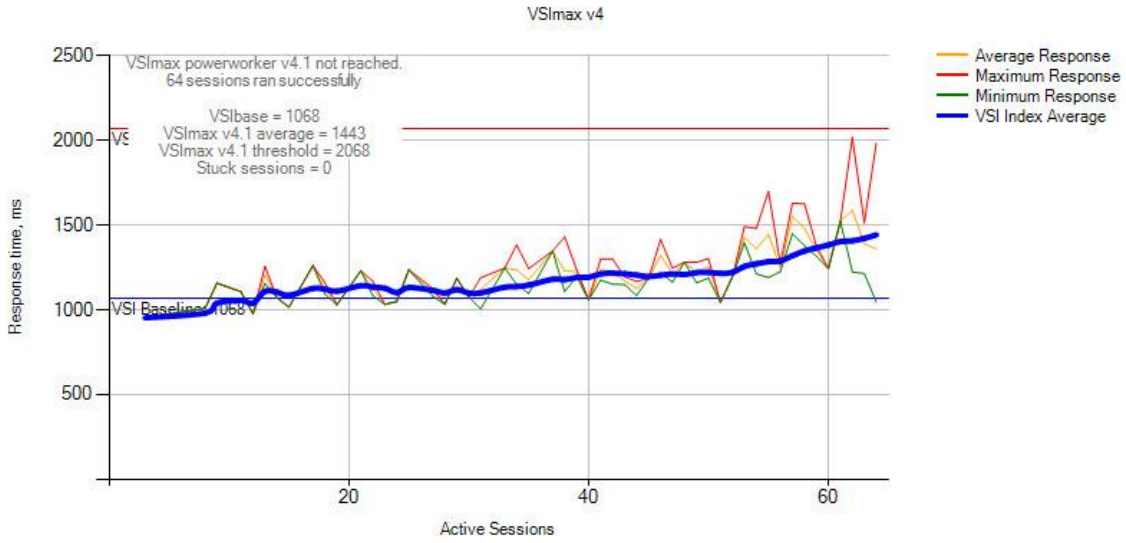


On average 28% GPU memory was used during steady state testing. GPU memory utilization will depend on the graphics applications itself. High GPU memory usage should not be expected with the workload.

NOTE: The server had two M10 graphic cards installed during testing; however, graphs from only one card are displayed since both performed similarly.

LoginVSI - VSI Max

The table below shows the results from the LoginVSI Analyzer.



Acknowledgements

Thank you to David Hulama of the Dell Wyse Technical Marketing team for his support and assistance with datacenter EUC programs at Dell.

Thank you to Gus Chavira, VMware Alliances Manager for CCC at Dell, for the design and development of the first release of the Dell Precision Appliance for Wyse and for his continued guidance and support of the program.

Thank you to Kristina Bako, Senior Systems Engineer in the Cloud Client Solutions Engineering Group at Dell, for her input and guidance for virtualized graphics and for validation testing of the Dell Precision Appliance for Wyse.

Thank you to Scott Stanford, Solutions Development Manager for the Dell Precision Appliance for Wyse program.

About the Authors

Jerry Van Blaricom is a Lead Architect in the Cloud Client Solutions Engineering Group at Dell. Jerry has extensive experience with the design and implementation of a broad range of enterprise systems and is focused on making Dell's virtualization offerings consistently best in class.

Peter Fine is the Chief Architect for datacenter EUC at Dell. Peter has extensive experience and expertise on the broader Microsoft, Citrix and VMware solutions software stacks as well as in enterprise virtualization, storage, networking and enterprise data center design.