Reference Architecture for Workloads using Lenovo ThinkAgile HX Series

Last update: 10 December 2020
Version 2.3

Provides a technical overview of Lenovo ThinkAgile HX Series appliances and certified nodes

Contains performance data and sizing recommendations

Shows variety of workloads that can be used in a hyper-converged environment

Explains reliability and performance features of hyper-converged appliances

Kenny Bain
Pawan Sharma
Chandrakandh Mouleeswaran
Dan Ionut Ghervase
Patrick Hartman

Frank Zimmer
Xiaotong Jiang
Srihari Angaluri
Markesha Parker
Cristian Ghetau

LENOVO PRESS

Table of Contents

1 Introduction 1

2 Technical overview of appliances 2

2.1 ThinkAgile HX series 2
2.2 Software components 5
2.3 Data network components 10
2.4 Hardware management network components 15
2.5 Reliability and performance features 16

3 Deployment models 20

3.1 SMB deployment model 20
3.2 Rack-scale deployment models 20

4 Citrix XenDesktop 24

4.1 Solution overview 24
4.2 Component model 25
4.3 Citrix XenDesktop provisioning 27
4.4 Management VMs 29
4.5 Graphics acceleration 35
4.6 Performance testing 36
4.7 Performance recommendations 38
4.8 Deployment ready solutions 42

5 Microsoft Exchange 43

5.1 Solution overview 43
5.2 Component model 44
5.3 30,000 Mailbox Exchange Performance with Hybrid Storage 46
5.4 60,000 Mailbox Exchange Performance with AF Storage 52
5.5 Exchange deployment best practices 58
5.6 Summary 59

6 Microsoft SQL Server 60
6.1 Solution overview 60
6.2 Component model 61
6.3 SQL Server deployment best practices 62
6.4 Performance test configuration 65
6.5 Performance test results 65
7 Red Hat OpenShift Container Platform 70
7.1 Solution Overview 70
7.2 Component Model 71
7.3 Operational model 74
7.4 Best practices 78
7.5 Deployment example 79
8 VMware Horizon 89
8.1 Solution overview 89
8.2 Component model 89
8.3 VMware Horizon provisioning 91
8.4 Management VMs 92
8.5 Graphics acceleration 93
8.6 Performance testing 94
8.7 Performance recommendations 95
8.8 Deployment ready solutions 98
9 VMware vCloud Suite 99
9.1 Solution Overview 99
9.2 Component model 100
9.3 Shared edge and compute cluster 105
9.4 Management cluster 106
9.5 Hybrid networking to public clouds 108
9.6 Systems management 109
9.7 Deployment example 115

Resources 121

1 Introduction

The intended audience for this document is technical IT architects, system administrators, and managers who are interested in executing workloads on the Lenovo ThinkAgile HX Series appliances and certified nodes.

ThinkAgile HX Series provides a hyper-converged infrastructure. Hyper-converged means incorporating multiple components like compute and storage into a single entity through software. A hyper-converged infrastructure seamlessly pools compute and storage to deliver high performance for the virtual workloads and provides flexibility to combine the local storage using a distributed file system to eliminate shared storage such as SAN or NAS. These factors make the solution cost effective without compromising the performance.

Chapter 2 provides a technical overview of ThinkAgile HX Series and explains why the combination of Lenovo servers and Nutanix software provides best of breed system performance and reliability. Chapter 3 provides some deployment models.

Each of the subsequent chapters in the document describes a particular virtualized workload and provides recommendations on what appliance model to use and how to size the appliance to that workload. Some best practice recommendations are also listed. ThinkAgile HX Series appliances and certified nodes are not limited to just the workloads described in this reference architecture and can execute any virtualized workload on the supported hypervisors.

This Reference Architecture describes seven workloads:

- Citrix XenDesktop
- Microsoft Exchange
- Microsoft SQL Server
- Red Hat OpenShift Container Platform
- SAP Business Applications
- VMware Horizon
- VMware vCloud Suite

2 Technical overview of appliances

This chapter provides an overview of the ThinkAgile HX Series appliances and certified nodes including the associated software, systems management, and networking. The last section provides an overview of the performance and reliability features.

2.1 ThinkAgile HX series

Lenovo ThinkAgile HX Series appliances and certified nodes are designed to help you simplify IT infrastructure, reduce costs, and accelerate time to value. These hyper-converged appliances from Lenovo combine industry-leading hyper-convergence software from Nutanix with Lenovo enterprise platforms. Several common uses are:

- Enterprise workloads
- Private and hybrid clouds
- Remote office and branch office (ROBO)
- Server virtualization
- Virtual desktop infrastructure (VDI)
- Small-medium business (SMB) workloads

Starting with as few as three nodes to keep your acquisition costs down, the Lenovo ThinkAgile HX Series appliances and certified nodes are capable of immense scalability as your needs grow.

Lenovo ThinkAgile HX Series appliances and certified nodes are available in five families that can be tailored to your needs:

- Lenovo ThinkAgile HX1000 Series: optimized for ROBO environments
- Lenovo ThinkAgile HX2000 Series: optimized for SMB environments
- Lenovo ThinkAgile HX3000 Series: optimized for compute-heavy environments
- Lenovo ThinkAgile HX5000 Series: optimized for storage-heavy workloads
- Lenovo ThinkAgile HX7000 Series: optimized for high-performance workloads

Table 1 shows the similarities and differences between ThinkAgile HX Series appliances and certified nodes.

Table 1: Comparison of ThinkAgile HX Series appliances and certified nodes

Feature	HX Series Appliances	HX Series certified nodes
Validated and integrated hardware and firmware	Yes	Yes
Certified and preloaded with Nutanix software	Yes	Yes
Includes Nutanix licenses	Yes	No
ThinkAgile Advantage Single Point of Support for quick 24/7 problem reporting and resolution	Yes	Yes
Includes deployment services	Optional	Optional
Supports ThinkAgile HX2000 Series	Yes	No

For more information about the system specifications and supported configurations, refer to the product guides for the Lenovo ThinkAgile HX Series appliances and certified nodes based on the Intel Xeon Scalable processor. For appliances see:

- Lenovo ThinkAgile HX1000 Series: lenovopress.com/lp0726
- Lenovo ThinkAgile HX2000 Series: lenovopress.com/lp0727
- Lenovo ThinkAgile HX3000 Series: lenovopress.com/lp0728
- Lenovo ThinkAgile HX5500 Series: lenovopress.com/lp0729
- Lenovo ThinkAgile HX7500 Series: lenovopress.com/lp0730
- Lenovo ThinkAgile HX7800 Series: lenovopress.com/lp0950

For certified nodes see:

- Lenovo ThinkAgile HX1001 Series: lenovopress.com/lp0887
- Lenovo ThinkAgile HX3001 Series: lenovopress.com/lp0888
- Lenovo ThinkAgile HX5501 Series: lenovopress.com/lp0889
- Lenovo ThinkAgile HX7501 Series: lenovopress.com/lp0890
- Lenovo ThinkAgile HX7800 Series: lenovopress.com/lp0951

The diagrams below show the Intel Xeon Scalable processor-based ThinkAgile HX Series appliances and certified nodes.

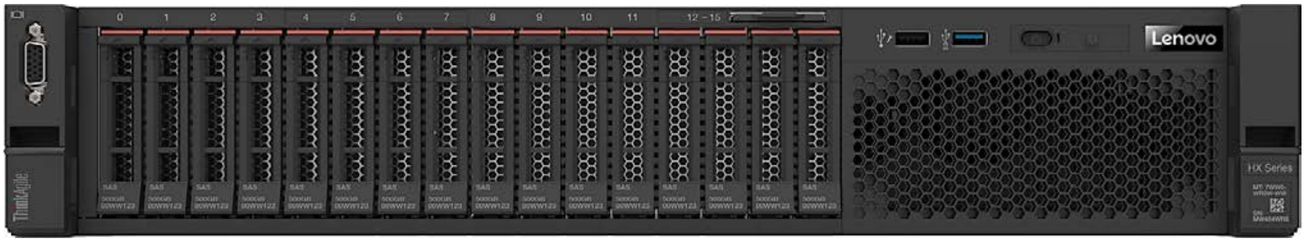
HX1320 or HX1321:

HX2320-E:

HX2720-E:

HX3320 or HX3321:

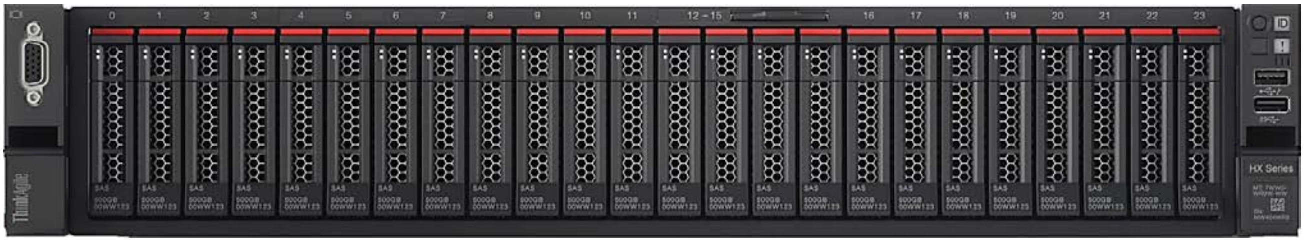
HX3520-G or HX3521-G:



HX3720 or HX3721:

HX1520-R, HX1521-R, HX5520, HX5521, HX5520-C, or HX5521-C:

HX7520 or HX7521:



HX7820 or HX7821:

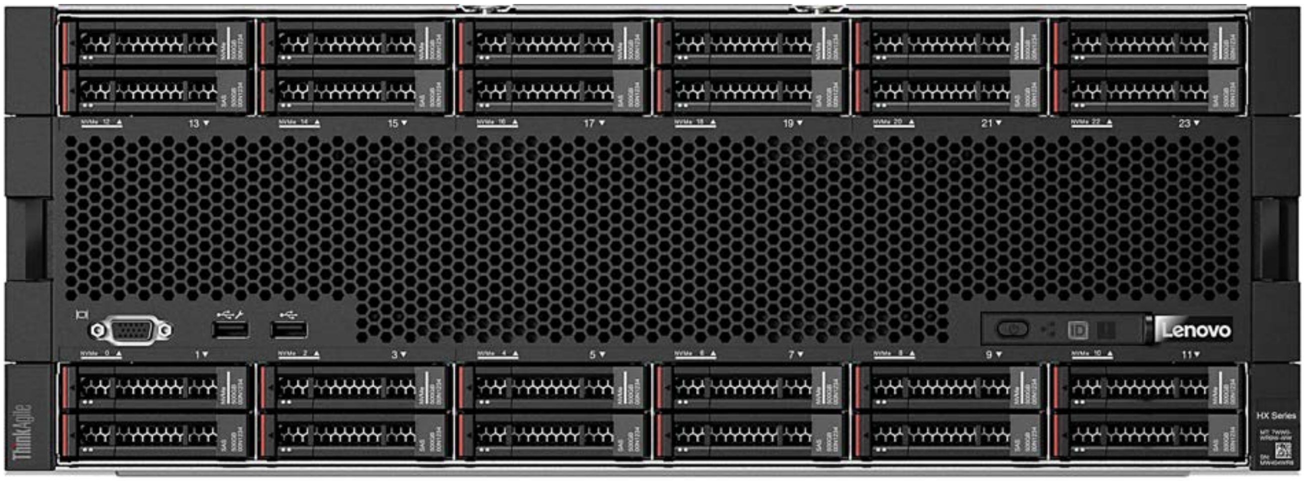


Table 2 provides a summary of the default configurations for the ThinkAgile HX Series appliances and certified nodes (including all-flash variations).

Table 2: Default configurations for ThinkAgile HX Series

Model	Intel Xeon processor	Memory (RDIMMs)	Storage controller	SSDs	HDDs	NIC
HX132x	1x 4110 8C	96GB (6x 16GB)	1x 430-8i	3.84TB (2x 1.92TB)	8TB (2x 4TB)	2x 10GbE RJ-45
HX152x-R	1x 4114 10C	192GB (12x 16GB)	1x 430-16i	3.84TB (2x 1.92TB)	60TB (10x 6TB)	2x 10GbE RJ-45
HX2320-E	2x 4108 8C	192GB (12x 16GB)	1x 430-8i	1.92TB (1x 1.92TB)	6TB (6x 1TB)	2x 10GbE RJ-45
HX2720-E	1x 4108 8C	192GB (12x 16GB)	1x 430-8i	1.92TB (1x 1.92TB)	4TB (4x 1TB)	2x 10GbE SFP+
HX332x Hybrid	2x 6136 12C	384GB (12x 32GB)	1x 430-16i	3.84TB (2x 1.92TB)	6TB (6x 1TB)	2x 10GbE SFP+
HX332x All Flash	2x 6136 12C	384GB (12x 32GB)	1x 430-16i	11.52TB (6x 1.92TB)	N/A	2x 10GbE SFP+
HX332x SAP HANA	2x 6136 12C	384GB (12x 32GB)	1x 430-16i	7.68TB (8x 960GB)	N/A	4x 10GbE SFP+
HX352x-G Hybrid	2x 6126 12C	384GB (12x 32GB)	1x 430-16i	3.84TB (2x 1.92TB)	12TB (12x 1TB)	4x 10GbE SFP+
HX352x-G All Flash	2x 6126 12C	384GB (12x 32GB)	1x 430-16i	23.04TB (12x 1.92TB)	N/A	4x 10GbE SFP+
HX372x Hybrid	2x 6126 12C	384GB (12x 32GB)	1x 430-8i	3.84TB (2x 1.92TB)	8TB (4x 2TB)	2x 10GbE SFP+
HX372x All Flash	2x 6126 12C	384GB (12x 32GB)	1x 430-8i	7.68TB (4x 1.92TB)	N/A	2x 10GbE SFP+
HX552x	2x 6140 18C	384GB (12x 32GB)	1x 430-16i	3.84TB (2x 1.92TB)	60TB (10x 6TB)	2x 10GbE SFP+
HX552x-C	1x 4110 8C	64GB (4x 16GB)	1x 430-16i	3.84TB (2x 1.92TB)	60TB (10x 6TB)	2x 10GbE SFP+
HX752x Hybrid	2x 8164 26C	768GB (24x 32GB)	3x 430-8i	7.68TB (4x 1.92TB)	32TB (16x 2TB)	4x 10GbE SFP+
HX752x All Flash	2x 8164 26C	768GB (24x 32GB)	3x 430-8i	34.56TB (18x 1.92TB)	N/A	4x 10GbE SFP+
HX752x SAP HANA	2x 8164 26C	768GB (24x 32GB)	3x 430-8i	15.36TB (8x 1.92TB)	N/A	4x 10GbE SFP+
HX782x Hybrid	2x 8180 28C	1536GB (24x 64GB)	2x 430-16i	28.8TB (12x 2.4TB)	N/A	4x 10GbE RJ45
HX782x All Flash	2x 8180 28C	1536GB (24x 64GB)	2x 430-16i	23.04TB (12x 1.92TB)	N/A	4x 10GbE RJ45
HX782x SAP HANA	2x 8180 28C	1536GB (24x 64GB)	2x 430-16i	38.4TB (10x 3.84TB)	N/A	4x 10GbE QSFP28

For best recipes of supported firmware and software, please see: https://datacentersupport.lenovo.com/de/en/solutions/ht505413.

2.2 Software components

This section gives an overview of the software components used in the solution.

2.2.1 Hypervisor

The ThinkAgile HX Series appliances and certified nodes (generally) support the following hypervisors:

- Nutanix Acropolis Hypervisor based on KVM (AHV)
- VMware ESXi 6.0 U3
- VMware ESXi 6.5 U2

The HX1520-R, HX5520-C, HX7820, and all SAP HANA models support only the following hypervisor:

- Nutanix Acropolis Hypervisor based on KVM (AHV)

The HX Series appliances come standard with the hypervisor preloaded in the factory. This software is optional for the ThinkAgile HX Series certified nodes.

2.2.2 Lenovo XClarity Administrator

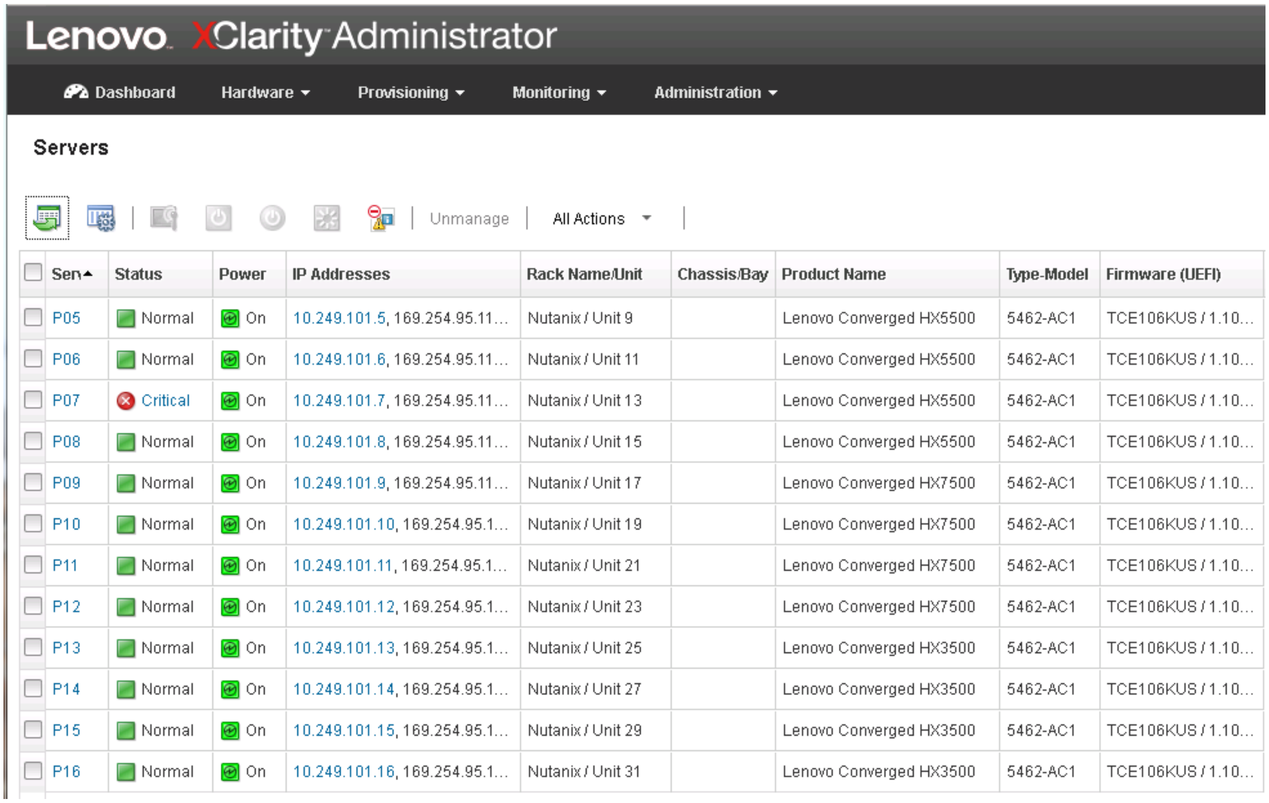
Lenovo XClarity Administrator is a centralized systems management solution that helps administrators deliver infrastructure faster. This solution integrates easily with Lenovo servers, ThinkAgile HX Series appliances and certified nodes, and Flex System, providing automated agent-less discovery, monitoring, firmware updates, and configuration management.

Lenovo XClarity Pro goes one step further and provides entitlement to additional functions such as XClarity Integrators for Microsoft System Center and VMware vCenter, XClarity Administrator Configuration Patterns and Service and Support.

Lenovo XClarity Administrator is an optional software component and can be used to manage firmware upgrades outside of the Nutanix Prism web console. Note that XClarity should not be used to install hypervisors and Nutanix Foundation should be used instead.

Lenovo XClarity Administrator is provided as a virtual appliance that can be quickly imported into a virtualized environment. XClarity can either be installed on a separate server or a server within a Nutanix cluster providing that the hardware management network with the server IMMs is routable from the server hosting the XClarity VM.

Figure 1 shows the Lenovo XClarity administrator interface.



Servers

□	Serv▲	Status	Power	IP Addresses	Rack Name/Unit	Chassis/Bay	Product Name	Type-Model	Firmware (UEFI)
□	P05	Normal	On	10.249.101.5, 169.254.95.11…	Nutanix / Unit 9		Lenovo Converged HX5500	5462-AC1	TCE106KUS / 1.10…
□	P06	Normal	On	10.249.101.6, 169.254.95.11…	Nutanix / Unit 11		Lenovo Converged HX5500	5462-AC1	TCE106KUS / 1.10…
□	P07	Critical	On	10.249.101.7, 169.254.95.11…	Nutanix / Unit 13		Lenovo Converged HX5500	5462-AC1	TCE106KUS / 1.10…
□	P08	Normal	On	10.249.101.8, 169.254.95.11…	Nutanix / Unit 15		Lenovo Converged HX5500	5462-AC1	TCE106KUS / 1.10…
□	P09	Normal	On	10.249.101.9, 169.254.95.11…	Nutanix / Unit 17		Lenovo Converged HX7500	5462-AC1	TCE106KUS / 1.10…
□	P10	Normal	On	10.249.101.10, 169.254.95.1…	Nutanix / Unit 19		Lenovo Converged HX7500	5462-AC1	TCE106KUS / 1.10…
□	P11	Normal	On	10.249.101.11, 169.254.95.1…	Nutanix / Unit 21		Lenovo Converged HX7500	5462-AC1	TCE106KUS / 1.10…
□	P12	Normal	On	10.249.101.12, 169.254.95.1…	Nutanix / Unit 23		Lenovo Converged HX7500	5462-AC1	TCE106KUS / 1.10…
□	P13	Normal	On	10.249.101.13, 169.254.95.1…	Nutanix / Unit 25		Lenovo Converged HX3500	5462-AC1	TCE106KUS / 1.10…
□	P14	Normal	On	10.249.101.14, 169.254.95.1…	Nutanix / Unit 27		Lenovo Converged HX3500	5462-AC1	TCE106KUS / 1.10…
□	P15	Normal	On	10.249.101.15, 169.254.95.1…	Nutanix / Unit 29		Lenovo Converged HX3500	5462-AC1	TCE106KUS / 1.10…
□	P16	Normal	On	10.249.101.16, 169.254.95.1…	Nutanix / Unit 31		Lenovo Converged HX3500	5462-AC1	TCE106KUS / 1.10…

Figure 1: XClarity Administrator interface

2.2.3 Nutanix Prism

Nutanix Prism gives administrators a simple and elegant way to manage virtual environments. Powered by advanced data analytics and heuristics, Prism simplifies and streamlines common workflows within a data center.

Nutanix Prism is a part of the Nutanix software preloaded on the appliances and offers the following features:

- Single point of control
  - Accelerates enterprise-wide deployment
  - Manages capacity centrally
  - Adds nodes in minutes
  - Supports non-disruptive software upgrades with zero downtime
  - Integrates with REST APIs and PowerShell
- Monitoring and alerting
  - Tracks infrastructure utilization (storage, processor, memory)
  - Centrally monitors multiple clusters across multiple sites
  - Monitors per virtual machine (VM) performance and resource usage
  - Checks system health
  - Generates alerts and notifications

- Integrated data protection
  - Offers customizable RPO/RTO and retention policies
  - Supports configurable per-VM replication (1:1, 1:many and many:1)
  - Provides efficient VM recovery
  - Deploys affordable data recovery (DR) and backup to the cloud
- Diagnostics and troubleshooting
  - Provides time-based historical views of VM activity
  - Performs proactive alert analysis
  - Correlates alerts and events to quickly diagnose issues
  - Generates actionable alerts and reduces resolution times
  - Analyzes trending patterns for accurate capacity planning

2.2.4 ThinkAgile Network Orchestrator for Nutanix

The Lenovo® ThinkAgile™ Network Orchestrator is a unique feature of the Lenovo RackSwitch CNOS (Cloud Network OS) network switch firmware that automatically provisions the switches as needed, on-the-fly, to support changes in the virtual network, such as the creation, moving, and shutdown of virtual machines, as well as manipulation of guest virtual machines on VLANs. These tasks are performed dynamically in response to PRISM commands, with the switches configured to detect the changes and act upon them. This software capability simplifies the server administrator's tasks by reducing the need to provision the switches, reducing maintenance windows, reducing human error, and dramatically saving time and administrative costs. The switches learn of changes in the environment from PRISM and dynamically modify their configurations as needed.

The ThinkAgile Network Orchestrator capability is supported with the CNOS version 10.3.2.0 or higher and the Acropolis hypervisor AOS version 5.0.2 and higher. See the following paper for more details:

lenovopress.com/lp0604-thinkagile-network-orchestrator-for-nutanix

2.2.5 Nutanix Foundation

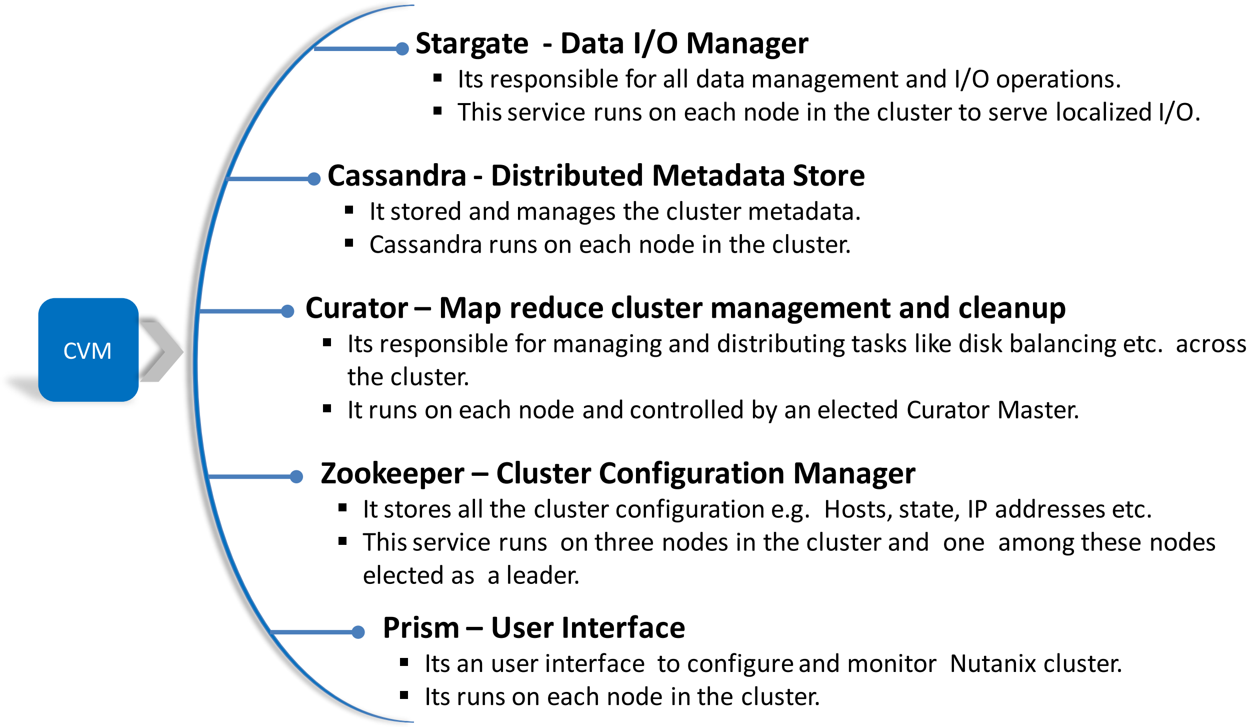
Nutanix Foundation is a separate utility that you use to orchestrate the installation of hypervisors and Nutanix software on one or more nodes. The maximum number of nodes that can be deployed at one time is 20.

Foundation is available both as a stand-alone VM and also integrated into the CVM. Because CVM is pre-installed in the factory, the CVM integration of Foundation simplifies the deployment and cluster creation of new servers delivered from the factory.

The dual M.2 boot drives must be configured as a RAID 1 mirrored array for installation to be successful.

2.2.6 Nutanix Controller VM

The Nutanix Controller VM (CVM) is the key to hyper-converged capability and each node in a cluster has its own instance. Figure 2 shows the main components of the CVM.



Stargate - Data I/O Manager

- Its responsible for all data management and I/O operations.
- This service runs on each node in the cluster to serve localized I/O.

Cassandra - Distributed Metadata Store

- It stored and manages the cluster metadata.
- Cassandra runs on each node in the cluster.

Curator – Map reduce cluster management and cleanup

- Its responsible for managing and distributing tasks like disk balancing etc. across the cluster.
- It runs on each node and controlled by an elected Curator Master.

Zookeeper – Cluster Configuration Manager

- It stores all the cluster configuration e.g. Hosts, state, IP addresses etc.
- This service runs on three nodes in the cluster and one among these nodes elected as a leader.

Prism – User Interface

- Its an user interface to configure and monitor Nutanix cluster.
- Its runs on each node in the cluster.

Figure 2: Controller VM components

The CVM works as interface between the storage and hypervisor to manage all I/O operations for the hypervisor and user VMs running on the nodes as shown in Figure 3.

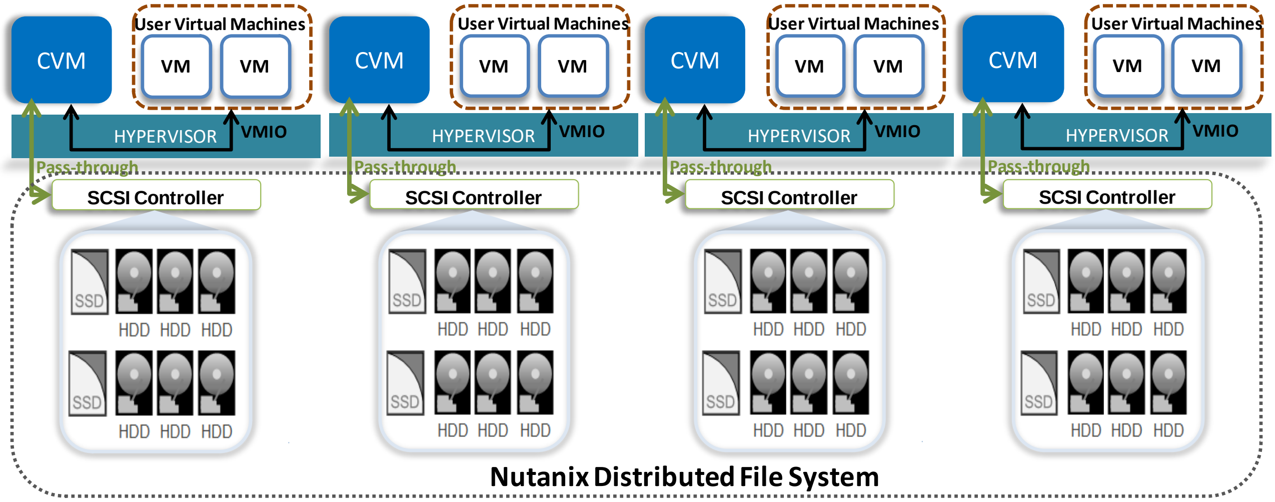


Figure 3: CVM interaction with Hypervisor and User VMs

CVM virtualizes all the local storage attached to each node in a cluster and presents it as centralized storage array using Nutanix Distributed File System (NDFS). All I/O operations are handled locally to provide the highest performance. See section 2.5 for more details on the performance features of NDFS.

2.3 Data network components

The data network is the fabric that carries all inter-node storage I/O traffic for the shared Lenovo HX distributed file system, in addition to the user data traffic via the virtual Network Interface Cards (NICs) exposed through the hypervisor to the virtual machines.

Each HX Series appliance contains between zero and two dual-port 10GbE network adapters as well as 4 on-board 1GbE ports. The hypervisors are configured by the Nutanix software so that the fastest network ports on the appliance are pooled for the data network. The hypervisor VM management network should use the same network. Because all of the network ports are pooled, each appliance only needs two network IP addresses; one for the hypervisor and one for the Nutanix CVM. These IP addresses should be all on the same subnet.

All storage I/O for virtual machines (VMs) running on a HX Series appliance node is handled by the hypervisor on a dedicated private network. The I/O request is handled by the hypervisor, which then forwards the request to the private IP on the local controller VM (CVM). The CVM then performs the remote data replication with other nodes in the cluster using its external IP address. In most cases, read request traffic is served locally and does not enter the data network. This means that the only traffic in the public data network is remote replication traffic and VM network I/O (i.e. user data). In some cases, the CVM will forward requests to other CVMs in the cluster, such as if a CVM is down or data is remote. Also, cluster-wide tasks, such as disk balancing, temporarily generate I/O traffic on the data network.

For more information on the network architecture see nutanixbible.com.

2.3.1 Data network switches

The following Lenovo 10GbE TOR switches are recommended for use in a HX Series cluster:

- Lenovo ThinkSystem NE1032 RackSwitch
- Lenovo RackSwitch G8272
- Lenovo ThinkSystem NE1072T RackSwitch
- Lenovo ThinkSystem NE2572 RackSwitch

Lenovo ThinkSystem NE1032 RackSwitch

The Lenovo ThinkSystem NE1032 RackSwitch (as shown in Figure 4) is a 1U rack-mount 10 Gb Ethernet switch that delivers lossless, low-latency performance with feature-rich design that supports virtualization, Converged Enhanced Ethernet (CEE), high availability, and enterprise class Layer 2 and Layer 3 functionality. The switch delivers line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data.

The NE1032 RackSwitch has 32x SFP+ ports that support 1 GbE and 10 GbE optical transceivers, active optical cables (AOCs), and direct attach copper (DAC) cables. The switch helps consolidate server and storage networks into a single fabric, and it is an ideal choice for virtualization, cloud, and enterprise workload solutions.

Figure 4: Lenovo ThinkSystem NE1032 RackSwitch

For more information, see this website: lenovopress.com/lp0605

Lenovo RackSwitch G8272

The Lenovo RackSwitch G8272 uses 10Gb SFP+ and 40Gb QSFP+ Ethernet technology and is specifically designed for the data center. It is an enterprise class Layer 2 and Layer 3 full featured switch that delivers line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data. Large data center-grade buffers help keep traffic moving, while the hot-swap redundant power supplies and fans (along with numerous high-availability features) help provide high availability for business sensitive traffic.

The RackSwitch G8272 (shown in Figure 5), is ideal for latency sensitive applications, such as high-performance computing clusters and financial applications. In addition to the 10 Gb Ethernet (GbE) and 40 GbE connections, the G8272 can use 1 GbE connections.

Figure 5: Lenovo RackSwitch G8272

For more information, see this website: lenovopress.com/tips1267

Lenovo ThinkSystem NE1072T RackSwitch

The Lenovo ThinkSystem NE1072T RackSwitch that uses 10GBASE-T and 40 Gb QSFP+ Ethernet technology is specifically designed for the data center. It is ideal for today's big data, cloud, and enterprise workload solutions. It is an enterprise class Layer 2 and Layer 3 full featured switch that delivers line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data. Large data center-grade buffers help keep traffic moving, while the hot-swap redundant power supplies and fans (along with numerous high-availability features) help provide high availability for business sensitive traffic.

The NE1072T RackSwitch (as shown in Figure 6) has 48x 1/10 Gb Ethernet (RJ-45) fixed ports and 6x QSFP+ ports that support 40 GbE optical transceivers, active optical cables (AOCs), and direct attach copper (DAC) cables. The QSFP+ ports can also be split out into four 10 GbE ports by using QSFP+ to 4x SFP+ DAC or active optical breakout cables.

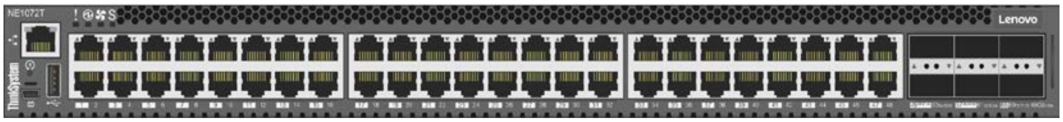


Figure 6: Lenovo ThinkSystem NE1072T RackSwitch

For more information, see this website: lenovopress.com/lp0607

Lenovo ThinkSystem NE2572 RackSwitch

The Lenovo ThinkSystem NE2572 RackSwitch is designed for the data center and provides 10 Gb/25 Gb Ethernet connectivity with 40 Gb/100 Gb Ethernet upstream links. It is ideal for big data, cloud, and enterprise workload solutions. It is an enterprise class Layer 2 and Layer 3 full featured switch that delivers line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data. Large data center-grade buffers help keep traffic moving, while the hot-swap redundant power supplies and fans (along with numerous high-availability software features) help provide high availability for business sensitive traffic.

The NE2572 RackSwitch (as shown in Figure 7) has 48x SFP28/SFP+ ports that support 10 GbE SFP+ and 25 GbE SFP28 optical transceivers, active optical cables (AOCs), and direct attach copper (DAC) cables. The switch also offers 6x QSFP28/QSFP+ ports that support 40 GbE QSFP+ and 100 GbE QSFP28 optical transceivers, active optical cables (AOCs), and direct attach copper (DAC) cables. The QSFP28/QSFP+ ports can also be split out into four 10 GbE (for 40 GbE QSFP+) or 25 GbE (for 100 GbE QSFP28) connections by using breakout cables.

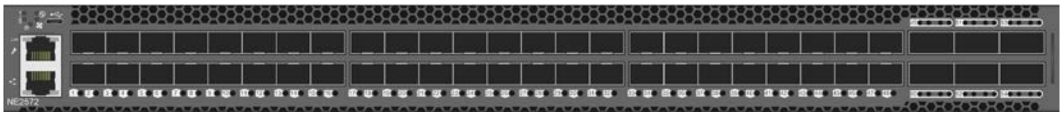


Figure 7: Lenovo ThinkSystem NE2572 RackSwitch

For more information, see this website: lenovopress.com/lp0608

2.3.2 VLANs

It is a networking best practice to use VLANs to logically separate different kinds of network traffic. The following standard VLANs are recommended:

- Management Used for all management traffic for the hypervisor
- Storage network Used for NDFS storage traffic

The following ESXi specific VLANs are recommended:

- vSphere vMotion Used to move VMs from one server to another.
- Fault Tolerance Used to support the fault tolerance (FT) feature of vSphere.

In addition, each workload application might require one or more VLANs for its logical networks. For larger networks with many workloads, it is easy to run out of unique VLANs. In this case, VXLANs could be used.

The procedure for configuring VLANs for HX Series appliances is outside of the scope of this document.

2.3.3 Redundancy

It is recommended that two top of rack (TOR) switches are used for redundancy in the data network. It is recommended to use two dual-port 10Gbps network adapters for workloads that require high throughput on the network or scale-out cluster deployments. This will effectively provide two redundant, bonded links per host for 20Gbps of bandwidth per logical link. Note that by default, the bonding configuration for ThinkAgile HX Series is active/passive, but this can be changed to active/active with the proper configuration on the hypervisor host and switch side.

In order to support the logical pairing of the network adapter ports and to provide automatic failover of the switches, the Lenovo ThinkSystem NE1032 RackSwitch and G8272 support virtual link aggregation groups (VLAGs). When VLAG is enabled over the inter-switch link (ISL) trunk, it enables logical grouping of these switches. When one of the switches is lost, or the uplink from the host to the switch is lost, the connectivity is automatically maintained over the other switch.

Figure 8 shows the two scenarios of single port and dual-port connectivity using the Lenovo RackSwitch G8272. Note the connections into the customer switch and also the extra link between the data switches and the management switch that is required for initial setup only.

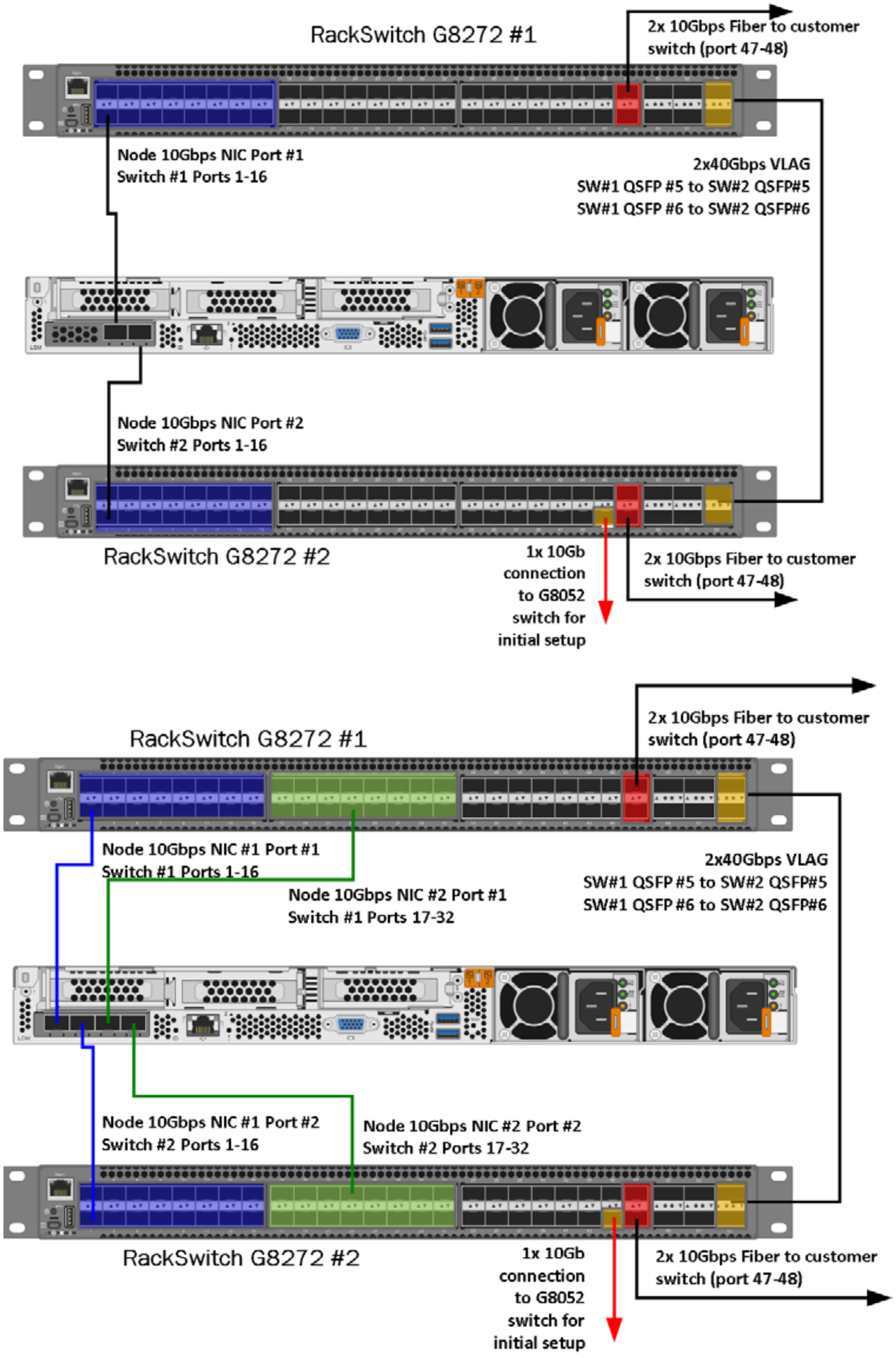

RackSwitch G8272 #1
2x 10Gbps Fiber to customer switch (port 47-48)
Node 10Gbps NIC Port #1
Switch #1 Ports 1-16
2x40Gbps VLAG
SW#1 QSFP #5 to SW#2 QSFP#5
SW#1 QSFP #6 to SW#2 QSFP#6
Node 10Gbps NIC Port #2
Switch #2 Ports 1-16
RackSwitch G8272 #2
1x 10Gb connection to G8052 switch for initial setup
2x 10Gbps Fiber to customer switch (port 47-48)
RackSwitch G8272 #1
2x 10Gbps Fiber to customer switch (port 47-48)
Node 10Gbps NIC #1 Port #1
Switch #1 Ports 1-16
Node 10Gbps NIC #2 Port #1
Switch #1 Ports 17-32
2x40Gbps VLAG
SW#1 QSFP #5 to SW#2 QSFP#5
SW#1 QSFP #6 to SW#2 QSFP#6
Node 10Gbps NIC #1 Port #2
Switch #2 Ports 1-16
Node 10Gbps NIC #2 Port #2
Switch #2 Ports 17-32
RackSwitch G8272 #2
1x 10Gb connection to G8052 switch for initial setup
2x 10Gbps Fiber to customer switch (port 47-48)

Figure 8: Data network with single and dual adapters

In addition, the Lenovo Cloud Network Operating System (CNOS) should be used on the G8272 switches. A detailed description of the CNOS operating system and its application to the HX Series deployments is provided in the following paper: Networking Guide for Lenovo ThinkAgile HX Series (CNOS Switch Firmware) lenovopress.com/lp0595-networking-guide-for-lenovo-converged-hx-series-nutanix-cnos.

2.4 Hardware management network components

The hardware management network is used for out-of-band access to ThinkAgile HX Series appliances and certified nodes via the optional Lenovo XClarity Administrator. It may also be needed to re-image an appliance. All systems management is handled in-band via Intelligent Platform Management Interface (IPMI) commands.

The dedicated Integrated Management Module (IMM) port on all of the Lenovo ThinkAgile HX series appliances and certified nodes needs to be connected to a 1GbE TOR switch as shown in Figure 9.

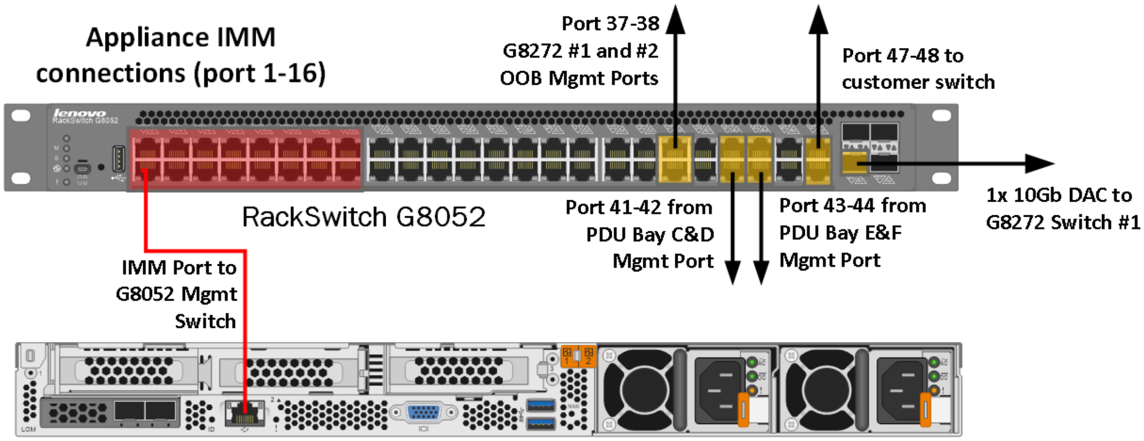

Appliance IMM connections (port 1-16)
Port 37-38 G8272 #1 and #2 OOB Mgmt Ports
Port 47-48 to customer switch
RackSwitch G8052
Port 41-42 from PDU Bay C&D Mgmt Port
Port 43-44 from PDU Bay E&F Mgmt Port
1x 10Gb DAC to G8272 Switch #1
IMM Port to G8052 Mgmt Switch

Figure 9: IMM 1GbE management network

2.4.1 Hardware management switches

The following Lenovo 1GbE TOR switches are recommended for use in a HX Series cluster:

- Lenovo RackSwitch G7028
- Lenovo RackSwitch G8052

Lenovo RackSwitch G7028

The Lenovo RackSwitch G7028 (as shown in Figure 10) is a 1 Gb top-of-rack switch that delivers line-rate Layer 2 performance at an attractive price. G7028 has 24 10/100/1000BASE-T RJ45 ports and four 10 Gb Ethernet SFP+ ports. It typically uses only 45 W of power, which helps improve energy efficiency.

Figure 10. Lenovo RackSwitch G7028

For more information, see this website: lenovopress.com/tips1268.

Lenovo RackSwitch G8052

The Lenovo System Networking RackSwitch G8052 (as shown in Figure 11) is an Ethernet switch that is designed for the data center and provides a virtualized, cooler, and simpler network solution. The Lenovo RackSwitch G8052 offers up to 48 1 GbE ports and up to four 10 GbE ports in a 1U footprint. The G8052 switch is always available for business-sensitive traffic by using redundant power supplies, fans, and numerous high-availability features.

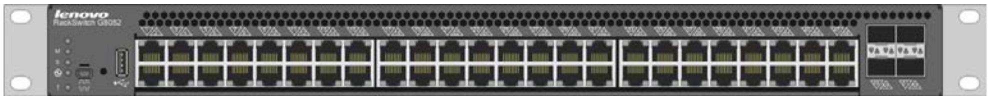


Figure 11: Lenovo RackSwitch G8052

For more information, see this website: lenovopress.com/tips0813.

2.5 Reliability and performance features

Reliability and excellent performance are important for any workload but particularly for hyper-converged infrastructures like ThinkAgile HX Series. These requirements are met through the following design features of Nutanix software combined with Lenovo Servers.

Hardware reliability

Lenovo uses the highest quality hardware components combined with firmware that is thoroughly tested. As a consequence Lenovo servers have been rated #1 in hardware reliability for the last 3 years. This is important as it lowers the frequency of a server failure which in turn lowers OPEX.

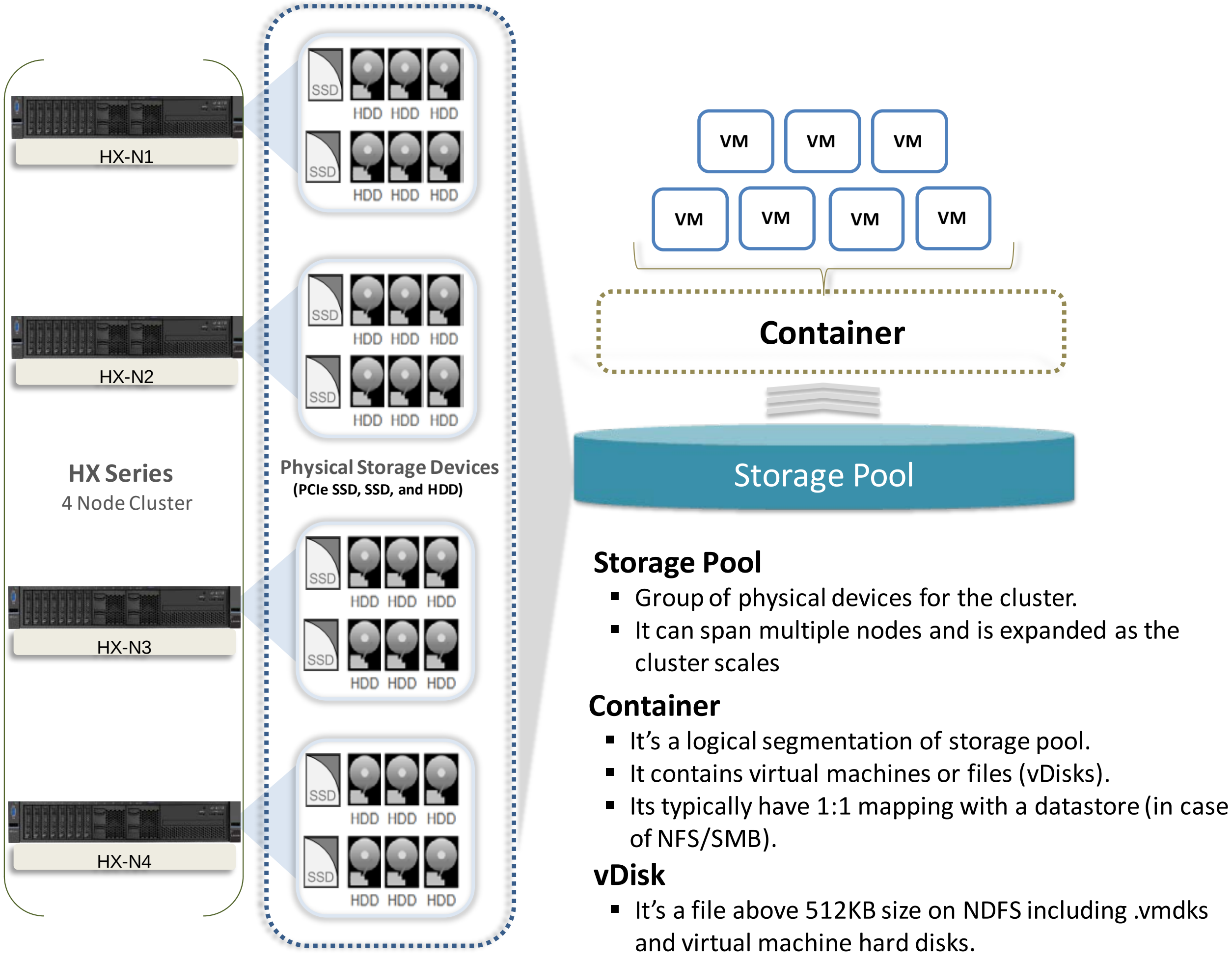
A HX appliance has redundant hardware components by including two power supplies, multiple chassis fans, two Intel CPUs, multiple memory DIMMs, multiple SSDs and HDDs, and optionally up to two dual-port network interface cards.

Hardware performance

The HX Series appliances have been carefully designed for performance. In addition to all of the usual attributes like processors and memory, the 24 drive HX7520 uses three HBA controllers instead of the one. As a consequence the latency is halved for some workloads that heavily utilize the cold tier. This allows a higher throughput and improved transaction rates.

Distributed file system

The Nutanix Distributed file system (NDFS) is an intelligent file system which virtualizes the local attached storage (SSD/HDD) on all the nodes in a cluster and presents it as single storage entity to cluster. Figure 12 shows the high level structure of NDFS:



Storage Pool

- Group of physical devices for the cluster.
- It can span multiple nodes and is expanded as the cluster scales

Container

- It's a logical segmentation of storage pool.
- It contains virtual machines or files (vDisks).
- Its typically have 1:1 mapping with a datastore (in case of NFS/SMB).

vDisk

- It's a file above 512KB size on NDFS including .vmdks and virtual machine hard disks.

Figure 12: Nutanix Distributed File System

Data protection via replication

The Nutanix platform replication factor (RF) and checksum is used to ensure data redundancy and accessibility in the event of a node or disk failure or corruption. It uses an OpLog which acts as a staging area for incoming writes on low latency SSDs which are then replicated to the OpLogs for one or two other Controller VMs before acknowledging a successful write. This approach ensures that data available in at least two to three different locations and is fault tolerant. While the data is being written a checksum is calculated and stored as part of its metadata.

In the case of a drive or node failure, that data is replicated out to more nodes to maintain the replication factor. A checksum is computed every time the data is read to ensure the data validity. If the checksum and data mismatch, then the data replica is read to replace the invalid copy.

Performance with data tiering

Nutanix uses a disk tiering concept in which disk resources (SSD and HDD) are pooled together to form a cluster wide storage tier. This tier can be accessed by any node within the cluster for data placement and can leverage the full tier capacity. The following data tiering functions are provided:

- The SSD on a local node always has the highest tier priority for write I/O.
- If the local node's SSD is full then the other SSDs in the cluster are used for I/O.
- The NDFS Information Lifecycle Management (ILM) component migrates cold data from the local SSD to HDD to free up SSD space. It also moves heavily accessed data to the local SSD to provide high performance.

Performance by data locality

Data locality is a crucial factor for cluster and VM performance. In order to minimize latency the CVM will work to ensure that all I/O happens locally. This ensures optimal performance and provides very low latencies and high data transfer speeds that cannot be achieved easily with shared storage arrays, even if all-flash.

The following occurs in case of a VM migration or high availability event that moves a VM from Node-A to Node-B:

- The VM's data is provided by the CVM running on Node-B.
- All write I/O requests occur locally i.e. to the local storage of Node-B.
- When a request comes for reading old data, the I/O request is forwarded by Node-B to Node-A. NDFS detects that the I/O request originated from different node and migrates the data locally in the background i.e. from Node-A to Node-B so that all subsequent read I/O operations are served locally. This approach (migration only on a read) helps to avoid network flooding.

Performance of snapshots and clones

NDFS provides support for offloaded snapshots and clones using a redirect-on-write algorithm. When a snapshot or clone is created, the base vDisk is marked as read only and another vDisk is created with read/write permissions as shown in Figure 13 and Figure 14 below.

At this point both vDisks have the same block map - a metadata mapping of the vDisk to its corresponding extents. This approach reduces the overhead of creating snapshots and allows snapshots to be taken very quickly with little performance impact.

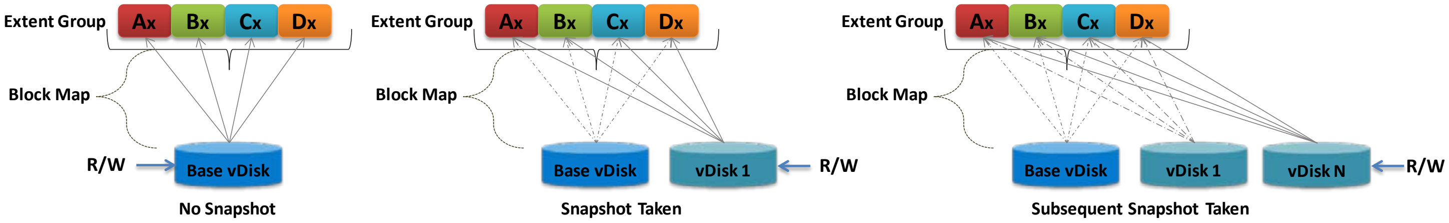


Figure 13: VM snapshots

When a VM is cloned the current block map is locked and then clones are created. These updates are metadata only so again no actual I/O takes place. The logic applies for clones of clones as well where a previously cloned VM acts as a base vDisk. All the clones inherit the prior block map and any new writes take place on the individual block maps.

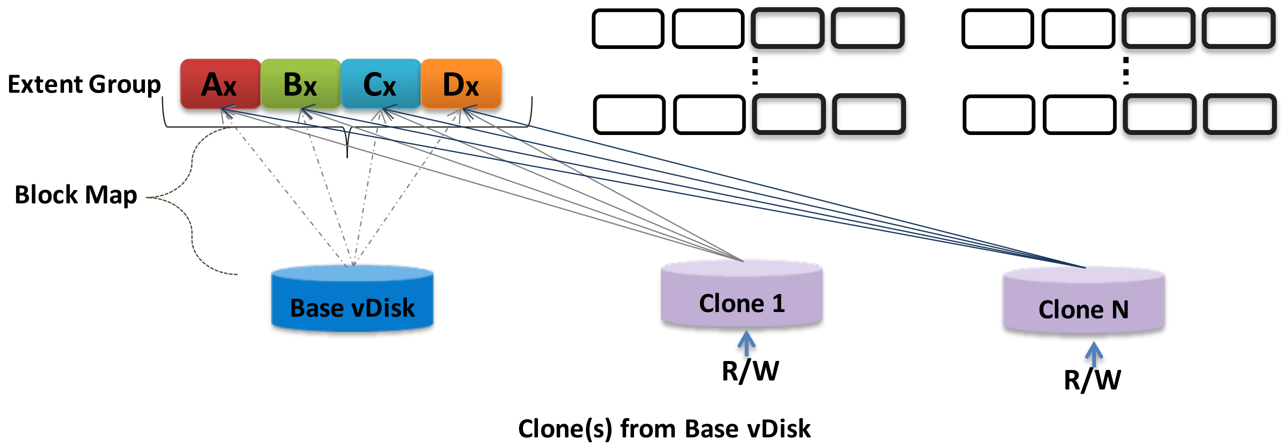


Figure 14: VM clones

Storage reduction via De-duplication and Compression

The Nutanix elastic de-duplication engine increases the effective capacity of a disk, as well as the RAM and cache of the system by removing duplicate data. It's an intelligent technology which performs following actions to increase storage efficiency:

- Sequential streams of data fingerprinted at 4K granularity
- Single instance of the shared VM data is loaded into the cache upon read
- Each node in a cluster performs its own fingerprinting and deduplication

The Nutanix capacity optimization engine is responsible for performing data transformations and compression to achieve data optimization. NDFS provides following compression methods:

- In-line compression sequential streams of data or large I/O sizes are compressed in memory before written to the disk
- Post-process compression whereby data is written in an uncompressed state and the curator framework is used to compress the data in a cluster wide manner

The Nutanix capacity optimization engine uses the Google snappy compression library to deliver good compression ratios with minimal compute overhead and very fast compression or decompression rates.

Elimination of "split-brain" errors

In a distributed system it is possible for one participant to become disconnected which will cause differences in the stored data. NDFS uses the proven "Paxos" algorithm to eliminate these "split-brain" issues by reaching a consensus (quorum) among the participants in a distributed system before the writes are made.

Drive reliability via active monitoring

The CVM actively monitors the performance of every drive in a node. The deterioration of a drive's performance may indicate that the drive is about to fail. The CVM proactively moves data off the drive before it fails and marks the drive offline and in need to replacement. The idea is to avoid the expensive data transfers to maintain data redundancy and possible loss of data.

3 Deployment models

This chapter provides recommended deployment models for different examples of using a HX Series cluster.

3.1 SMB deployment model

There are two specific models of ThinkAgile HX Series appliances targeted for the small-medium business (SMB) environment. These are the HX2320-E in a 1U form factor and HX2720-E in a 2U4N form factor.

As described in the product guide for the HX2000 Series appliances, the cluster size is limited to 4 nodes. This example deployment for a SMB customer with low performance requirements includes 4 HX2720-E nodes in a 2U chassis and a Lenovo RackSwitch G7028 1GbE switch. Figure 15 shows the front view.

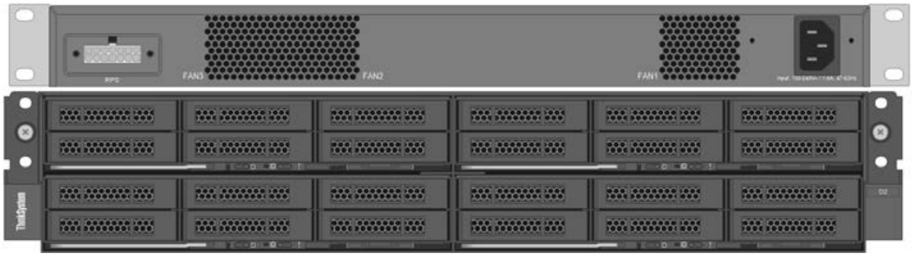


Figure 15: Front view of SMB Deployment

Figure 16 shows the rear view and the cabling for both the data (blue, red) and hardware management (green) networks into the same switch. It is also typical in these environments that the data network and management network IP addresses are in the same subnet.

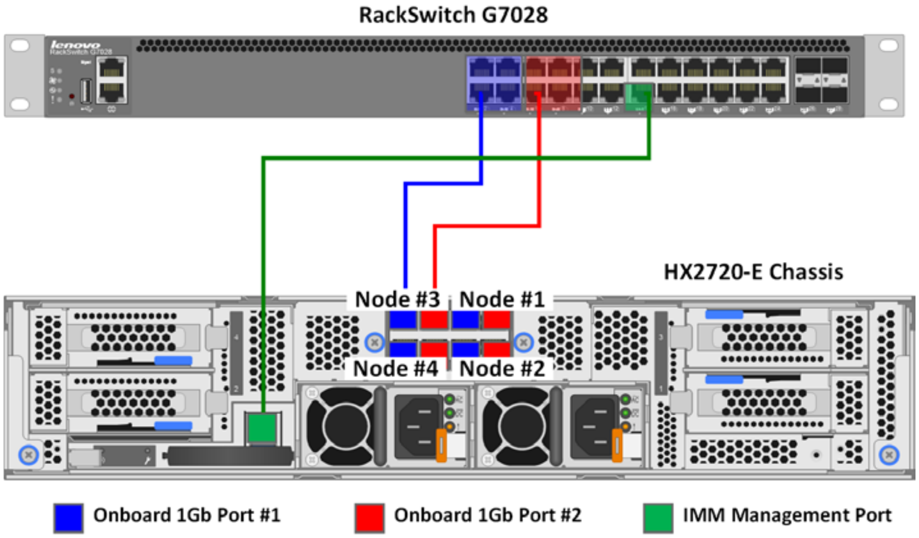


Figure 16: Rear view of SMB Deployment

3.2 Rack-scale deployment models

This section gives some example deployment models.

3.2.1 VDI deployment model

This deployment model applies equally for either Citrix XenDesktop (see chapter 4) or VMware Horizon (see chapter 0). In this example the requirement is to support 5000 stateless virtual desktops. Each Windows 10 desktop virtual machine (VM) requires 3GB of RAM and 1 vCPU (similar to Office worker profile).

In order to promote reusability, the management nodes are configured the same as the compute nodes running the virtual desktops. Because of the low requirement for storage, the HX3720 appliance is used and configured as follows:

- 2x Intel Xeon Scalable 6130 processors with 768 GB of system memory
- 2x 1.92TB SSD for cache
- 4x 2TB SATA HDD

Using the sizing tables presented in chapter 4 and chapter 0, it is recommended to have on average 180 virtual desktops per node. This translates to twenty-eight HX3720 appliances. In this configuration up to 5 nodes could be out of service (which is very unlikely) and that would leave twenty-three nodes to service the 5000 virtual desktops. This gives a 1:6 ratio for redundancy of the nodes.

In addition three nodes are needed for the VDI management VMs and to provide adequate failover characteristics. A fourth node could be used as a quick deploy spare.

Figure 17 shows the deployment for 5000 VDI users with thirty-two HX3720 appliances (8 chassis) and the TOR switches. Each of the compute nodes is numbered C1 to C28 and each of the management nodes is numbered M1 to M3.

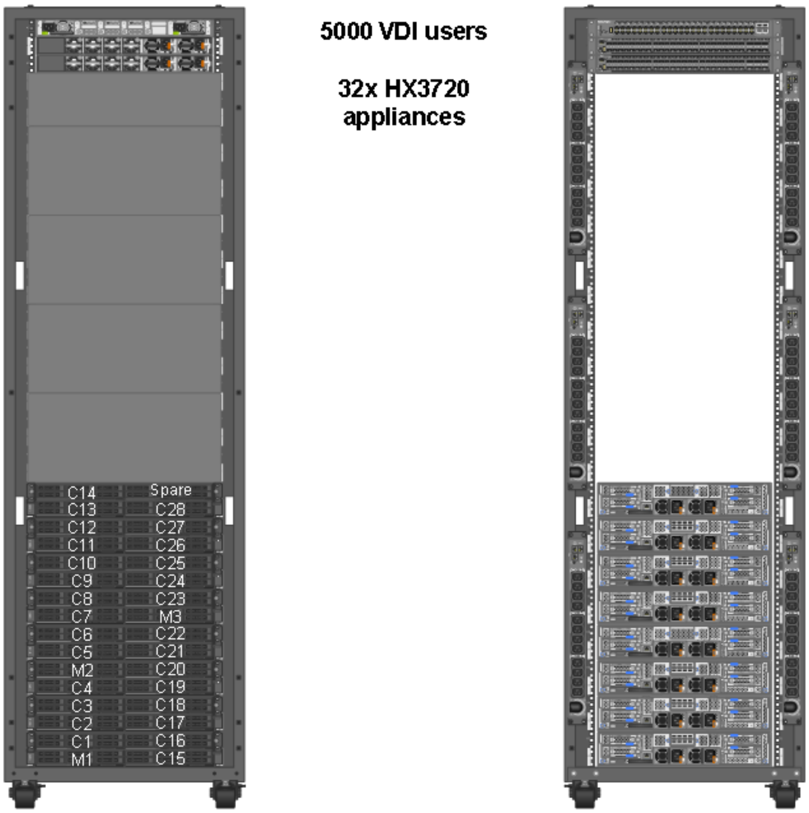


Figure 17: Example deployment model for 5000 VDI users

3.2.2 Microsoft Exchange and SQL Server deployment model

In this example deployment model, the requirement is to support 70,000 mailboxes of 1GB each and 8 applications that use Microsoft SQL Server for the database.

The 70,000 mailboxes are supported using 8 HX7520 appliances. An additional 4 HX7520 appliances are used for sixteen Microsoft SQL Server VMs which use Microsoft AlwaysOn availability groups (AAGs) for database redundancy. The VMs for the web front-ends are redundantly deployed on 4 HX3320 appliances.

The HX7520 appliance is configured as follows:

- 2x Intel Xeon Scalable 8170 processors with 768 GB of system memory
- 4x 1.92TB SSD for cache
- 20x 2TB SATA HDD

The HX3320 appliance is configured as follows:

- 2x Intel Xeon Scalable 6130 processors with 768 GB of system memory
- 2x 1.92TB SSD for cache
- 8x 2TB SATA HDD

Figure 18 shows the deployment of the 12 HX7520 appliances, 4 HX3320 appliances, and the TOR switches.

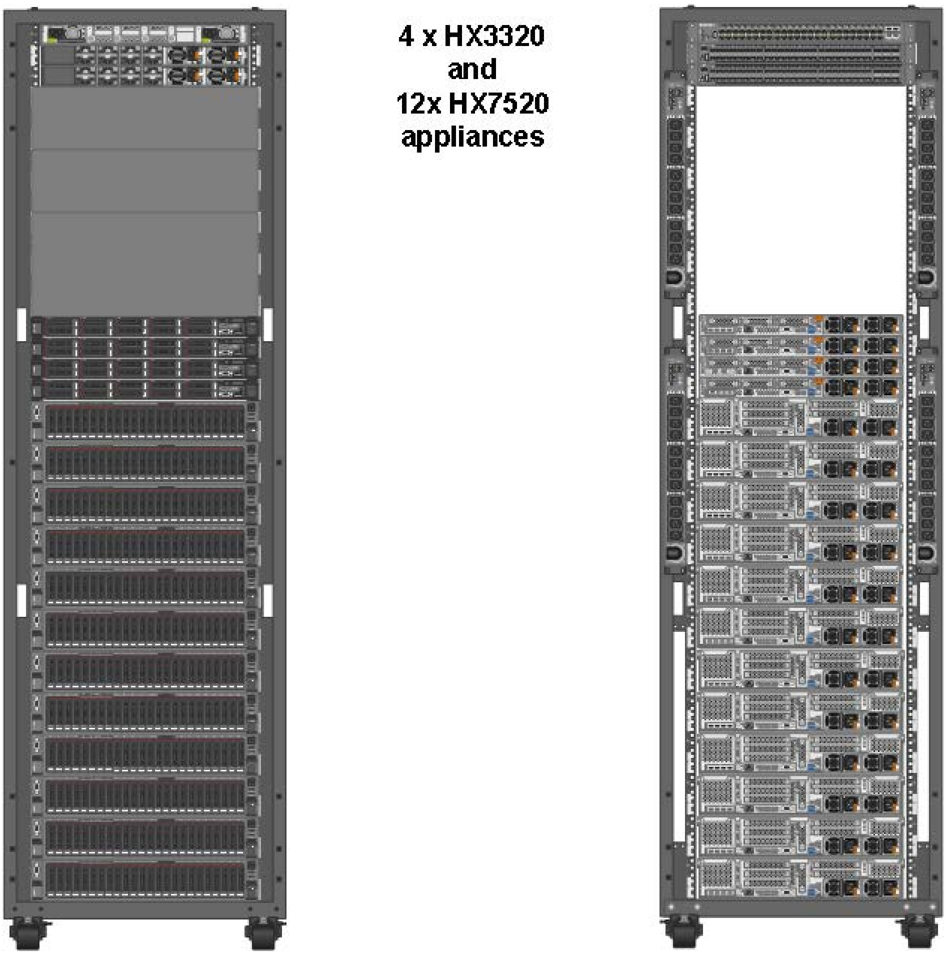


Figure 18: Example deployment model for Microsoft Exchange and SQL Server

3.2.3 VMware vCloud Suite deployment model

In this example deployment model, the requirement is to support VDI users with graphics acceleration using a VMware vCloud Suite. For more details see section 10.7 on page 115.

The vCloud edge-compute cluster uses 12 HX3520-G appliances configured as follows:

- 2x Intel Xeon Scalable 6130 processors with 384 GB of system memory
- 2x 1.92TB SSD for cache
- 6x 2TB SATA HDD
- 2x M60 GPU adapter

The vCloud management cluster uses four HX3320 appliances configured as follows:

- 2x Intel Xeon Scalable 6130 processors with 384 GB of system memory
- 2x 1.92TB SSD for cache
- 8x 2TB SATA HDD

Figure 19 shows the deployment of the 12 HX3520-G appliances, 4 HX3320 appliances, and the TOR switches.

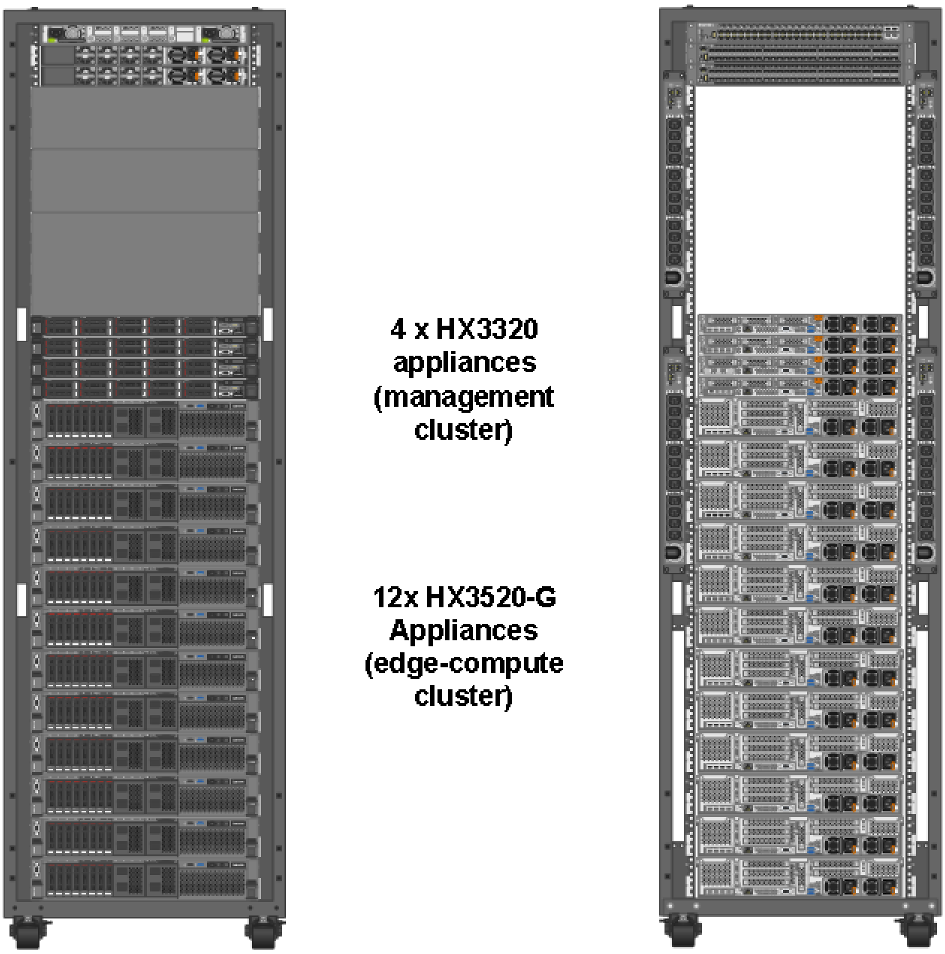

4 x HX3320 appliances (management cluster)
12x HX3520-G Appliances (edge-compute cluster)

Figure 19: Example deployment model for VMware vCloud

4 Citrix XenDesktop

Citrix XenDesktop is a suite of virtualization software which delivers Windows virtual desktops as well as virtual applications to meet the demands of any use case. It is based on the unified FlexCast Management Architecture (FMA) platform. See this website for more details: citrix.com/products/xendesktop.

4.1 Solution overview

Figure 20 provides an architecture overview of Lenovo hyper-converged Nutanix solution's reference architecture with Citrix XenDesktop 7.15 on VMware ESXi 6.7 U1 hypervisor. This chapter does not address the general issues of multi-site deployment and network management and limits the description to the components that are inside the customer's intranet.

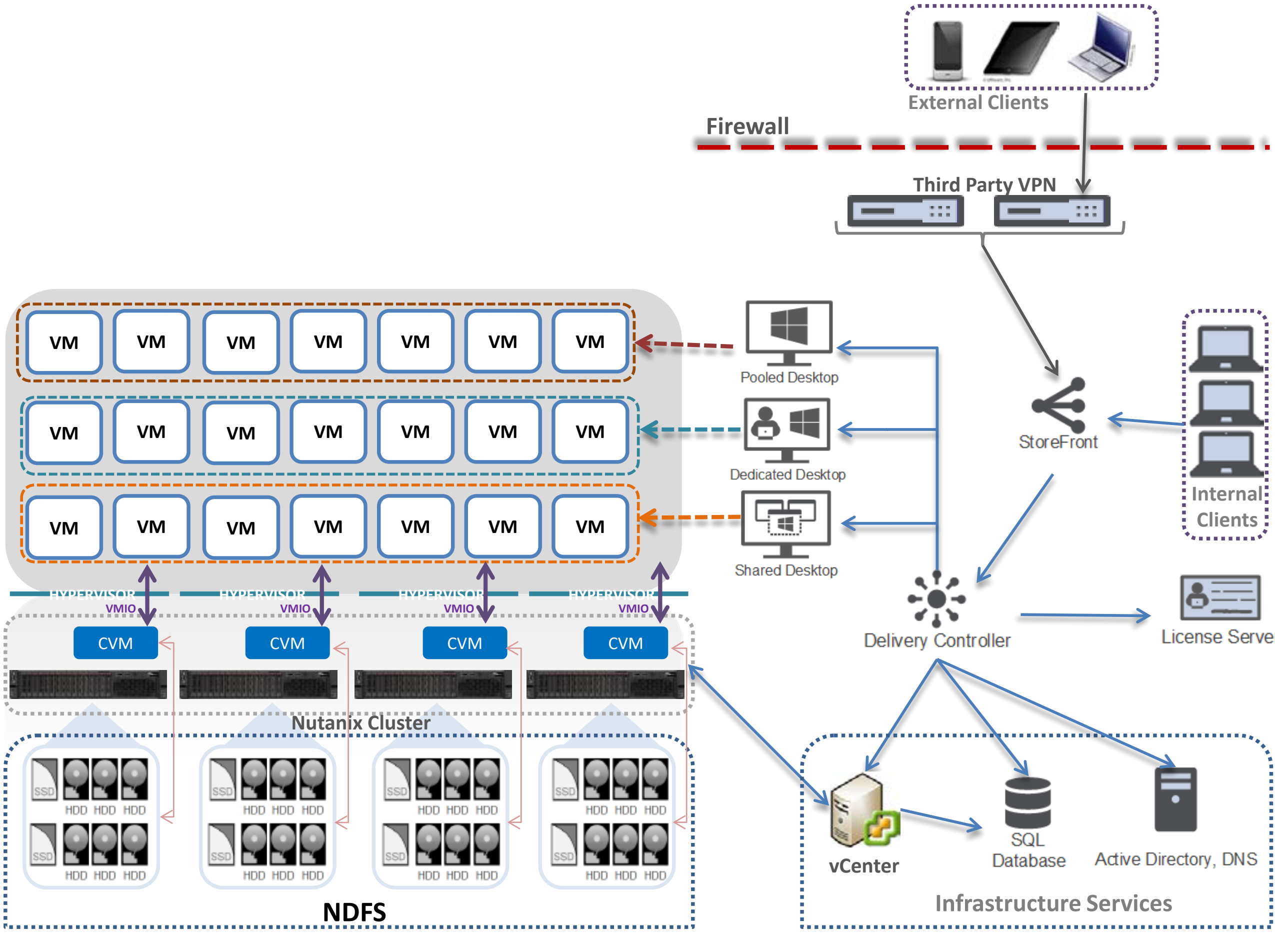


Figure 20: Lenovo ThinkAgile HX Series solution with Citrix XenDesktop

Pooled desktops are stateless (non-persistent) virtual desktops and dedicated desktops are persistent. Shared desktops are used for hosted shared desktops or hosted shared applications.

4.2 Component model

Figure 21 is a layered component view for the Citrix XenDesktop virtualization infrastructure.

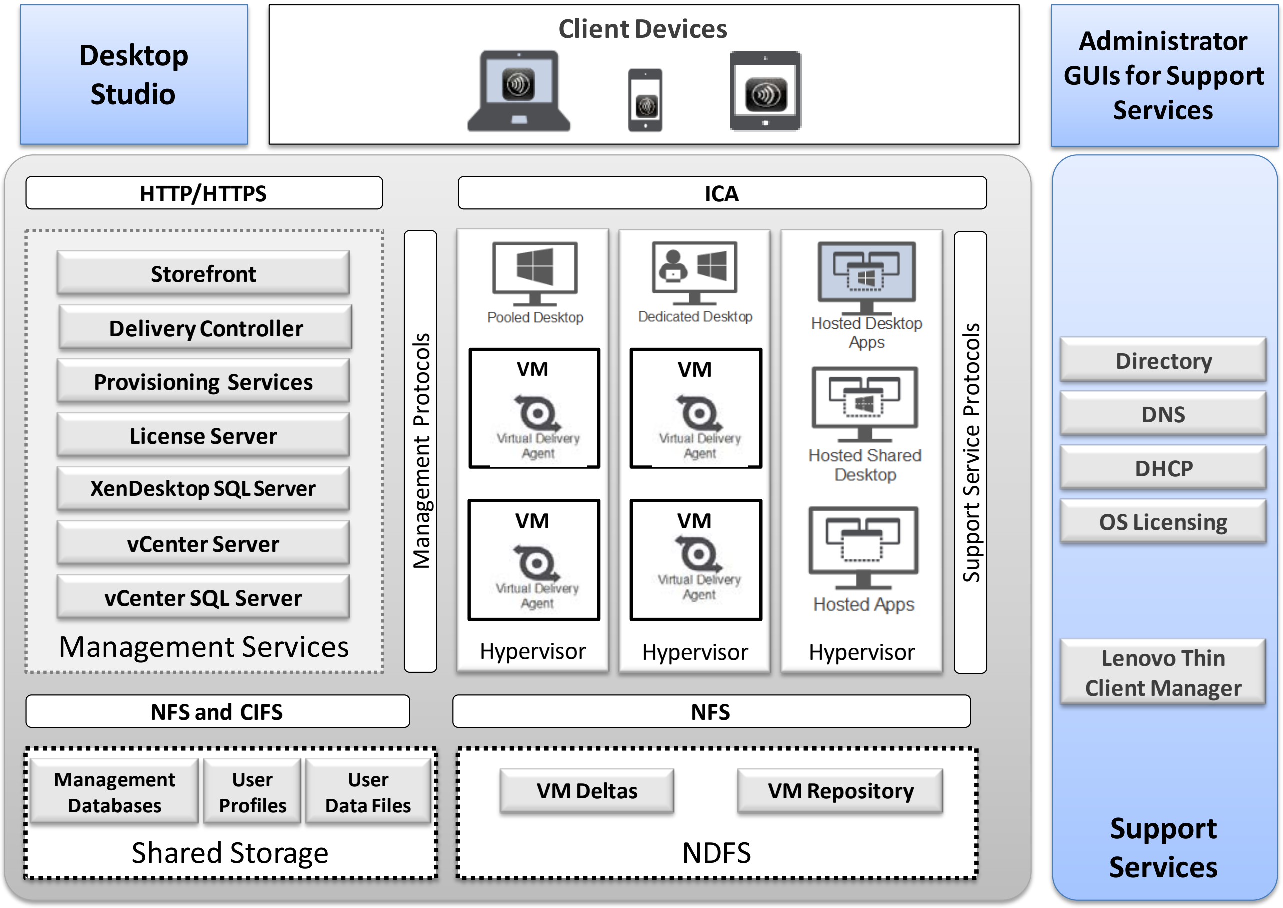


Figure 21: Component model with Citrix XenDesktop

Citrix XenDesktop features the following main components:

Desktop Studio Desktop Studio is the main administrator GUI for Citrix XenDesktop. It is used to configure and manage all of the main entities, including servers, desktop pools and provisioning, policy, and licensing.

Storefront Storefront provides the user interface to the XenDesktop environment. The Web Interface brokers user authentication, enumerates the available desktops and, upon start, delivers a .ica file to the Citrix Receiver on the user‘s local device to start a connection. The Independent Computing Architecture (ICA) file contains configuration information for the Citrix receiver to communicate with the virtual desktop. Because the Web Interface is a critical component, redundant servers must be available to provide fault tolerance.

Delivery controller

The Delivery controller is responsible for maintaining the proper level of idle desktops to allow for instantaneous connections, monitoring the state of online and connected desktops, and shutting down desktops as needed.

A XenDesktop farm is a larger grouping of virtual machine servers. Each delivery controller in the XenDesktop acts as an XML server that is responsible for brokering user authentication, resource enumeration, and desktop starting. Because a failure in the XML service results in users being unable to start their desktops, it is recommended that you configure multiple controllers per farm.

PVS and MCS

Provisioning Services (PVS) is used to provision stateless desktops at a large scale. Machine Creation Services (MCS) is used to provision dedicated or stateless desktops in a quick and integrated manner. For more information, see “Citrix XenDesktop provisioning” section on page 27.

License Server

The Citrix License Server is responsible for managing the licenses for all XenDesktop components. XenDesktop has a 30-day grace period that allows the system to function normally for 30 days if the license server becomes unavailable. This grace period offsets the complexity of otherwise building redundancy into the license server.

XenDesktop SQL Server

Each Citrix XenDesktop site requires an SQL Server database that is called the data store, which used to centralize farm configuration information and transaction logs. The data store maintains all static and dynamic information about the XenDesktop environment. Because the XenDesktop SQL server is a critical component, redundant servers must be available to provide fault tolerance.

vCenter Server

By using a single console, vCenter Server provides centralized management of the virtual machines (VMs) for the VMware ESXi hypervisor. VMware vCenter can be used to perform live migration (called VMware vMotion), which allows a running VM to be moved from one physical server to another without downtime.

Redundancy for vCenter Server is achieved through VMware high availability (HA). The vCenter Server also contains a licensing server for VMware ESXi.

vCenter SQL Server

vCenter Server for VMware ESXi hypervisor requires an SQL database. The vCenter SQL server might be Microsoft® Data Engine (MSDE), Oracle, or SQL Server. Because the vCenter SQL server is a critical component, redundant servers must be available to provide fault tolerance. Customer SQL databases (including respective redundancy) can be used.

Client devices Citrix XenDesktop supports a broad set of devices and all major device operating platforms, including Apple iOS, Google Android, and Google ChromeOS. XenDesktop enables a rich, native experience on each device, including support for gestures and multi-touch features, which customizes the experience based on the type of device. Each client device has a Citrix Receiver, which acts as the agent to communicate with the virtual desktop by using the ICA/HDX protocol.

Thin-client Manager The Lenovo Thin-client Manager (LTM) is used to manage and support Lenovo thin-client devices individually or in groups.

VDA Each VM needs a Citrix Virtual Desktop Agent (VDA) to capture desktop data and send it to the Citrix Receiver in the client device. The VDA also emulates keyboard and gestures sent from the receiver. ICA is the Citrix remote display protocol for VDI.

Citrix Receiver Citrix Receiver is the client software that provides access to applications, desktops and data easily and securely from any device, including smartphones, tablets, PCs and Macs

For more information, see the Lenovo Client Virtualization base reference architecture document that is available at this website: lenovopress.com/lp0756.

4.3 Citrix XenDesktop provisioning

Citrix XenDesktop features the following primary provisioning components for desktops and applications:

- Provisioning Services (PVS)
- Machine Creation Services (MCS)

4.3.1 Provisioning services

Hosted VDI desktops can be deployed with or without Citrix PVS. The advantage of PVS is that you can stream a single desktop image to create multiple virtual desktops on one or more servers in a data center. Figure 22 shows the sequence of operations that are executed by XenDesktop to deliver a hosted VDI virtual desktop.

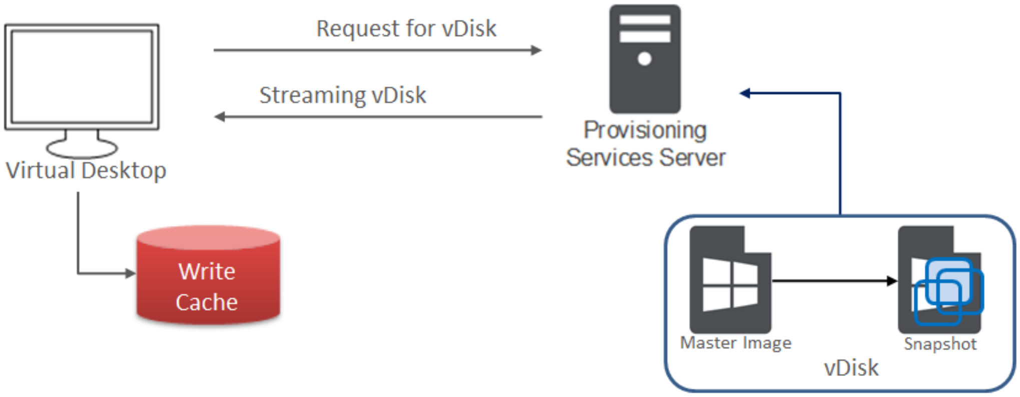


Figure 22: Using PVS for a stateless model

When the virtual disk (vDisk) master image is available from the network, the VM on a target device no longer needs its local hard disk drive (HDD) to operate; it boots directly from the network and behaves as if it were running from a local drive on the target device, which is why PVS is recommended for stateless virtual desktops. PVS often is not used for dedicated virtual desktops because the write cache is not stored on shared storage.

PVS is also used with Microsoft Roaming Profiles (MSRPs) so that the user's profile information can be separated out and reused. Profile data is available from CIFS based shared storage.

It is a best practice to use snapshots for changes to the master VM images and also keep copies as a backup.

4.3.2 Machine creation services

Unlike PVS, MCS does not require more servers. Instead, it uses integrated functionality that is built into the hypervisor and communicates through the APIs. Each desktop has one difference disk and one identity disk (as shown in Figure 23). The difference disk is used to capture any changes that are made to the master image. The identity disk is used to store information, such as device name and password.

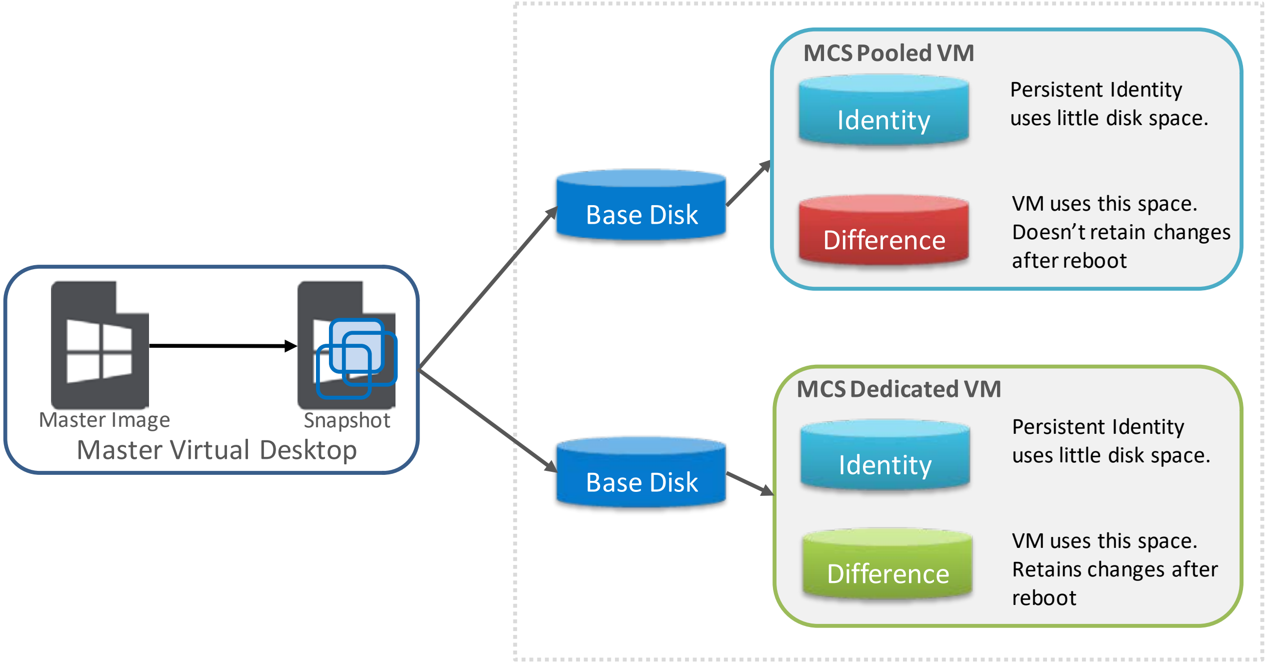

MCS Pooled VM
Identity
Persistent Identity uses little disk space.
Difference
VM uses this space. Doesn't retain changes after reboot
Base Disk
Master Image
Snapshot
Master Virtual Desktop
Base Disk
MCS Dedicated VM
Identity
Persistent Identity uses little disk space.
Difference
VM uses this space. Retains changes after reboot

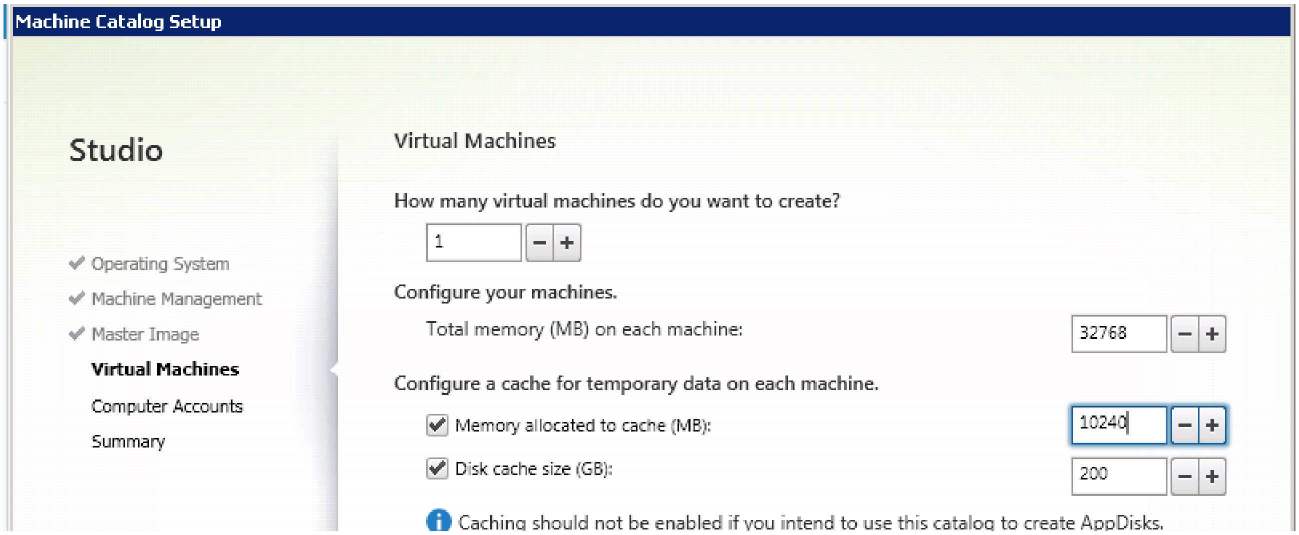
Figure 23: MCS image and difference/identity disk storage model

The following types of image assignment models for MCS are available:

- Pooled-random: Desktops are assigned randomly. When they log off, the desktop is free for another user. When rebooted, any changes that were made are destroyed.
- Pooled-static: Desktops are permanently assigned to a single user. When a user logs off, only that user can use the desktop, regardless if the desktop is rebooted. During reboots, any changes that are made are destroyed.
- Dedicated: Desktops are permanently assigned to a single user. When a user logs off, only that user can use the desktop, regardless if the desktop is rebooted. During reboots, any changes that are made persist across subsequent restarts.

MCS thin provisions each desktop from a master image by using built-in technology to provide each desktop with a unique identity. Only changes that are made to the desktop use more disk space.

There is a new caching option in Citrix XenDesktop 7.9 for Pooled and Hosted Shared desktops. Figure 24 shows a screenshot of the option.



Machine Catalog Setup

Studio

- ✓ Operating System
- ✓ Machine Management
- ✓ Master Image
- **Virtual Machines**
- Computer Accounts
- Summary

Virtual Machines

How many virtual machines do you want to create? 1

Configure your machines.

Total memory (MB) on each machine: 32768

Configure a cache for temporary data on each machine.

☑ Memory allocated to cache (MB): 10240

☑ Disk cache size (GB): 200

Caching should not be enabled if you intend to use this catalog to create AppDisks.

Figure 24: Caching option for Pooled and Hosted Shared desktops

4.4 Management VMs

A key part of the Citrix XenDesktop environment is the various management VMs used to manage the VDI infrastructure and user VMs. Table 3 lists the VM requirements and performance characteristics of each management service for Citrix XenDesktop.

Table 3: Characteristics of XenDesktop and ESXi management services

Management service VM	Virtual processors	System memory	Storage	Windows OS	HA needed	Performance characteristic
Delivery controller	4	8 GB	60 GB	2012 R2	Yes	5000 user connections
Storefront	4	4 GB	60 GB	2012 R2	Yes	30,000 connections per hour
Citrix licensing server	2	4 GB	60 GB	2012 R2	No	170 licenses per second
XenDesktop SQL server	2	8 GB	60 GB	2012 R2	Yes	5000 users
PVS servers	4	32 GB	60 GB (depends on number of images)	2012 R2	Yes	Up to 1000 desktops, memory should be a minimum of 2 GB plus 1.5 GB per image served

Management service VM	Virtual processors	System memory	Storage	Windows OS	HA needed	Performance characteristic
vCenter server	8	16 GB	60 GB	2012 R2	No	Up to 2000 desktops
vCenter SQL server	4	8 GB	200 GB	2012 R2	Yes	Double the virtual processors and memory for more than 2500 users

Table 4 lists the number of management VMs for each size of users following the recommendations for high availability and performance. The number of vCenter servers is half of the number of vCenter clusters because each vCenter server can handle two clusters of up to 1000 desktops.

Table 4: Management VMs needed

XenDesktop management service VM	300 users	600 users	1200 users	3000 users
Delivery Controllers Includes Citrix Licensing server Includes Web server	2 (1+1) Y Y	2 (1+1) Y Y	2 (1+1) N N	2 (1+1) N N
Web Interface	N/A	N/A	2 (1+1)	2 (1+1)
Citrix licensing servers	N/A	N/A	1	1
XenDesktop SQL servers	2 (1+1)	2 (1+1)	2 (1+1)	2 (1+1)
PVS servers for stateless case only	2 (1+1)	2 (1+1)	2 (1+1)	2 (1+1)
ESXi management service VM	**300 users**	**600 users**	**1200 users**	**3000 users**
vCenter servers	1	1	2	2
vCenter SQL servers	2 (1+1)	2 (1+1)	2 (1+1)	2 (1+1)

It is assumed that common services, such as Microsoft Active Directory, Dynamic Host Configuration Protocol (DHCP), domain name server (DNS), and Microsoft licensing servers exist in the customer environment.

There are 3 ways to deploy the management VMs for Citrix XenDesktop:

- Separate servers outside HX cluster
- Servers within the HX cluster
- In the Citrix Cloud using the Citrix Workspace Appliance (see section 4.4.1)

These management VMs can be run on separate servers from the HX series cluster or within the cluster itself. Separating out the VMs means that the management VMs can be separately managed and sized to the requirements and dedicated servers used for the user VMs. Putting all of the VMs together in one cluster means that the compute servers will execute less user VMs and additional resources are needed for the much larger and more granular management VMs. Lenovo recommends that the management and user VMs are separated for all but the smallest deployments (i.e. less than 600 users).

4.4.1 Citrix Workspace Appliance

Citrix Workspace Appliance (CWA) hosts the management VMs for XenDesktop in the Citrix Cloud; for a service fee. This significantly reduces the skills, effort, and hardware needed to manage a XenDesktop environment. The customer only needs to configure a cluster of servers to host user VMs for dedicated or stateless desktops.

Citrix installs and manages the following components in Citrix Cloud:

Delivery Controller

The Delivery Controller is the central control layer component in a deployment. The Controller’s services communicate through the Cloud Connectors in each resource location to:

- Distribute applications and desktops.
- Authenticate and manage user access.
- Broker connections between users and their virtual desktops and applications.
- Optimize use connections, and load-balance these connections.
- Track which users are logged on and where, which session resources the users have, and if users need to reconnect to existing applications. This includes managing the state of desktops, starting and stopping them based on demand and administrative configuration.

Data from the Controller services is stored in a Microsoft SQL Server site database. A deployment also uses a Configuration Logging database, plus a monitoring database used by Director.

When the Citrix Cloud is hosted on Microsoft Azure, SQL Server on Azure VMs (IaaS) is supported. Azure SQL (PaaS) Database is not supported.

Citrix license management

License management functionality communicates with the Controller to manage licensing for each user’s session and allocate license files. The customer administrator does not need to configure or manage anything with licensing. All that work is done automatically in Citrix Cloud.

Citrix Studio

Studio is the management console you use to configure and manage connections, machine catalogs, and Delivery Groups. Studio launches when you select **Manage** in the Citrix Cloud console.

Citrix Director

Director enables IT support and help desk teams to monitor an environment, troubleshoot issues before they become system-critical, and perform support tasks for end users. Displays include:

- Real-time session data from the Broker Service in the Controller, which includes data from the broker agent in the VDA.
- Historical data from the Monitor Service in the Controller.
- Data about HDX traffic (also known as ICA traffic).

Director provides the data when you select **Monitor** in the Citrix Cloud console.

Workspace configuration

From Workspace configuration in Citrix Cloud, you can:

- Specify which services are integrated with Citrix Workspace.
- Customize the URL that your subscribers use to access their workspace.
- Customize the appearance of subscribers’ workspaces, such as logos, color, and preferences.
- Specify how subscribers authenticate to their workspace, such as using Active Directory or Azure Active Directory.
- Specify external connectivity for resource locations used by your subscribers.

Citrix StoreFront

Citrix StoreFront is the predecessor to Citrix Workspace, and is used in on-premises deployments. It offers several similar features to Citrix Workspace, including subscriber access to apps and desktops, authentication, and other subscriber data that ensures consistent experience across devices.

In a service deployment, you can optionally install StoreFront servers in resource locations. Having local stores can help deliver apps and desktops during network outages.

Citrix Gateway

When users connect from outside the corporate firewall, Citrix Virtual Apps and Desktops can use Citrix Gateway technology to secure these connections with TLS. The Citrix Gateway or VPX virtual appliance is an SSL VPN appliance deployed in the DMZ. It provides a single secure point of access through the corporate firewall.

Although Citrix installs and manages Citrix Gateway in Citrix Cloud, you can also optionally install Citrix Gateway in resource locations.

Components in resource locations (on-premises):

A resource location contains resources required to deliver services to your subscribers (users). You manage these resources from Citrix Cloud. Resource locations contain different resources depending on which Citrix Cloud services you’re using and the services you want to provide to your users.

To interact with Citrix Cloud, each resource location needs Cloud Connectors and access to a Microsoft Active Directory domain.

In a Citrix Virtual Apps and Desktops service deployment, a resource location contains items from the access layer and resource layer:

- Cloud Connectors
- Active Directory domain controller
- Virtual Delivery Agents (VDAs)
- Hypervisors that provision VDAs and store their data, if used
- Citrix Gateway (optional)

- StoreFront servers (optional)

Cloud Connector

Every resource location contains at least one Cloud Connector (two or more are recommended for redundancy). A Cloud Connector is the communications channel between the components in the Citrix Cloud and components in the resource location. In the resource location, the Cloud Connector acts as a proxy for the Delivery Controller in Citrix Cloud.

You install Cloud Connectors from the Citrix Cloud console. Citrix then manages and updates the Cloud Connectors automatically.

Virtual Delivery Agents (VDAs)

Each physical or virtual machine that delivers applications and desktops must have a VDA. The VDA registers with a Cloud Connector. After registration, connections are brokered from those resources to users. VDAs establish and manage the connection between the machine and the user device, and apply policies that are configured for the session.

The VDA communicates session information to the Cloud Connector through a broker agent in the VDA. The broker agent hosts multiple plugins and collects real-time data.

VDAs are available for Windows server and desktop operating systems. VDAs for Windows server operating systems allow multiple users to connect to the server at one time. VDAs for Windows desktop operating systems allow only one user to connect to the desktop at a time. Linux VDAs are also available.

Throughout this documentation, “VDA” refers to the agent and the machine on which it is installed.

Hypervisors and cloud services

A hypervisor or cloud service contains the VDAs that host applications and desktops.

To provision virtual machines that deliver applications and desktops, you can use:

- Machine Creation Services: The MCS technology is built into Studio and is accessed automatically through the Citrix Cloud Console. MCS creates copies of a master image to create and provision VMs.
- Citrix Provisioning (formerly Provisioning Services): The Citrix Provisioning technology streams a master image to user devices. Citrix Provisioning doesn’t require a hypervisor, so you can provision physical machines.
- Another provisioning tool of your choice.

Although many deployments use hypervisors, you don’t need one if:

- Your applications and desktops are hosted on physical machines.
- You use Citrix Provisioning to provision VMs.
- You want to deploy Remote PC Access, which enables employees to remotely access their physical PCs.

Active Directory

Although not a Citrix component, Microsoft Active Directory is required for authentication and authorization in any deployment. The Kerberos infrastructure in Active Directory is used to guarantee the authenticity and confidentiality of communications with Citrix Cloud.

Items that help deliver desktops and applications

As part of delivering applications and desktops to users in a production environment, you configure the following items.

Host connection

A host connection enables communication between components in the control plane (Citrix Cloud) and VDAs in a hypervisor or cloud service. Connection specifications include:

- The address and credentials to access the host
- Which tool you use to create VMs
- The storage method to use, and the machines to use for storage
- Which network the VMs will use

Machine catalog

A machine catalog is a collection of virtual or physical machines that have the same operating system type: server or desktop.

If you use VMs, you can create a master image (also known as template) on the hypervisor or cloud service, and install a VDA on the master image. You can also install applications on the master image, if you want them to appear on all machines created from that image and don’t want to virtualize them. Then, you create a catalog using a Citrix tool (MCS or Citrix Provisioning) or your own tools. With Citrix tools, the catalog creation process provisions identical VMs from that image.

If you use your own tools to provision VMs, or if you use physical machines, the catalog creation process adds those machines to the catalog.

For technical details about the Citrix provisioning tools, see Citrix Virtual Apps and Desktops Image Management.

Delivery Group

A Delivery Group specifies:

- One or more machines from a machine catalog.
- Optionally, users who are allowed to access those machines. Alternatively, you can specify users through the Citrix Cloud console.
- Optionally, which applications and desktops users can access. Alternatively, you can specify applications through the Citrix Cloud console.

Delivering applications and desktops

Delivery methods describes the choices available to deliver applications and desktops to users.

Service Level Agreement

The Citrix Virtual Apps and Desktops service (the Service) is designed using industry best practices to achieve cloud scale and a high degree of service availability.

For complete details about Citrix's commitment for availability of Citrix Cloud services, see the Service Level Agreement.

Performance against this goal can be monitored on an ongoing basis at https://status.cloud.com.

For more information on CWA and the Citrix Cloud, see docs.citrix.com/en-us/citrix-cloud.html.

4.5 Graphics acceleration

This section is specific to the Lenovo ThinkAgile HX3520-G and ThinkAgile HX3521-G that support GPU acceleration. The VMware ESXi hypervisor supports the following options for graphics acceleration:

- Dedicated GPU with one GPU per user, which is called virtual dedicated graphics acceleration (vDGA) mode.
- GPU hardware virtualization (vGPU) that partitions each GPU.
- Shared GPU with users sharing a GPU, which is called virtual shared graphics acceleration (vSGA) mode and is not recommended because of user contention for shared use of the GPU.

The vDGA (or pass-through) option has a low user density as it restricts a single user to access each very powerful GPU. This option is not flexible and is no longer cost effective even for high-end power users. Therefore vDGA is no longer recommended especially given that the performance of the equivalent vGPU mode is similar.

When using the vGPU option with ESXi 6.5 and the latest drivers from NVidia, it is necessary to change the default GPU mode from "Shared" (vSGA) to "Shared Direct" (vGPU) for each GPU using VMware vCenter. This enables the correct GPU support for the VMs which would otherwise result in the VM not powering on correctly and the standard "graphics resources not available" error message. The host needs to be rebooted for the changes to take effect.

The performance of graphics acceleration was tested using the Lenovo ThinkSystem SR650 servers. Each server supports up to two GPU adapters. The Heaven benchmark is used to measure the per user frame rate for different GPUs, resolutions, and image quality. This benchmark is graphics-heavy and is fairly realistic for designers and engineers. Power users or knowledge workers usually have less intense graphics workloads and can achieve higher frame rates. Table 5 lists the results of the Heaven benchmark as FPS that are available to each user with the GRID 2,0 M60 adapter by using vGPU mode with DirectX 11.

Table 5: Performance of GRID 2.0 M60 vGPU modes with DirectX 11

Quality	Tessellation	Anti-Aliasing	Resolution	M60-8Q	M60-4Q	M60-2Q	M60-4A	M60-2A
High	Normal	0	1280x1024	Untested	Untested	32.07	59.94	33.12
High	Normal	0	1680x1050	Untested	50.43	25.68	N/A	N/A
High	Normal	0	1920x1200	Untested	41.82	21.72	N/A	N/A
Ultra	Extreme	8	1280x1024	Untested	Untested	18.24	36.91	18.76
Ultra	Extreme	8	1680x1050	57.61	29.79	14.40	N/A	N/A
Ultra	Extreme	8	1920x1080	50.50	26.18	12.63	N/A	N/A
Ultra	Extreme	8	1920x1200	46.17	23.01	Untested	N/A	N/A
Ultra	Extreme	8hai	2560x1600	27.46	14.19	Untested	N/A	N/A

Lenovo recommends that a medium to high powered CPU, such as the Xeon Scalable 6130, is used for accelerated graphics applications tend to also require extra load on the processor. For vGPU mode, Lenovo recommends at least 384GB of server memory. Because there are many variables when graphics acceleration is used, Lenovo recommends that testing is done in the customer environment to verify the performance for the required user workloads.

4.6 Performance testing

This section describes the performance benchmarking tool and the results obtained for different configurations of a cluster of 4 Lenovo ThinkAgile HX3320 appliances. This chapter applies equally to the ThinkAgile HX3321 certified nodes.

4.6.1 Login VSI benchmarking tool

Login VSI is a vendor-independent benchmarking tool that is used to objectively test and measure the performance and scalability of server-based Windows desktop environments. Leading IT analysts recognize and recommend Login VSI as an industry-standard benchmarking tool for client virtualization and can be used by user organizations, system integrators, hosting providers, and testing companies.

Login VSI provides multiple workloads to simulate real user work and suitable in performing load test, benchmarking and capacity planning for VDI environments. Table 6 lists the characteristics of the Login VSI 4.1 workloads that are used in the Lenovo testing.

Table 6. Login VSI Workload Comparison

Workload Name	Login VSI Version	Apps Open	CPU Usage	Disk Reads	Disk Writes	IOPS	Memory	vCPU
Office worker	4.1	5-8	82%	90%	101%	8.1	2GB	1vCPU
Knowledge worker	4.1	5-9	100%	100%	100%	8.5	2GB	2vCPU
Power worker	4.1	8-12	119%	133%	123%	10.8	3GB	3vCPU

The VSImax score parameter (the number indicates user density) is used to determine the performance of a particular system configuration. The following parameters and rules are used for Login VSI tests:

- User login interval: 30 seconds per node
- Workload: Office Worker, Knowledge Worker, or Power User
- All virtual desktops were pre-booted before the tests
- The number of powered-on VMs was adjusted to stay within a 10% margin of VSImax to avoid unreasonable overhead by “idling” virtual machines
- VSImax score is derived using the “classic model” calculation

4.6.2 Performance results for virtual desktops

This section shows the virtual desktop performance results for Lenovo ThinkAgile HX3320 appliances each configured with dual Xeon Scalable 6230 processors, 768 GB of memory, two 1.92 TB SATA SSDs, and six 2 TB SATA disk drives.

The recommended configuration of the Nutanix CVM is as follows:

• vCPU	12
• CPU Reservation	10000 MHz
• Memory	24GB
• Memory Reservation	24GB
• NUMA	No Affinity
• Advance CPU – Scheduling Affinity	No Affinity

Table 7 lists the Login VSI performance results of a HX Series appliance 4 node cluster using VMware ESXi 6.7 U1 and Windows 10 with the latest security fixes and the new Xeon Scalable Processors generation 2.

Table 7: Login VSI Performance with VMware ESXI 6.7 U1

Processor	**Workload**	**Stateless**	**Dedicated**
Two Scalable 6230 processors 2.10 GHz, 20C 125W	Office worker	860 users	848 users
Two Scalable 6230 processors 2.10 GHz, 20C 125W	Knowledge worker	705 users	744 users
Two Scalable 6230 processors 2.10 GHz, 20C 125W	Power worker	599 users	607 users

Table 8 lists the Login VSI performance results of a HX Series appliance 4 node cluster using Nutanix AHV 5.1.3 and Windows 10. These results will be updated with the Intel 6230 processor at a later time.

Table 8: Login VSI Performance with Nutanix AHV 5.1.3

Processor	**Workload**	**Stateless**	**Dedicated**
Two Scalable 6130 processors 2.10 GHz, 16C 125W	Office worker	870 users	909 users
Two Scalable 6130 processors 2.10 GHz, 16C 125W	Knowledge worker	622 users	649 users

4.6.3 Performance results from boot storm testing

A boot storm occurs when a substantial number of VMs are all booted within a short period of time. Booting a large number of VMs simultaneously requires large IOPS otherwise the VMs become slow and unresponsive.

Different numbers of VMs were booted on a cluster of 4 HX3320 hybrid appliances. The VMs were unpowered in vCenter and the boot storm created by powering on all of the VMs simultaneously. The time for all of the VMs to become visible in Citrix XenDesktop was measured.

Figure 25 shows a comparison of the boot times for a variety of VMs. With an even spread of VMs on each node, the boot time for the VMs on each node was similar to the overall cluster boot time.

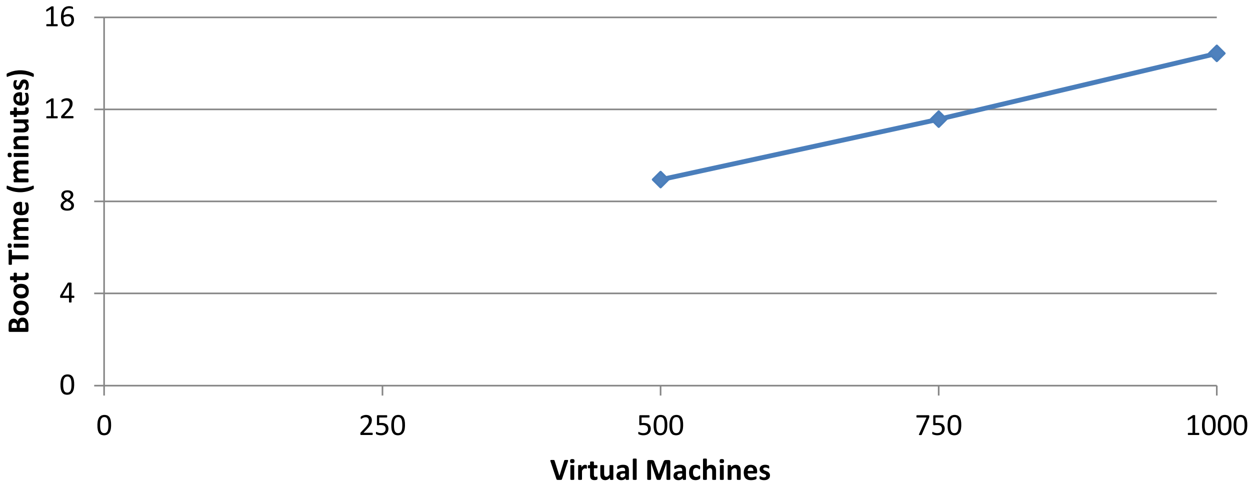


Figure 25: Boot storm comparison

4.7 Performance recommendations

This section provides sizing performance recommendations and best practices.

4.7.1 Sizing recommendations for virtual desktops

The default recommendation is two Xeon Scalable 6130 processors and 768 GB of system memory because this configuration provides the best coverage and density for a range of users. Assuming there is enough storage configured for the VMs, this configuration is recommended for any of the HX 3000 Series appliances.

For an office worker, Lenovo testing shows that 150 users per server is a good baseline and has an average of 83% usage of the processors in the server. If a server goes down, users on that server must be transferred to the remaining servers. For this degraded failover case, Lenovo testing shows that 180 users per server have an average of 88% usage of the processors. It is important to keep this 25% headroom on servers to cope with possible failover scenarios. Lenovo recommends a general failover ratio of 5:1. By using a target of 150 users per server, the maximum number of office workers is 9,600 in a 64 node cluster.

For a knowledge worker, Lenovo testing shows that 125 users per server is a good baseline and has an average of 80% usage of the processors in the server. For the degraded failover case, Lenovo testing shows that 150 users per server have an average of 86% usage of the processors. By using a target of 125 users per server, the maximum number of knowledge workers is 8000 in a 64 node cluster.

For a power worker, Lenovo testing shows that 100 users per server is a good baseline and has an average of 75% usage of the processors in the server, For the degraded failover case, Lenovo testing shows that 125 users per server have an average of 84% usage of the processors. By using target of 100 users per server, the maximum number of power workers is 6400 in a 64 node cluster.

Table 9 summarizes the processor usage with ESXi for the recommended user counts for normal mode and failover mode.

Table 9: Processor usage

Processor	Workload	Users per Server	CPU Utilization
Two 6230	Office worker	150 users – Normal Mode	83%
Two 6230	Office worker	180 users – Failover Mode	88%
Two 6230	Knowledge worker	125 users – Normal Mode	80%
Two 6230	Knowledge worker	150 users – Failover Mode	86%
Two 6230	Power worker	100 users – Normal Mode	75%
Two 6230	Power worker	125 users – Failover Mode	84%

Table 10 lists the recommended number of virtual desktops per server for different workload types and VM memory sizes. The number of users is reduced in some cases to fit within the available memory and still maintain a reasonably balanced system of compute and memory.

Table 10: Recommended number of virtual desktops per server

Workload	Office worker	Knowledge worker	Power worker
Processor	Two 6230	Two 6230	Two 6230
VM memory size	3 GB	4 GB	5 GB
System memory	768 GB	768 GB	768 GB
Memory overhead of CVM	24 GB	24 GB	24 GB
Desktops per server (normal mode)	150	125	100
Desktops per server (failover mode)	180	150	125

Table 11 lists the approximate number of compute servers that are needed for different numbers of users and Office worker workloads.

Table 11: Compute servers needed for Office workers and different numbers of users

Office workers	300 users	600 users	1200 users	3000 users
Compute servers @150 users (normal)	3	5	8	20
Compute servers @180 users (failover)	2	4	7	17

Table 12 lists the approximate number of compute servers that are needed for different numbers of users and Knowledge worker workloads.

Table 12: Compute servers needed for Knowledge workers and different numbers of users

Knowledge workers	300 users	600 users	1200 users	3000 users
Compute servers @125 users (normal)	3	5	10	24

Compute servers @150 users (failover)	2	4	8	20

Table 12 lists the approximate number of compute servers that are needed for different numbers of users and power worker workloads.

Table 13: Compute servers needed for Power workers and different numbers of users

Power workers	**300 users**	**600 users**	**1200 users**	**3000 users**
Compute servers @100 users (normal)	3	6	12	30
Compute servers @125 users (failover)	3	5	10	24

4.7.2 Best practices

The number of desktops that can be run on a specific server depends upon the available system memory, compute power of the processors, and number of logons per second during a logon storm. For a cost-effective solution, the maximum number of users should be put on each server to balance processor, memory, storage I/O, and networking. Lenovo recommends using all flash appliances for situations where the user logon rate is high or time to reboot all the VMs on a node must be less than 10 minutes.

Another important consideration for compute servers is system memory. For stateless users, the typical range of memory that is required for each desktop is 2 GB - 4 GB. For dedicated users, the range of memory for each desktop is 2 GB - 6 GB. In general, power users that require larger memory sizes also require more virtual processors. This reference architecture standardizes on 2 GB per desktop as the minimum requirement of a Windows 10 desktop. The virtual desktop memory should be large enough so that swapping is not needed and vSwap can be disabled.

It is a best practice not to overcommit on memory as swapping to disk can have a severe effect on performance; a better strategy is to give each desktop more memory. Alternatively, a monitoring tool can be run to gather information about existing desktops. The desktop memory size that is required does not necessarily have to match the memory supplied in a desktop machine; it can be larger or smaller.

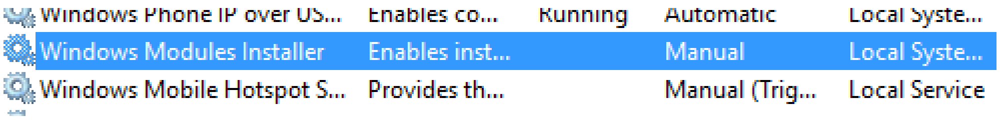
Lenovo recommends the use of VLANs to partition the network traffic. The following VLANs should be used:

- User (for web protocols, display protocols, and support service protocols)
- Management (for management protocols)
- Storage (for NDFS)

Lenovo recommends to always perform user virtualization, even if users have dedicated desktops. This separation of user-specific data makes it much easier to manage and perform upgrades.

Windows 10 was used for all of the performance testing. In general Windows 10 requires 10% to 20% more compute power than Windows 7. The following optimizations are recommended for the Windows 10 base image:

- Applied #VDILIKEAPRO Tuning Template(developed by loginVSI) – see the following for more details: loginvsi.com/blog/520-the-ultimate-windows-10-tuning-template-for-any-vdi-environment
- Set Adobe acrobat as a default app for PDF files using steps in following webpage: adobe.com/devnet-docs/acrobatetk/tools/AdminGuide/pdfviewer.html
- Disabled **Windows Modules installer** service on the base image because the CPU utilization can remain high after rebooting all the VMs. By default this service is set to manual rather than disabled.



Windows Phone IP over US...	Enables co...	Running	Automatic	Local Syste...
Windows Modules Installer	Enables inst...		Manual	Local Syste...
Windows Mobile Hotspot S...	Provides th...		Manual (Trig...	Local Service

Please refer to the websites below for best practices and optimizations recommended by Citrix:

- Windows 10 Optimization Guide: support.citrix.com/article/CTX216252

- Citrix Virtual Desktop Handbook 7.x:
  support.citrix.com/article/CTX139331

4.8 Deployment ready solutions

There are 3 deployment ready solutions for Citrix XenDesktop in the Lenovo Data Center Solution Configurator (DCSC): https://dcsc.lenovo.com/#/reference-solution.

Solution Identifier	SID0000115	SID0000117	SID0000119
ThinkAgile HX node	HX3720 or HX3721	HX3320 or HX3321	HX3520-G or HX3521-G
Office Workers (3GB)	180 users/node	180 users/node	N/A
Knowledge workers (4GB)	150 users/node	150 users/node	N/A
Power users (5GB)	N/A	125 users/node	N/A
Power user with GPU (6GB)	N/A	N/A	48 users/node
Light Cad with GPU (8GB)	N/A	N/A	16 users/node
Heavy Cad with GPU (16GB)	N/A	N/A	8 users/node

5 Microsoft Exchange

Microsoft Exchange Server 2016 is the market leader in enterprise messaging and collaboration. With increasing processor performance, the primary design goal for Exchange 2016 is simplicity of scale, hardware utilization, and failure isolation. With Exchange 2016, the number of server roles is reduced to two: Edge Transport server and the Mailbox server which includes client access and mailbox services. This chapter shows the performance of two scenarios using a 4 node cluster of Lenovo ThinkAgile HX7520 appliances: 30,000 mailboxes with hybrid storage and 60,000 mailboxes with all flash storage. This chapter applies equally to the ThinkAgile HX7521 certified nodes.

5.1 Solution overview

Figure 26 shows the architectural overview of the Microsoft Exchange solution. This chapter does not address integrating Exchange with unified messaging solutions and handling edge transport routing and distribution.

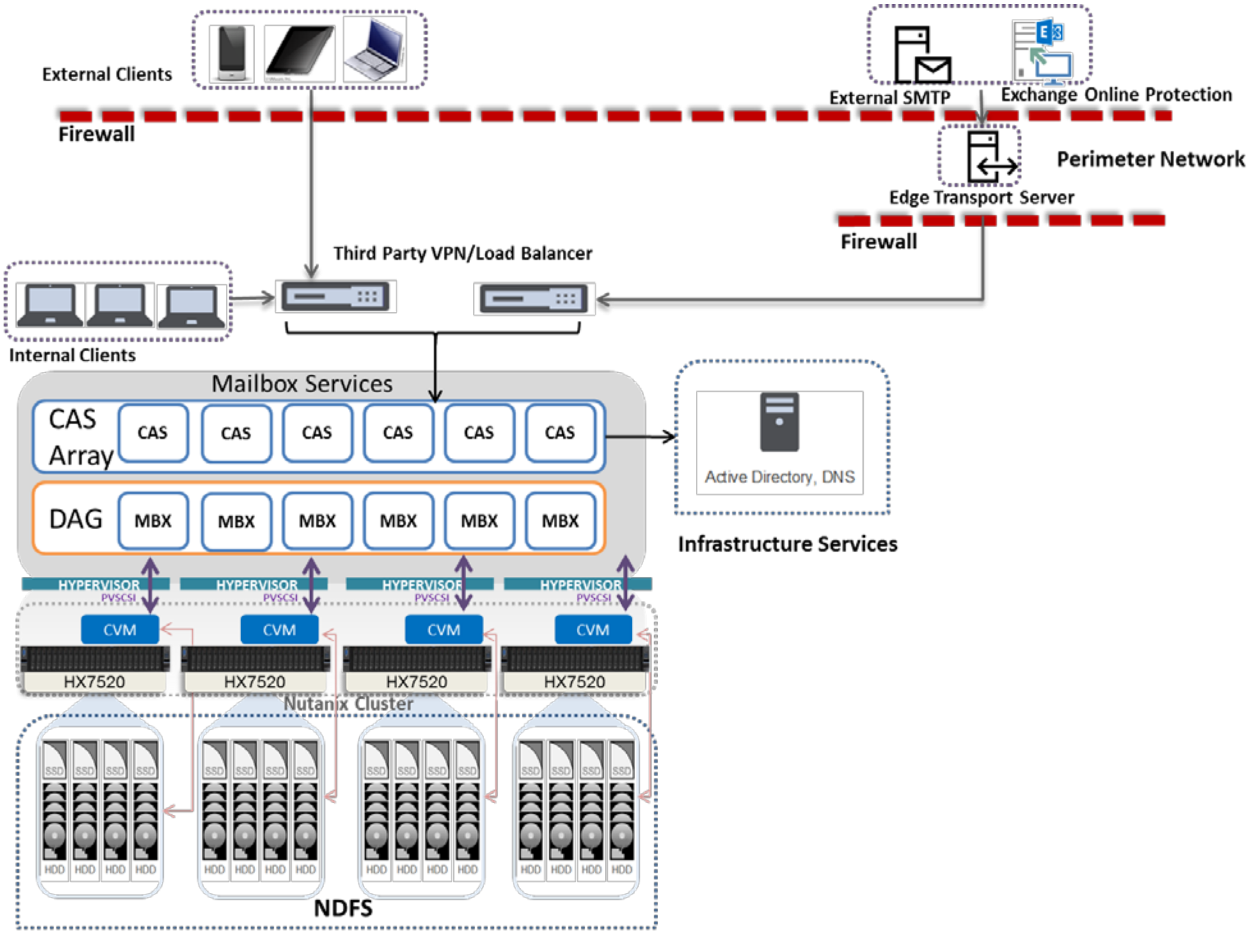


Figure 26. Lenovo ThinkAgile HX Series solution with Microsoft Exchange

The Client Access Server (CAS) role which is no longer a separate entity and is automatically installed with the Mailbox Server role still provides client protocols and unified messaging support. The Mailbox Server (MBX) role provides all of the data processing services. All external mail through SMTP is now handled via a separately installed Edge Transport Server, which usually resides within the perimeter network.

For load balancing into the CAS layer either a network load balancer can be used with a CAS array object or a layer 4 or layer 7 load balancer can be used without the need for configuring a CAS array.

5.2 Component model

This section describes the logical component view of the Exchange Server environment. Figure 27 shows a high-level component model.

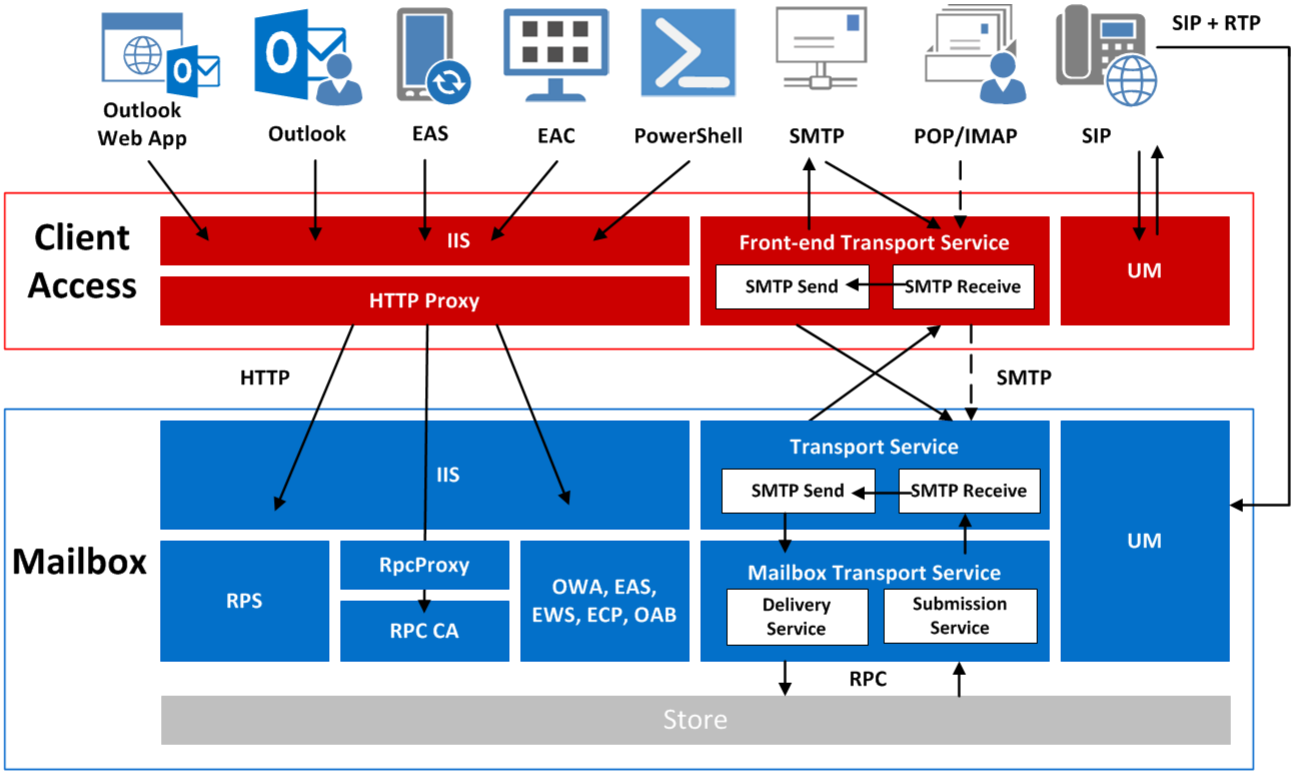


Figure 27. Exchange Server logical component view

The following basic concepts and terminology are used throughout this section:

Exchange Admin Center (EAC) – The EAC is the web-based management console in Microsoft Exchange Server that is optimized for on-premises, online, and hybrid Exchange deployments. The EAC replaces the Exchange Management Console (EMC) and the Exchange Control Panel (ECP), which were the two interfaces used to manage Exchange Server 2010.

Exchange Control Panel (ECP) – The ECP is a web application that runs on a Client Access Server and provides services for the Exchange organization.

Exchange Web Services (EWS) – EWS provides the functionality to enable client applications to communicate with the Exchange server.

Internet Information Services (IIS) – IIS is an extensible web server that was created by Microsoft for use with Windows NT family.

Internet Message Access Protocol (IMAP) – IMAP is a communications protocol for email retrieval and storage developed as an alternative to POP.

Microsoft Exchange ActiveSync (EAS) – EAS is a communications protocol that is designed for the synchronization of email, contacts, calendar, tasks, and notes from a messaging server to a smartphone or other mobile device.

Microsoft Outlook® Web App (OWA) – OWA (formerly Outlook Web Access) is a browser-based email client with which users can access their Microsoft Exchange Server mailbox from almost any web browser.

Offline Address Book (OAB) – The OAB is a copy of an address list collection that was downloaded so a Microsoft Outlook user can access the address book while disconnected from the server. Microsoft Exchange generates the new OAB files and then compresses the files and places them on a local share.

Outlook Anywhere – Outlook Anywhere is a service that provides RPC/MAPI connectivity for Outlook clients over HTTP or HTTPS by using the Windows RPC over HTTP component.

Post Office Protocol (POP) – The POP is an application-layer Internet standard protocol that is used by local email clients to retrieve email from a remote server over a TCP/IP connection

Real-time Transport Protocol (RTP) – RTP is a network protocol for delivering audio and video over IP networks.

Remote PowerShell (RPS) – RPS allows you to use Windows PowerShell on your local computer to create a remote Shell session to an Exchange server if you do not have the Exchange management tools installed.

RPC Client Access (RPC) – In Microsoft Exchange Server 2007, the Client Access server role was introduced to handle incoming client connections to Exchange mailboxes. Although most types of client connections were made to the Client Access server, Microsoft Office Outlook still connected directly to the Mailbox server when it was running internally with the MAPI protocol.

A new service was introduced with Exchange Server 2010 to allow these MAPI connections to be handled by the Client Access server. The RPC Client Access service provides data access through a single, common path of the Client Access server, with the exception of public folder requests (which are still made directly to the Mailbox server). This change applies business logic to clients more consistently and provides a better client experience when failover occurs.

Remote Procedure Call over HTTP – In Exchange 2016 this feature has been replaced by **MAPI over HTTP** offering improvements over the traditional Outlook anywhere (RPC over HTTP). In Exchange 2016 the MAPI over HTTP feature is enabled by default because Exchange 2016 which does not allow direct RPC connectivity.

Session Initiation Protocol (SIP) – SIP is a protocol that is used for starting, modifying, and ending an interactive user session that involves multimedia elements, such as video, voice, and instant messaging.

Simple Mail Transfer Protocol (SMTP) – SMTP is an Internet standard for email transmission.

Unified Messaging (UM) – UM allows an Exchange Server mailbox account that was enabled for UM to receive email, voice, and fax messages in the Inbox.

5.3 30,000 Mailbox Exchange Performance with Hybrid Storage

This section describes the deployment configuration and performance of a 4 node cluster using Lenovo ThinkAgile HX7520 appliances with hybrid storage to support 30,000 mailboxes.

5.3.1 DAG architecture overview

The following section illustrates the Lenovo ThinkAgile HX7520 based Exchange 2016 mailbox resiliency solution Lenovo implemented for ESRP testing.

Compute per node:	2 x Intel(R) Xeon(R) Platinum 8170 CPU @ 2.10GHz Processors
RAM per node:	768GB
Raw storage per node:	SSD: 4 x 1.92TB SATA SSDs – 7.68TB HDD: 20 x 2TB SATA HDDs – 40TB
Raw storage per cluster:	SSD: 16 x 1.92 TB SATA SSDs – 31TB HDD: 80 x 2TB SATA HDDs – 160 TB

Figure 28 describes the high level DAG architecture of the 30,000 mailbox virtualized Exchange 2016 mailbox resiliency solution. This solution comprises of two Lenovo ThinkAgile HX7520 clusters with 4 nodes per cluster. Each node uses the Nutanix Acropolis Hypervisor (AHV).

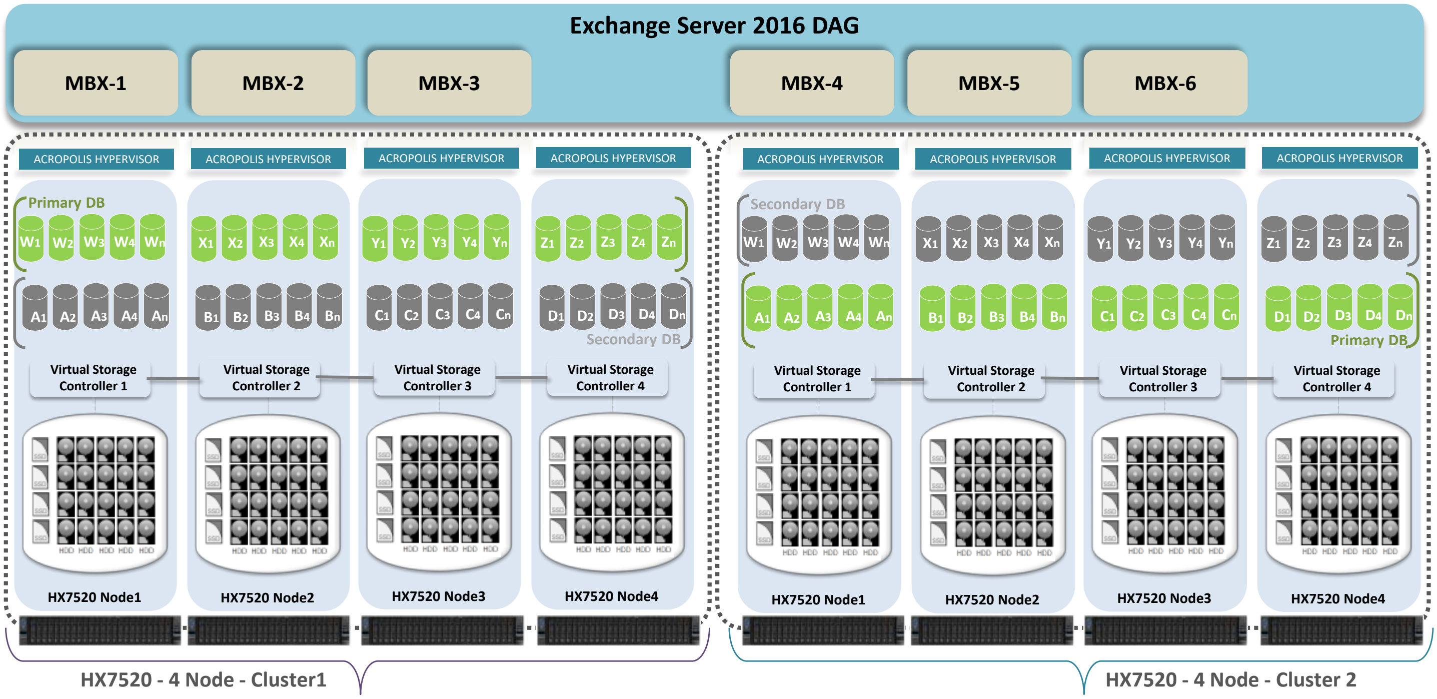


Figure 28. DAG Architecture using Exchange 2016 and Lenovo ThinkAgile HX7520 (30,000 Mailboxes)

The DAG has six Exchange 2016 mailbox servers and two database copies. The two database copies were placed on two physically isolated clusters. On both Cluster1 and Cluster2, the Mailbox server (MBX) VMs were created on three nodes. All the Database/Logs volumes were connected to Mailbox servers using the iSCSI protocol. The primary database copy and secondary database copy are stored on two physically separated and isolated clusters. The two clusters can be located at the same datacentre or two different datacenters.

Each Mailbox Server VM is configured as follows:

- 40 vCPUs
- 96 GB RAM
- Windows Server 2016
- Microsoft Exchange 2016 with 10,000 mailboxes

The ESRP-Storage program focuses on storage solution testing to address performance and reliability issues with storage design. However, storage is not the only factor to take into consideration when designing a scale up Exchange solution. Other factors which affect the server scalability are: server processor utilization, server physical and virtual memory limitations, resource requirements for other applications, directory and network service latencies, network infrastructure limitations, replication and recovery requirements, and client usage profiles. All these factors are beyond the scope for ESRP-Storage. Therefore, the number of mailboxes hosted per server as part of the tested configuration may not necessarily be viable for some customer deployment.

For more information on identifying and addressing performance bottlenecks in an Exchange system, please refer to Microsoft's Troubleshooting Microsoft Exchange Server Performance, available at http://technet.microsoft.com/en-us/library/dd335215.aspx.

5.3.2 Targeted customer profile

The target customer profile for a medium enterprise Microsoft Exchange 2016 environment is as follows:

- 30,000 mailboxes of 1GB
- 6x Exchange 2016 servers (3x Tested)
- 0.06 IOPS per mailbox
- 24/7 background database maintenance
- Mailbox resiliency factor of 2
- 10 databases per host

5.3.3 Tested deployment environment

The section describes the tested deployment environment.

Simulated exchange configuration

The following table summarizes the simulated Exchange configuration.

Number of Exchange mailboxes simulated	30,000
Number of Database Availability Groups (DAGs)	1
Number of servers/DAG	6 (3 tested)
Number of active mailboxes/server	5,000 (5,000 active and 5,000 passive mailboxes per server and tested 10,000 active mailboxes per server)

Number of databases/host	10
Number of copies/database	2
Number of mailboxes/database	1000
Simulated profile: I/O’s per second per mailbox (IOPS, include 20% headroom)	0.06 IOPs / Mailbox
Database/LUN size	1.5 TB
Total database size for performance testing	30 TB
% storage capacity used by Exchange database[1]	38.3%

Storage hardware

The following table summarizes the storage hardware.

Storage Connectivity (Fiber Channel, SAS, SATA, iSCSI)	iSCSI
Storage model and OS/firmware revision	HX7520 running Acropolis 5.1.3
Storage cache	47.69 GB per node
Number of storage controllers	4x virtual controller virtual machines
Number of storage ports	4 x 10 Gbe Port
Maximum bandwidth of storage connectivity to host	40 Gbps per node
Switch type/model/firmware revision	Lenovo RackSwitch G8272 (10GbE) Firmware version: 7.7.5
HBA model and firmware	Lenovo ThinkSystem 430-8i HBA
Number of HBA’s/host	3
Host server type	3x Lenovo ThinkAgile HX7520 (2 x Intel(R) Xeon(R) Platinum 8170 CPU @ 2.10GHz) 768 GB RAM
Total number of disks tested in solution	96 (4 node cluster)
Maximum number of spindles can be hosted in the storage	96 (cluster can be scaled to 40+ nodes)

[1] Storage performance characteristics change based on the percentage utilization of the individual disks. Tests that use a small percentage of the storage (~25%) may exhibit reduced throughput if the storage capacity utilization is significantly increased beyond what is tested in this paper.

Storage software

The following table summarizes the storage software.

HBA driver	Nutanix Virt I/O SCSI Pass-thru Driver 62.62.101.5800
HBA QueueTarget Setting	N/A
HBA QueueDepth Setting	N/A
Hypervisor	Nutanix Acropolis Hypervisor (AHV)
Exchange VM guest OS	Windows Server 2016
ESE.dll file version	15.00.0847.030
Replication solution name/version	N/A

Storage disk configuration (mailbox store disks)

The following table summarizes the storage disk configuration.

Disk type, speed and firmware revision	Per Node: 4x Intel 1.92TB 6Gbps SATA G3HS 2.5" SSD 20x 2 TB 7.2K 6Gbps NL SATA 2.5" G3HS 512e HDD
Raw capacity per disk (GB)	2048 GB (2 TB)
Number of physical disks in test	96 (16x 1.92 TB + 80x 2 TB)
Total raw storage capacity (GB)	195,297GB
Disk slice size (GB)	N/A
Number of slices or disks per LUN	N/A
Raid level	Nutanix Replication Factor 2 (RAID 1)
Total formatted capacity	78.33 TB
Storage capacity utilization	41%
Database capacity utilization	15.7%

5.3.4 Performance test results

This section provides a high-level summary of the results of executing the Microsoft ESRP storage test version 4.0 on the configuration of 4 Lenovo ThinkAgile HX7520 appliances as described in the previous section. ESRP storage test results include reliability, storage performance, and database backup/restore.

Note that the ESRP program is not designed to be a benchmarking program and tests are not designed to get the maximum throughput for a giving solution. Rather, the program is focused on producing recommendations from vendors for the Exchange application. Therefore, the data presented in this document should not be used for direct comparisons among the solutions and customers should not quote the data directly for their pre-deployment verifications. It is recommended that a proof of concept is carried out to validate the storage design for a specific customer environment.

The results in this section were developed by Lenovo and reviewed by the Microsoft Exchange Product team.

Reliability

Several of the tests in the ESP test framework are used to check reliability and run for 24 hours. The test objective is to verify that the storage can handle high I/O workloads for extensive periods. Log and database files are analyzed for integrity after the stress test to ensure there is no database or log corruption.

Executing this test on the Lenovo ThinkAgile HX7520 appliances showed:

- No errors reported in the saved event log file.
- No errors reported during the database and log checksum process.

Storage performance results

The primary storage performance test in the ESP test framework is designed to exercise the storage with a maximum sustainable Exchange I/O pattern for 2 hours. The purpose is to reveal how long it takes for the storage to respond to I/O operations under a load.

Table 14 shows the sum of I/O's and the average latency across all storage groups on a per server basis.

Table 14: Individual Server Performance

	Node 1	**Node 2**	**Node 3**
Database I/O			
Database Disks Transfers/sec	1564.80	1460.86	1464.32
Database Disks Reads/sec	1061.10	991.52	993.425
Database Disks Write/sec	503.66	469.34	470.90
Average Database Disk Read Latency (ms)	6.90	7.55	7.56
Average Database Disk Write Latency (ms)	3.90	4.56	4.66
Transaction Log I/O			
Log Disks Writes/sec	11.8	11.03	11.05
Average Log Disk Write Latency(ms)	0.98	0.91	0.91

Table 15 shows the sum of I/O's and the average latency across the 3 primary servers in the solution.

Table 15: Aggregate Server Performance

Database I/O	
Database Disks Transfers/sec	4489.98
Database Disks Reads/sec	3046.05
Database Disks Writes/sec	1443.89
Average Database Disk Read Latency (ms)	7.35
Average Database Disk Write Latency (ms)	4.39
Transaction Log I/O	
Log Disks Writes/sec	33.89
Average Log Disk Write Latency (ms)	1.70

Database backup/recovery performance

Several of the tests in the ESP test framework are used to measure the sequential read rate of the database files and the recovery/replay performance (playing transaction logs into the database).

The database read-only performance test measures the maximum rate at which databases could be backed up using Microsoft Volume Shadow Copy Service (VSS). Table 16 shows the average read performance for a backing up a single database file and all ten database files on a single node.

Table 16: Database backup read-only performance results

MB read/sec per database	67.07
MB read/sec total per node (10 databases)	670.70

Transaction log recovery/replay performance

The test is to measure the maximum rate at which the log files can be played against the databases. Table 17 shows the average rate for 500 log files played in a single storage group. Each log file is 1 MB in size.

Table 17: Transaction Log Recovery/Replay Performance

Average time to play one Log file (sec)	0.42

5.4 60,000 Mailbox Exchange Performance with AF Storage

This section describes the deployment configuration and performance of a 4 node cluster using Lenovo ThinkAgile HX7520 appliances with all flash storage to support 60,000 mailboxes.

5.4.1 DAG architecture overview

The following section illustrates the Lenovo ThinkAgile HX7520 based Exchange 2016 mailbox resiliency solution Lenovo implemented for ESRP testing.

Compute per node:	2 x Intel(R) Xeon(R) Platinum 8170 CPU @ 2.10GHz Processors
RAM per node:	768GB
Raw storage per node:	SSD: 24 x 1.92TB SATA SSDs – 46.08TB
Raw storage per cluster:	SSD: 96 x 1.92 TB SATA SSDs – 184.32TB

Figure 28 describes the high level DAG architecture of the 60,000 mailbox virtualized Exchange 2016 mailbox resiliency solution. This solution comprises of two Lenovo ThinkAgile HX7520 clusters with 4 nodes per cluster. Each node uses the Nutanix Acropolis Hypervisor (AHV).

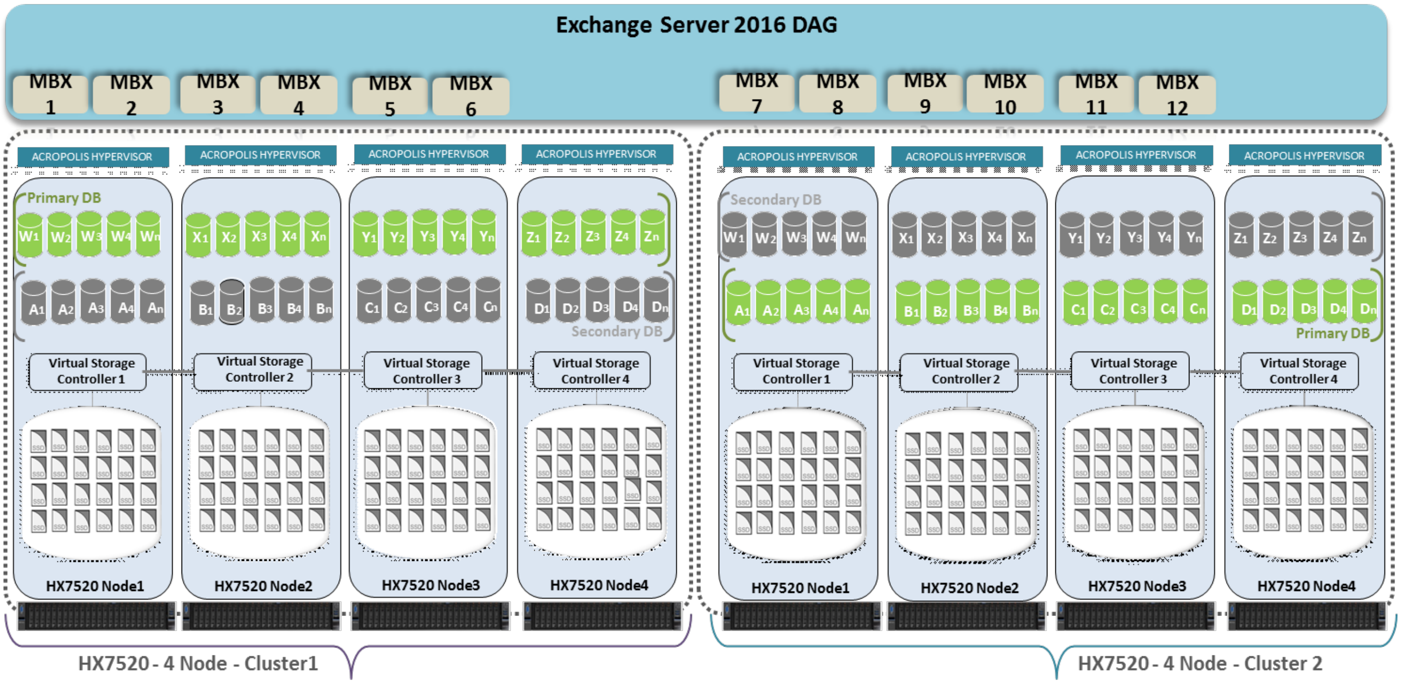


Figure 29. DAG Architecture using Exchange 2016 and Lenovo ThinkAgile HX7520 (60,000 Mailboxes)

The DAG has twelve Exchange 2016 mailbox servers and two database copies. The two database copies were placed on two physically isolated clusters. On both Cluster1 and Cluster2, the Mailbox server (MBX) VMs were created on three nodes. All the Database/Logs volumes were connected to Mailbox servers using the iSCSI protocol. The primary database copy and secondary database copy are stored on two physically separated and isolated clusters. The two clusters can be located at the same datacentre or two different datacenters.

Each Mailbox Server VM is configured as follows:

- 40 vCPUs
- 96 GB RAM
- Windows Server 2016
- Microsoft Exchange 2016 with 10,000 mailboxes

The ESRP-Storage program focuses on storage solution testing to address performance and reliability issues with storage design. However, storage is not the only factor to take into consideration when designing a scale up Exchange solution. Other factors which affect the server scalability are: server processor utilization, server physical and virtual memory limitations, resource requirements for other applications, directory and network service latencies, network infrastructure limitations, replication and recovery requirements, and client usage profiles. All these factors are beyond the scope for ESRP-Storage. Therefore, the number of mailboxes hosted per server as part of the tested configuration may not necessarily be viable for some customer deployment.

For more information on identifying and addressing performance bottlenecks in an Exchange system, please refer to Microsoft's Troubleshooting Microsoft Exchange Server Performance, available at http://technet.microsoft.com/en-us/library/dd335215.aspx.

5.4.2 Targeted customer profile

The target customer profile for a medium enterprise Microsoft Exchange 2016 environment is as follows:

- 60,000 mailboxes of 0.75GB
- 12x Exchange 2016 servers (6x Tested)
- 0.06 IOPS per mailbox
- 24/7 background database maintenance
- Mailbox resiliency factor of 2
- 10 databases per host

5.4.3 Tested deployment environment

The section describes the tested deployment environment.

Simulated exchange configuration

The following table summarizes the simulated Exchange configuration.

Number of Exchange mailboxes simulated	60,000
Number of Database Availability Groups (DAGs)	1
Number of servers/DAG	12 (6 tested)
Number of active mailboxes/server	10,000 (10,000 active and 10,000 passive mailboxes per server and tested 20,000 active mailboxes per server)
Number of databases/host	10

Number of copies/database	2
Number of mailboxes/database	1000
Simulated profile: I/O’s per second per mailbox (IOPS, include 20% headroom)	0.06 IOPs / Mailbox
Database/LUN size	800 GB
Total database size for performance testing	32 TB
% storage capacity used by Exchange database[2]	44.8%

Storage hardware

The following table summarizes the storage hardware.

Storage Connectivity (Fiber Channel, SAS, SATA, iSCSI)	iSCSI
Storage model and OS/firmware revision	HX7520 running Acropolis 5.1.3
Storage cache	N/A
Number of storage controllers	4x virtual controller virtual machines
Number of storage ports	4 x 10 Gbe Port
Maximum bandwidth of storage connectivity to host	40 Gbps per node
Switch type/model/firmware revision	Lenovo RackSwitch G8272 (10GbE) Firmware version: 7.7.5
HBA model and firmware	Lenovo ThinkSystem 430-8i HBA
Number of HBA’s/host	3
Host server type	3x Lenovo ThinkAgile HX7520 (2 x Intel(R) Xeon(R) Platinum 8170 CPU @ 2.10GHz) 768 GB RAM
Total number of disks tested in solution	96 (4 node cluster)
Maximum number of spindles can be hosted in the storage	96 (cluster can be scaled to 40+ nodes)

[2] Storage performance characteristics change based on the percentage utilization of the individual disks. Tests that use a small percentage of the storage (~25%) may exhibit reduced throughput if the storage capacity utilization is significantly increased beyond what is tested in this paper.

Storage software

The following table summarizes the storage software.

HBA driver	Nutanix Virt I/O SCSI Pass-thru Driver 62.62.101.5800
HBA QueueTarget Setting	N/A
HBA QueueDepth Setting	N/A
Hypervisor	Nutanix Acropolis Hypervisor (AHV)
Exchange VM guest OS	Windows Server 2016
ESE.dll file version	15.00.0847.030
Replication solution name/version	N/A

Storage disk configuration (mailbox store disks)

The following table summarizes the storage disk configuration.

Disk type, speed and firmware revision	Per Node: 24 x Intel 1.92TB 6Gbps SATA G3HS 2.5" SSD
Raw capacity per disk (GB)	1966 GB (1.92 TB)
Number of physical disks in test	96 x 1.92 TB
Total raw storage capacity (GB)	188,736 GB
Disk slice size (GB)	N/A
Number of slices or disks per LUN	N/A
Raid level	Nutanix Replication Factor 2 (RAID 1)
Total formatted capacity	71.33 TB
Storage capacity utilization	39%
Database capacity utilization	17.6%

5.4.4 Performance test results

This section provides a high-level summary of the results of executing the Microsoft ESRP storage test version 4.0 on the configuration of 4 Lenovo ThinkAgile HX7520 appliances as described in the previous section. ESRP storage test results include reliability, storage performance, and database backup/restore.

Note that the ESRP program is not designed to be a benchmarking program and tests are not designed to get the maximum throughput for a giving solution. Rather, the program is focused on producing recommendations from vendors for the Exchange application. Therefore, the data presented in this document should not be used for direct comparisons among the solutions and customers should not quote the data directly for their pre-deployment verifications. It is recommended that a proof of concept is carried out to validate the storage design for a specific customer environment.

The results in this section were developed by Lenovo and reviewed by the Microsoft Exchange Product team.

Reliability

Several of the tests in the ESP test framework are used to check reliability and run for 24 hours. The test objective is to verify that the storage can handle high I/O workloads for extensive periods. Log and database files are analyzed for integrity after the stress test to ensure there is no database or log corruption.

Executing this test on the Lenovo ThinkAgile HX7520 appliances showed:

- No errors reported in the saved event log file.
- No errors reported during the database and log checksum process.

Storage performance results

The primary storage performance test in the ESP test framework is designed to exercise the storage with a maximum sustainable Exchange I/O pattern for 2 hours. The purpose is to reveal how long it takes for the storage to respond to I/O operations under a load.

Table 18 shows the sum of I/O's and the average latency across all storage groups on a per server basis.

Table 18: Individual Server Performance

	Node 1	**Node 2**	**Node 3**
Database I/O			
Database Disks Transfers/sec	3559.83	3441.53	3684.05
Database Disks Reads/sec	2530.87	2445.83	2514.00
Database Disks Write/sec	1210.10	1171.54	1260.44
Average Database Disk Read Latency (ms)	6.44	7.56	6.40
Average Database Disk Write Latency (ms)	8.14	10.24	8.26
Transaction Log I/O			
Log Disks Writes/sec	15.65	15.12	15.97
Average Log Disk Write Latency(ms)	2.58	2.31	2.63

Table 19 shows the sum of I/O’s and the average latency across the 3 primary servers in the solution.

Table 19: Aggregate Server Performance

Database I/O	
Database Disks Transfers/sec	10685.41
Database Disks Reads/sec	7490.70
Database Disks Writes/sec	3642.08
Average Database Disk Read Latency (ms)	6.8
Average Database Disk Write Latency (ms)	8.88
Transaction Log I/O	
Log Disks Writes/sec	46.74
Average Log Disk Write Latency (ms)	2.51

Database backup/recovery performance

Several of the tests in the ESP test framework are used to measure the sequential read rate of the database files and the recovery/replay performance (playing transaction logs into the database).

The database read-only performance test measures the maximum rate at which databases could be backed up using Microsoft Volume Shadow Copy Service (VSS). Table 20 shows the average read performance for a backing up a single database file and all 20 database files on a single node.

Table 20: Database backup read-only performance results

MB read/sec per database	133.32
MB read/sec total per node (10 databases)	1330.05

Transaction log recovery/replay performance

The test is to measure the maximum rate at which the log files can be played against the databases. Table 21 shows the average rate for 500 log files played in a single storage group. Each log file is 1 MB in size.

Table 21: Transaction Log Recovery/Replay Performance

Average time to play one Log file (sec)	3.25

5.5 Exchange deployment best practices

This section describes recommended best practices for Microsoft Exchange mailboxes. See also this website for Nutanix Best Practices Guide: Virtualizing Microsoft Exchange: go.nutanix.com/virtualizing-microsoft-exchange-converged-infrastructure.html.

5.5.1 Data optimization

By default all Nutanix storage containers are thin provisioned which reduces unused capacity and automatically provisions additional storage capacity when needed. It is also very easy to add additional storage capacity for mailboxes by simply adding nodes to the cluster. It is also possible to set a storage reservation to guarantee a minimum amount of storage capacity.

Data compression can be used to further increase data capacity especially for data that is less frequently accessed. Lenovo recommends enabling compression with a delay of 1440 minutes (1 day) which minimizes the performance impact on I/O writes.

Data de-duplication is not recommended and should be disabled for active Exchange mailboxes because of the frequency of changes. Note that de-duplication may be beneficial for backup volumes which are not changed very often.

A resiliency factor of 2 is the default. This provides a minimum level of data redundancy but a resiliency factor of 3 might be important in some environments. Using erasure coding saves significant storage capacity but it is only recommended for archive data.

5.5.2 Cluster high availability

The minimum number of nodes in each cluster is 3 and should be at least 4 to provide failover. The following high availability features are recommended for an AHV-based cluster:

A database availability group (DAG) is the base component of the high availability and site resilience framework that is built into Microsoft Exchange Server. A DAG is a group of up to 16 mailbox servers that hosts a set of mailbox databases and provides automatic database-level recovery from failures that affect individual servers or databases.

A DAG is a boundary for mailbox database replication, database and server switchovers, failovers, and an internal component called Active Manager. Active Manager, which runs on every server in a DAG, manages switchovers and failovers.

Any server in a DAG can host a copy of a mailbox database from any other server in the DAG. When a server is added to a DAG, it works with the other servers in the DAG to provide automatic recovery from failures that affect mailbox databases (such as a disk failure or server failure).

Lenovo recommends a DAG configuration of 2 database copies and optionally one lagged copy. With a data resiliency factor of 2, the effective number of copies of each mailbox is 4 and this allows two disk failures without losing data.

DR across datacenters can also be done using DAGs assuming there is sufficient band-width between the sites. The scenarios for active-active and active-passive DR sites using DAGs are outside the scope of this document.

5.5.3 Other best practices

Consider the following points regarding virtualizing Exchange:

- All Exchange server roles should be supported in a single VM.
- Some hypervisors include features for taking snapshots of VMs. However, VM snapshots are not application aware, and the use of snapshots can have unintended and unexpected consequences for a server application that maintains state data, such as Exchange. Therefore, making VM snapshots of an Exchange guest VM is not supported.

5.6 Summary

This chapter shows the performance of two scenarios using a 4 node cluster of Lenovo ThinkAgile HX7520 appliances: 30,000 mailboxes with hybrid storage and 60,000 mailboxes with all flash storage. The performance advantage of an all flash configuration is clearly shown as it easily supports twice as many mailboxes as the hybrid drive configuration by using two VMs per node instead of one. Testing with Exchange 2013 shows very similar performance results to Exchange 2016.

The cluster of 4 Lenovo ThinkAgile HX7520 appliances proved more than capable of handling the high IOPs generated by both scenarios. Part of the reason for this is because the Lenovo ThinkAgile HX7520 appliances use 3 HBAs for the 24 drives. For best performance, it is recommended to use the Nutanix Windows drivers instead of the default Windows drivers for iSCSI. See download.nutanix.com/mobility/1.1.1/Nutanix-VirtIO-1.1.1.msi.

For more details, see the following ESRP reports published by Lenovo and approved by Microsoft:

- 60,000 Mailbox Resiliency Solution for Microsoft Exchange 2016 using all flash Lenovo ThinkAgile HX7520 Appliances and AHV: lenovopress.com/lp0838
- 30,000 Mailbox Resiliency Solution for Microsoft Exchange 2016 using Lenovo ThinkAgile HX7520 Appliances and AHV: lenovopress.com/lp0820
- 30,000 Mailbox Resiliency Solution for Microsoft Exchange 2013 using Lenovo ThinkAgile HX7520 Appliances and AHV: lenovopress.com/lp0819

6 Microsoft SQL Server

Microsoft SQL Server is a database platform for large-scale online transaction processing (OLTP), data warehousing, and a business intelligence platform for data integration, analysis, and reporting solutions. It uses a common set of tools to deploy and manage databases for in-house and cloud environments.

6.1 Solution overview

Figure 30 shows high level architecture of Microsoft SQL Server on Lenovo ThinkAgile HX7520 appliances. This chapter applies equally to the ThinkAgile HX7521 certified nodes.

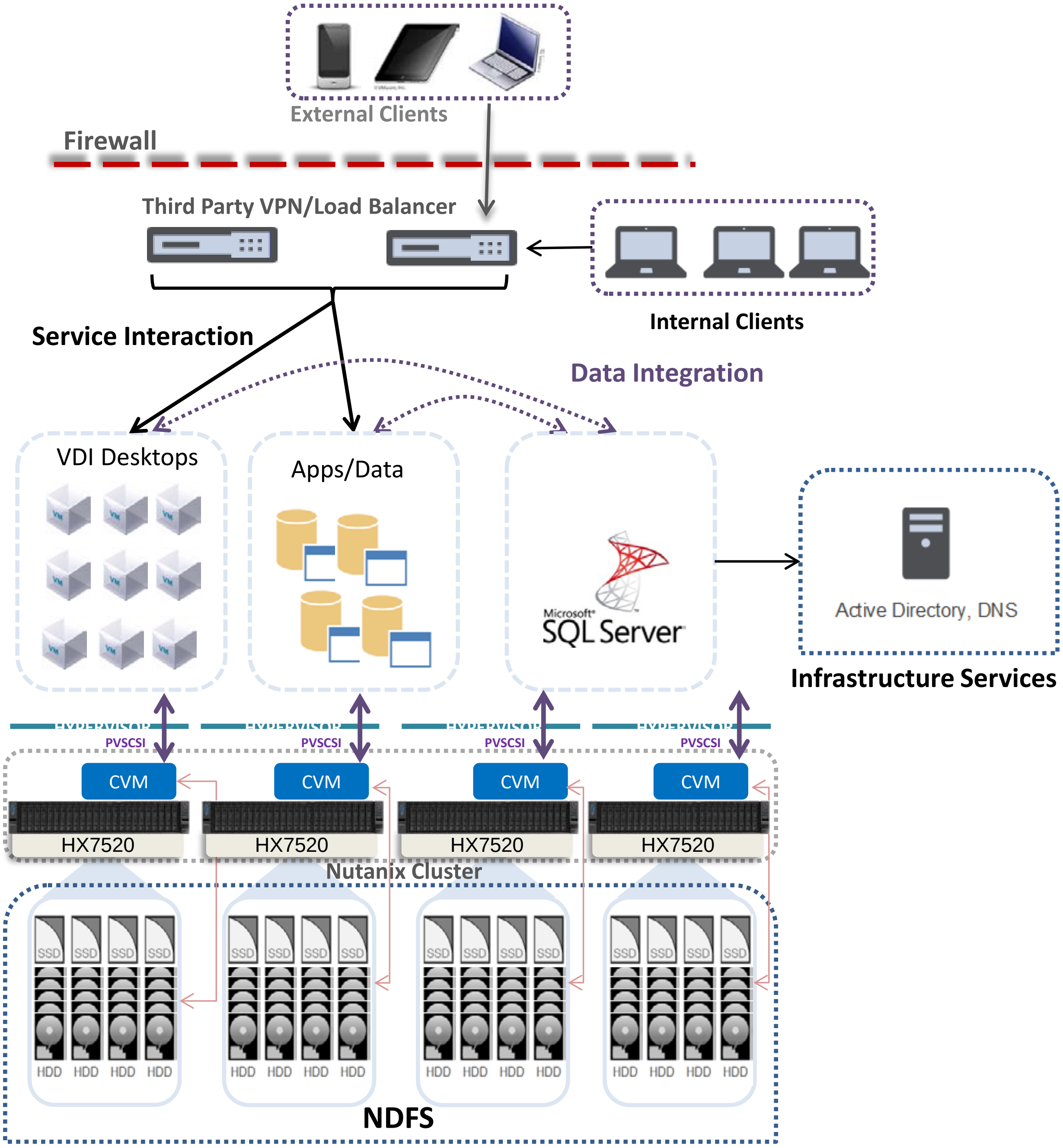
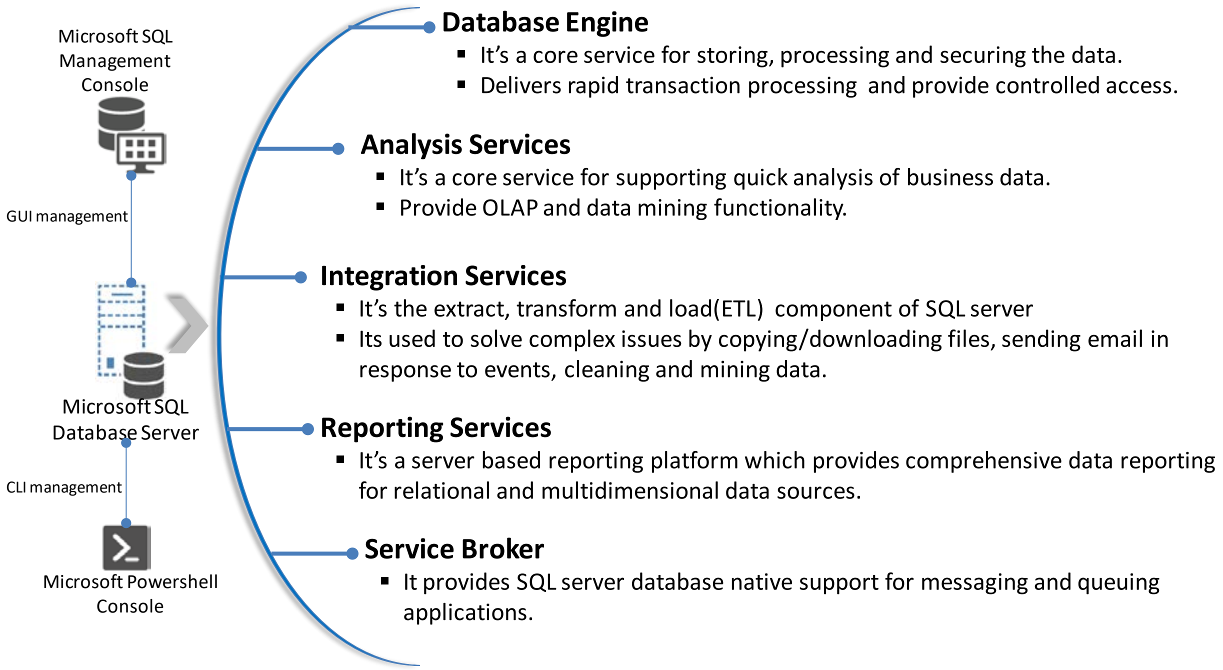


Figure 30: Lenovo ThinkAgile HX Series solution with Microsoft SQL Server

Microsoft SQL Server 2017 can be deployed and can operate in combination with other hosted applications and provides a single scalable platform for all deployments.

6.2 Component model

Figure 31 is a layered component view for Microsoft SQL Server.


Microsoft SQL Management Console
GUI management
Microsoft SQL Database Server
CLI management
Microsoft Powershell Console

Database Engine

- It's a core service for storing, processing and securing the data.
- Delivers rapid transaction processing and provide controlled access.

Analysis Services

- It's a core service for supporting quick analysis of business data.
- Provide OLAP and data mining functionality.

Integration Services

- It's the extract, transform and load(ETL) component of SQL server
- Its used to solve complex issues by copying/downloading files, sending email in response to events, cleaning and mining data.

Reporting Services

- It's a server based reporting platform which provides comprehensive data reporting for relational and multidimensional data sources.

Service Broker

- It provides SQL server database native support for messaging and queuing applications.

Figure 31: Component model with Microsoft SQL Server

Microsoft SQL Server features the following main components:

Database Engine This part of SQL Server actually creates and drives relational databases.

Analysis Services SQL Server Analysis Services (SSAS) is the data analysis component of SQL Server. It can create OLAP (OnLine Analytical Processing) cubes — sophisticated programming objects for organizing data inside a relational database — and do data mining (pulling relevant data out of a database in response to an ad-hoc question).

Integration Services SQL Server Integration Services (SSIS) performs the extract-transform-load (ETL) process that cleans up and formats raw data from source systems for inclusion in the database as ready-to-use information.

Reporting Services SQL Server Reporting Services (SSRS) provides reporting regardless of a database's operating system.

Service Broker SQL Server Service Broker provides native support for messaging and queuing applications which makes it easier to build distributed and reliable applications that use the Database Engine components.

6.3 SQL Server deployment best practices

This section describes recommended best practices to provide data optimization and high availability of Microsoft SQL Server. See also this website for Nutanix Best Practices: Microsoft SQL server 2017: go.nutanix.com/microsoft-sql-server-converged-infrastructure.html.

6.3.1 Data optimization

By default all Nutanix storage containers are thin provisioned which reduces unused capacity and automatically provisions additional storage capacity when needed. It is also very easy to add additional storage capacity for databases by simply adding nodes to the cluster. It is also possible to set a storage reservation amount to guarantee a minimum amount of storage capacity.

Data compression can be used to further increase data capacity especially for data that is less frequently accessed. Lenovo recommends enabling compression with a delay of 1440 minutes (1 day) which minimizes the performance impact on I/O writes.

Data de-duplication is not recommended and should be disabled for SQL Server because of the frequency of changes. Note that de-duplication may be beneficial for backup volumes which are not changed very often.

A resiliency factor of 2 is the default. This provides a minimum level of data redundancy but a resiliency factor of 3 might be important in some environments. Using erasure coding saves significant storage capacity but it only recommended for archive data.

6.3.2 Cluster high availability

The minimum number of nodes in each cluster is 3 and should be at least 4 to provide failover. The following high availability features are recommended for an ESXi-based cluster:

- VMware vSphere high availability (HA) for failover
- VMware vSphere distributed resource scheduler (DRS) for load balancing
- Microsoft AlwaysOn availability groups (AAGs) for data redundancy

VMware vSphere HA pools VMs into a cluster to increase data resiliency. If a host fails, VMware HA moves the VMs to other hosts with spare capacity. Lenovo recommends enabling the “Admission Control Setting” and using the “Admission Control Policy” to set the percentage of cluster resources reserved as failover spare capacity.

VMware vSphere DRS can be used to group ESXi hosts into resource clusters to provide highly available resources and balance workloads. In order to keep the active working set for each SQL Server VM local to the node, Lenovo recommends creating a host group for each node and a “should” rule that keeps each SQL Server VM on a 1 to 1 ratio with the ESXi host. The hosts should be configured with a minimum of N+1 availability.

Because of the Nutanix architecture, Lenovo recommends spreading SQL Server instances across multiple VMs, each with adequate vdisks, rather than using several SQL Server instances on a single VM. Deploying a greater number of smaller VMs also allows advanced hypervisor features such as the distributed resource scheduler (DRS) to better manage workloads. It is easier to place small VMs than very large VMS. The distribution of SQL instance ultimately depends on performance requirements and database size.

The Microsoft AlwaysOn availability groups (AAGs) is a recommended high availability mechanism. It uses a shared-nothing approach where transactions are replicated to other nodes so each node has a full copy of the database and transaction logs. This provides a very fast failover mechanism. The DRS anti-affinity rules need to be used to ensure that the SQL Server VMs are placed on different physical hosts.

DR across datacenters can also be done using AAGs assuming there is sufficient band-width between the sites. The scenarios for active-active and active-passive DR sites using AAGs are outside the scope of this document.

6.3.3 Virtual Disk Configuration

Multiple virtual disks should be used to keep SQL binaries, database, and database logs files and achieve optimal SQL performance. All four SCSI controllers should be utilized and virtual disks should be distributed evenly across controllers as shown in Table 22.

Table 22: Mapping of virtual disks for ESXi and cluster size

Virtual Disk	vSCSI Controller Type	Controller #	Cluster size
Operating System	LSI Logic SAS	0	4 KB
SQL Binaries	LSI Logic SAS	0	4 KB
Backup/Restore	LSI Logic SAS	0	64 KB
SQL Database - 1	VMware Paravirtual	1	64 KB
SQL Database - 2	VMware Paravirtual	1	64 KB
TempDB - 1	VMware Paravirtual	2	64 KB
TempDB - 2	VMware Paravirtual	2	64 KB
TempDB log files	VMware Paravirtual	3	64 KB
Database log files	VMware Paravirtual	3	64 KB

All SQL database and log drives should be formatted with 64KB NTFS cluster size as it enhances the I/O performance without adding any overhead. The OS and SQL binary drives should be formatted with the standard 4KB NTFS cluster size. Drives space utilization should not be above 80% to achieve optimal performance.

To maximize the storage performance of SQL Server VMs, Lenovo recommends using the ESXi Paravirtual SCSI (PVSCSI) adapters. Each PVSCSI adapter can support up to 15 VMDKs.

Lenovo recommends at least eight vdisk per SQL Server VM, spread intelligently across the four available SCSI controllers. Figure 32 shows a mixture of LSI Logic SAS and PVSCSI controllers.

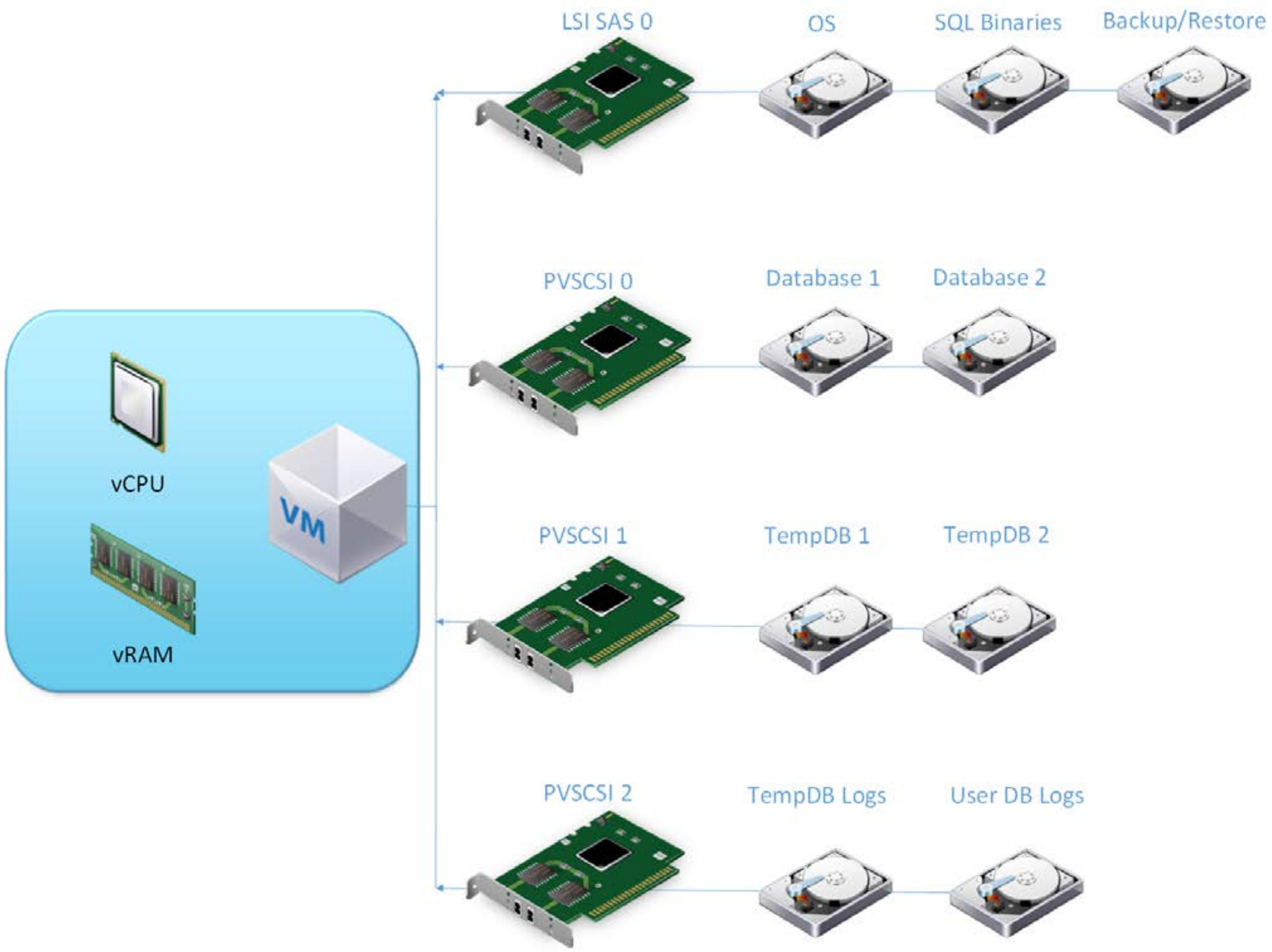


Figure 32: Example controller configuration

6.3.4 SQL Server Files

To achieve high performance, the database should be split into multiple files across multiple virtual disks. In general, one database file per vCPU is ideal. For example a VM with 4 vCPUs hosting 400GB database could be split into four 100GB database files and should spread evenly across the two virtual disks.

For write intensive databases, it is recommended to distribute the database files over four or more virtual disks as it improves the write performance on the back-end and delivers consistent performance. To avoid complexity, the PAGEIOLATCH_XX values should be monitored for contention.

SQL log files (DB and TempDB) are written sequentially, so using multiple log files wouldn't improve the performance. Using a single log file per database is recommended.

TempDB is used as scratch space by the applications and is one of the most important factors of SQL performance. The number of TempDB files to be used is based on the vCPU count. If the vCPUs are less than 8, then configure the same number of TempDB files. If the number of vCPUs is higher than 8, then start with 8 TempDB files and monitor the contention for in-memory allocation (PAGELATCH_XX). The number of TempDB file should be increased in increments of four until contention is eliminated.

It is recommended to create all TempDBs with the same size and not allow for autogrowth. The TempDB file sizing should be done based on required application and is usually 1-10% of the database size.

6.4 Performance test configuration

This section describes the test configuration for Microsoft SQL Server 2017 using a cluster of 4 Lenovo ThinkAgile HX752x appliances. Each appliance has the following configuration:

- 2 x Intel Xeon Scalable 8276 (28 cores @ 2.2 GHz) processors
- 768GB RAM
- 96 x 1.92TB SATA SSDs
- ESXi 6.7 U1
- Set UEFI to “Performance Bias” – for more information, see lenovopress.com/lp0780

From Microsoft SQL licensing guide we see that from SQL Server 2012 onwards, there are two enterprise licenses – Core Based and SERVER/CAL Based. As per the Microsoft license document Compute capacity limits by edition of SQL Server, we found that CAL based license limits SQL Enterprise to use only 20 physical processors. Hence, we will perform two tests based on purchased license, one with core based license using 28 vCPUs per SQL Server, and another test with SERVER/CAL license using only 12 vCPUs per SQL Server.

Each host has ESXi 6.7 U1 and is configured with HA and DRS. The cluster has 8 SQL Server VMs, each configured as follows:

- 28 vCPUs
- 192 GB RAM
- Windows Server 2016
- Microsoft SQL 2017 Enterprise - evaluation

The Nutanix CVM is configured as follows:

- 12 vCPUs
- 96 GB RAM
- CPU affinity 0-55

See also the optimization recommendation in the HammerDB optimization document: hammerdb.com/hammerdb_mssql_oltp_best_practice.pdf.

6.5 Performance test results

This section provides a high-level summary of the results of executing the HammerDB test suite. HammerDB is a graphical open source database load testing and benchmarking tool for Linux and Windows to test databases running on any operating system. HammerDB is automated, multi-threaded and extensible with dynamic scripting support. See this website for more details: hammerdb.com.

This test concentrated on executing 8 VMs on a four node cluster of Lenovo ThinkAgileX752x appliances with two VMs per node. Each cluster has a raw storage of 92TB of SSDs.

A 5,000 scale OLTP database was used for each VM, which equates to about 500GB of database records on the cluster per VM and a total storage of 4TB in the cluster.

Separate load servers were used to simulate the user load. Each OLTP database workload was simulated by running 400 users simultaneously for a total of 3,200 users across the 8 VMs. Each test run had a 5 minute ramp up phase and executed for 15 minutes to simulate 320,000 transactions per user.

Three different scenarios were tested using the ThinkAgile HX752x all flash configuration:

- 12 vCPUs per SQL Server VM
- 28 vCPUs per SQL Server VM
- 44 vCPUs per SQL Server VM

6.5.1 Testing with 12 vCPUs

The first scenario uses 12 vCPUs per VM from the available 56 cores (28 per socket). In this situation a single SQL SERVER/CAL license can be used. It is limited to 20 vCPUs and in this case the physical CPU resources are not overcommitted. Each SQL VM is forced to stay on one vNUMA.

Figure 33 shows the results from this test. The two SQL Server VMs on a single host, one per socket, have a flat profile with a similar transaction rate of 1.8 to 1.9 million transactions per minute.

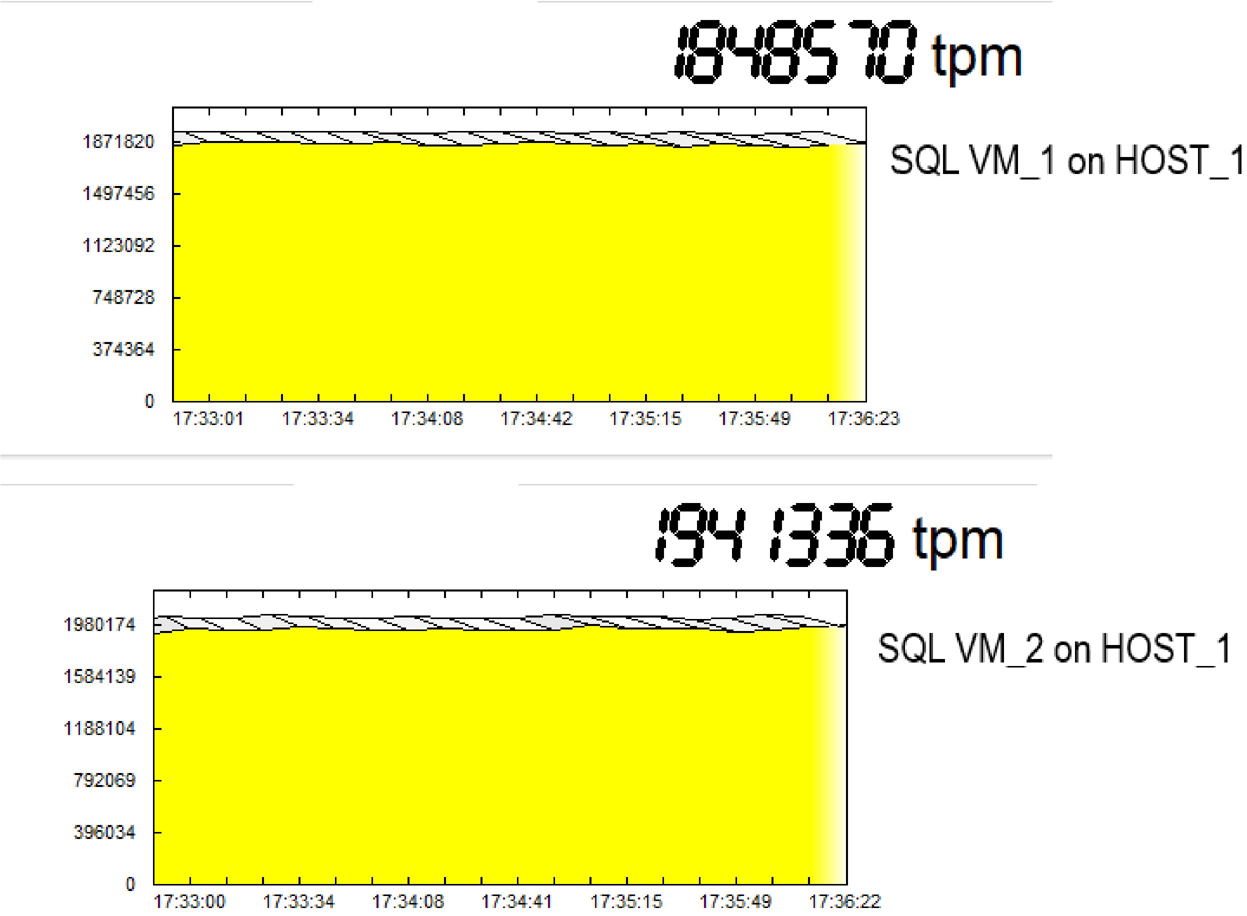


Figure 33: TPM profile obtained during 12 vCPUs tests on two VMs which are situated on same host

Table 23 shows the transactions per minute and new operations per minute results of the HammerDB test using 2 VMs each with 12vCPUs.

Table 23: Microsoft SQL Server results with cluster of four HX752x appliances and 12 vCPUs

Node	VM Name	HammerDB TPM	HammerDB NOPM
Node 1	SQL Instance - 1	1,634,813	355,652
	SQL Instance - 2	1,730,826	376,263
Node 2	SQL Instance - 3	1,668,535	362,940
	SQL Instance - 4	1,701,626	369,669
Node 3	SQL Instance - 5	1,615,699	351,488
	SQL Instance - 6	1,697,651	369,355
Node 4	SQL Instance - 7	1,603,921	348,847
	SQL Instance - 8	1,703,324	370,336
	Total	13,356,395	2,904,550

The 4 node cluster can execute over 13 million transactions per minute and almost 3 million new orders per minute. The CPU utilization is 90-95%.

6.5.2 Testing with 28 vCPUs

The second scenario uses 28 vCPUs per VM from the available 56 cores (28 per socket). This also requires using a CPU based SQL license as the SQL SERVER/CAL license is limited to 20 vCPUs.

In this scenario one socket has a SQL Server VM and the second socket has both the SQL Server VM and the Nutanix CVM completing for resources. The total number of vCPUs is 28 + 28 + 12 = 68, which exceeds the physical cores. In this case hyper-threading can be used to provide the necessary resources but it is less efficient. Benchmarks and actual user experiences have demonstrated that at most, a 30 percent gain can be obtained when transactions are extremely small, e.g., they execute in under 200 milliseconds. Hence, there are small turbulences in the graph for the second socket as there are some transactions waiting for free threads in the CPU.

Figure 34 shows the results from this test. The transaction rates on the second socket is less than the first one because of the CVM. Nevertheless, the transaction rate is twice that achieved with 12 vCPUs per VM because of the higher number of vCPUs.

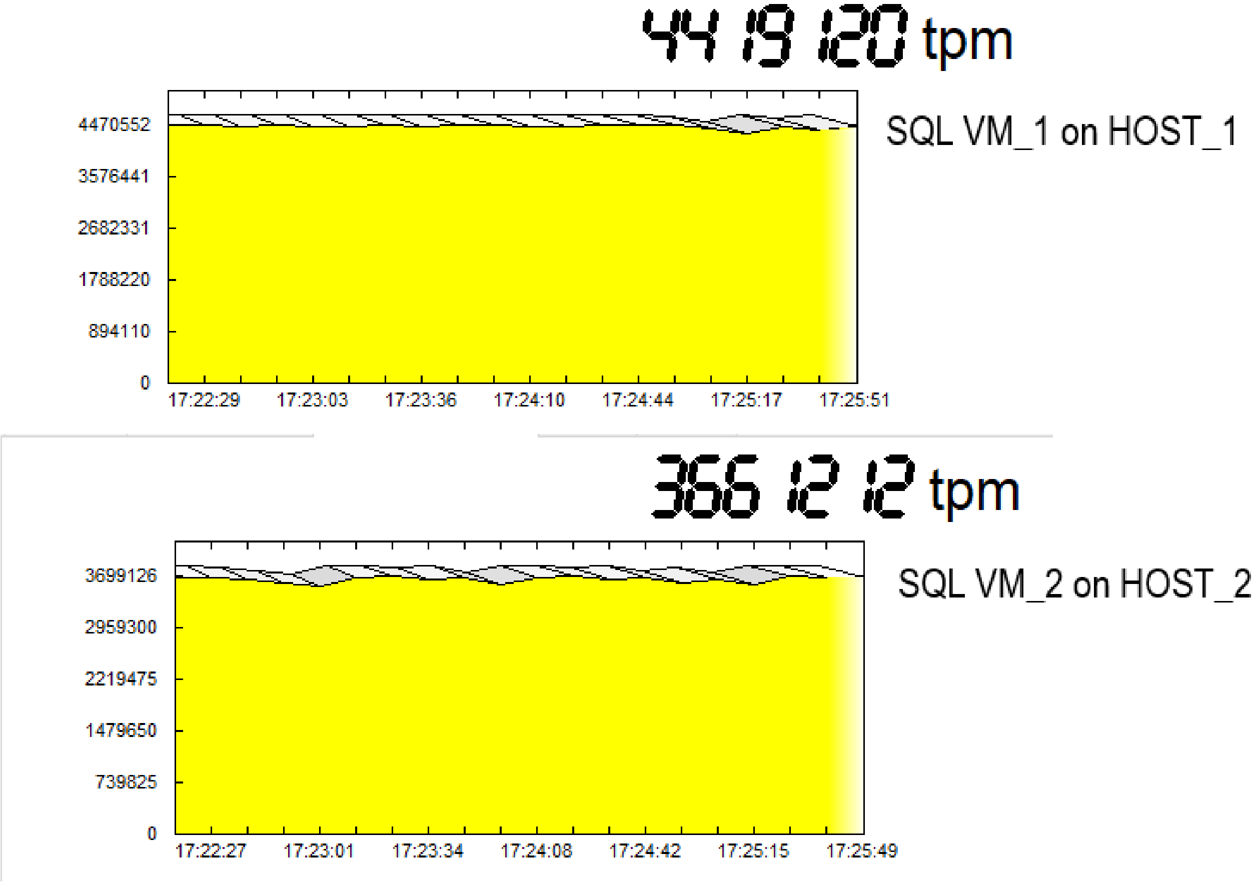


Figure 34: TPM profile obtained during 28 vCPUs tests on two VMs which are situated on same host

Table 24 shows the transactions per minute and new operations per minute results of the HammerDB test using 2 VMs each with 28vCPUs.

Table 24: Microsoft SQL Server results with cluster of four HX752x appliances and 28 vCPUs

Node	VM Name	HammerDB TPM	HammerDB NOPM
Node 1	SQL Instance - 1	3,928,848	854,444
	SQL Instance - 2	3,260,314	708,869
Node 2	SQL Instance - 3	3,863,147	839,716
	SQL Instance - 4	3,091,193	672,483
Node 3	SQL Instance - 5	3,807,745	828,074
	SQL Instance - 6	3,222,624	701,694
Node 4	SQL Instance - 7	3,850,576	837,641
	SQL Instance - 8	3,377,472	734,849
	Total	28,401,919	6,177,770

The 4 node cluster can execute over 28 million transactions per minute and over 6 million new orders per minute. The CPU utilization is 95-99%.

6.5.3 Testing with 44 vCPUs

The third scenario uses 44 vCPUs per VM from the available 56 cores (28 per socket). When hyper-threading is activated, one socket is using 44 vCPUs for the SQL Server VM. The second socket is using 56 vCPUs (44 for the SQL Server VM and 12 for the CVM). However, in this case the threading is not keeping up with transactions.

Figure 35 shows a significant decrease in transaction per minute compared with 28 vCPUs per VM (see Figure 34). The profile curve is also not very flat and consistent in the steady state condition.

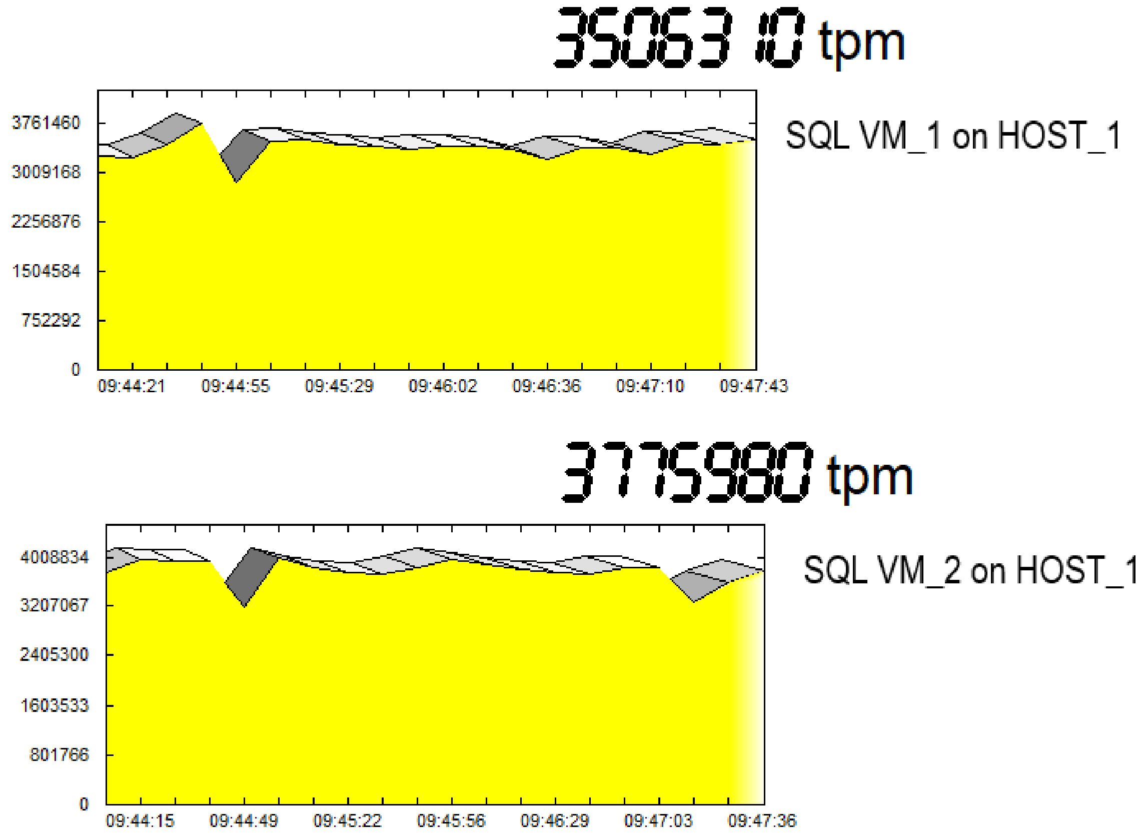


Figure 35: TPM profile obtained during 44 vCPUs tests on two VMs which are situated on same host

6.5.4 Conclusions

The best choice of processor partly depends on which SQL license is purchased. The SQL SERVER/CAL license limits the number of cores to that available from the license. The CPU license does not have this restriction. In either case Lenovo recommends adding 12 more cores per socket for the CVM and ESXi.

Lenovo recommends tying the number of vCPUs to the processor core count. When the number of vCPUs exceeds the core count, hyper-threading is used. Depending on the workload, hyper-threading can increase system performance by up to 30%.However in some cases and when the number of vCPUs is too large, it can increase contention and thereby cause a drop in overall performance.

7 Red Hat OpenShift Container Platform

The Red Hat OpenShift Container Platform is a complete container application platform that provides all aspects of the application development process in one consistent solution across multiple infrastructure footprints. OpenShift integrates all of the architecture, processes, platforms, and services needed to help development and operations teams traverse traditional siloed structures and produce applications that help businesses succeed.

7.1 Solution Overview

Figure 36 below shows the high level architecture of the Red Hat OpenShift Container Platform and the core building blocks based on Lenovo ThinkAgile HX cluster. OpenShift is a platform designed to orchestrate containerized workloads across a cluster of nodes. The system uses the Kubernetes as the core container orchestration engine, which manages the Docker container images and their lifecycle. Lenovo ThinkAgile HX cluster provides NFS storage and VM running environment. CVM virtualizes all the local storage attached to each node in a cluster and presents it as centralized storage array using Nutanix Distributed File System (NDFS)

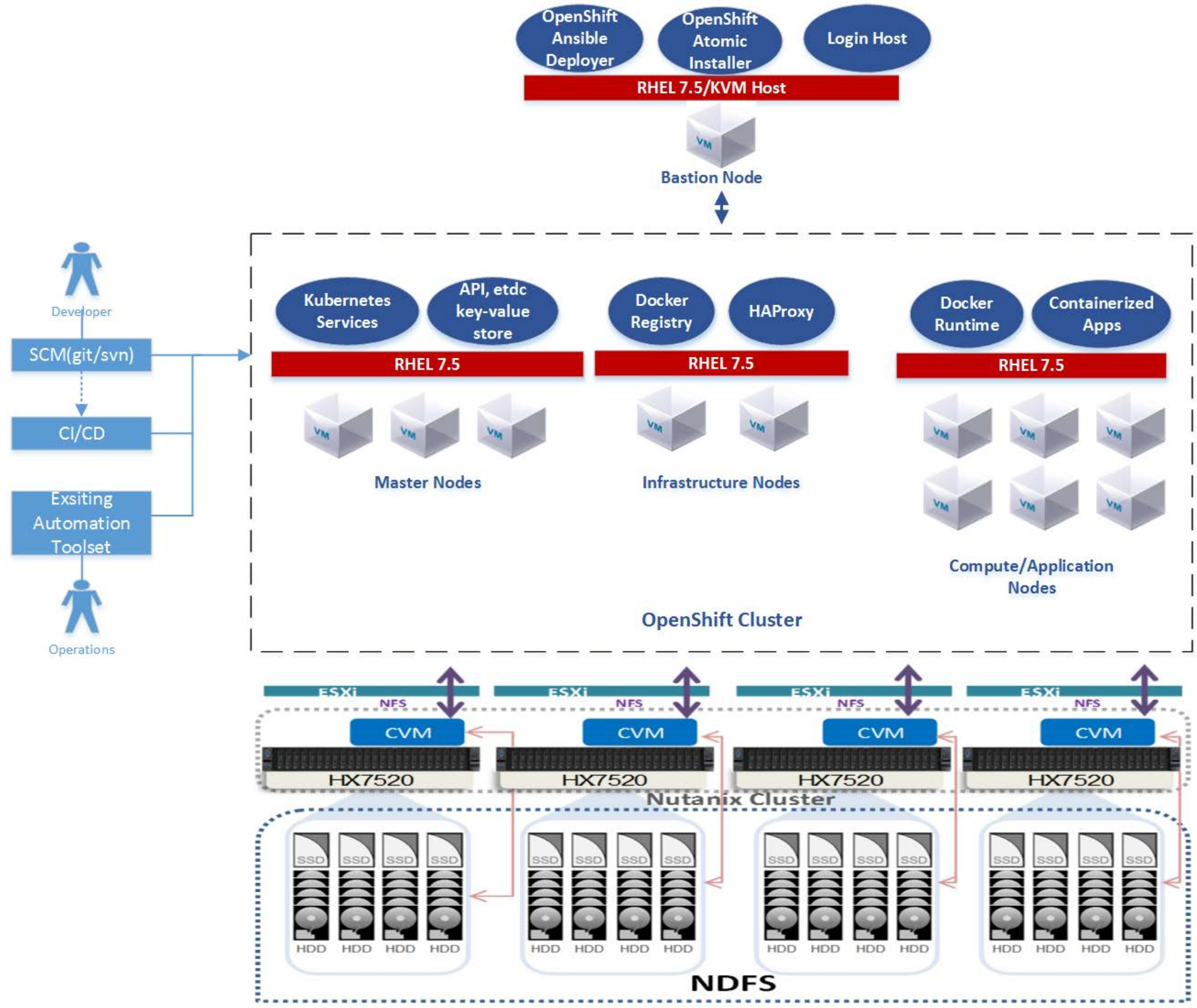


Figure 36. Red Hat OpenShift Container Platform Architecture

The VM node configuration of the OpenShift platform is based on the Kubernetes cluster architecture. The master node is the primary node on which the Kubernetes scheduler, along with the distributed cluster data store (etcd), the REST API services, and other associated management services run. In a product environment, you need to ensure high availability of the master services through replicating the services to multiple VM nodes and implementing monitoring and load-balancing services such as Keepalived and HAproxy. The infrastructure nodes can be used in a product setting to implement such services.

Application nodes (or just shown as Node in the diagram) run the users containerized applications on top of the Docker container environment.

7.2 Component Model

As shown in Figure 37, this chapter describes the components and logical architecture of the Red Hat OpenShift solution.

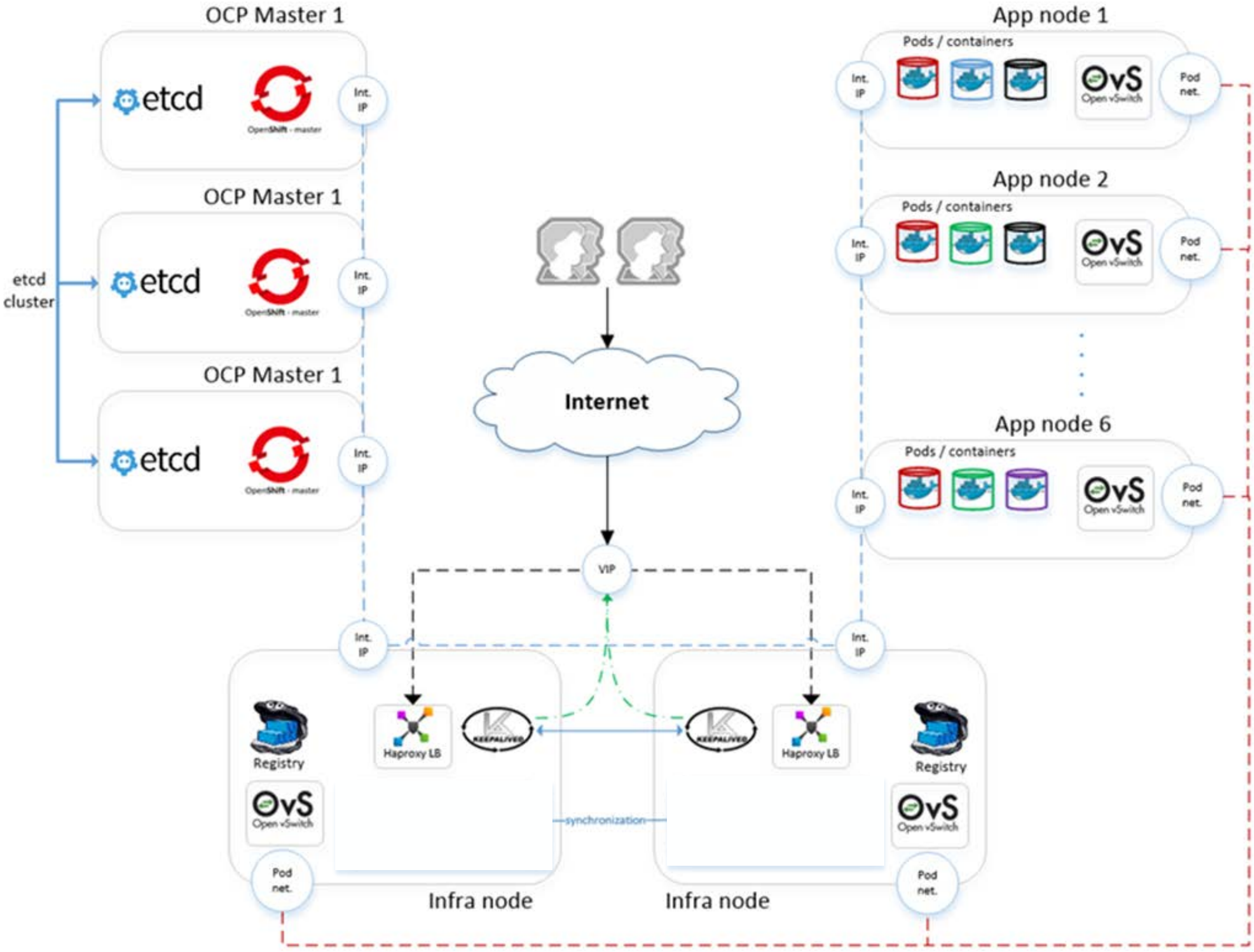


Figure 37. Red Hat OpenShift Container Platform logical architecture

All the OpenShift nodes are connected via the internal network, where they can communicate with each other. Furthermore, Open vSwitch creates its own network for OpenShift pod-to-pod communication. Because of the multi-tenant plugin, Open vSwitch pods can communicate to each other only if they share the same project namespace. There is a virtual IP address managed by Keepalived on two *infrastructure* hosts for external access to the OpenShift web console and applications. Lastly, there is a Red Hat OpenShift Container

Storage server that shares disk space with Docker Registry for Docker image storage. This storage is backed up by Red Hat OpenShift Container Storage, so Docker Registry storage can be easily switched in case of a node failure.

7.2.1 OpenShift infrastructure components

Figure 38 shows the four types of OpenShift nodes: *bastion*, *master*, *infrastructure,* and *application* in ThinkAgile HX cluster. In this solution, OpenShift platform use NFS storage in ThinkAgile HX cluster to store images and deploy applications. A lightweight DNS server, dnsmasq, is running in OpenShift platform. User can also use Active Directory and DNS server in ThinkAgile HX cluster to provide AD and DNS services when a more stable and large-scale OpenShift is required.

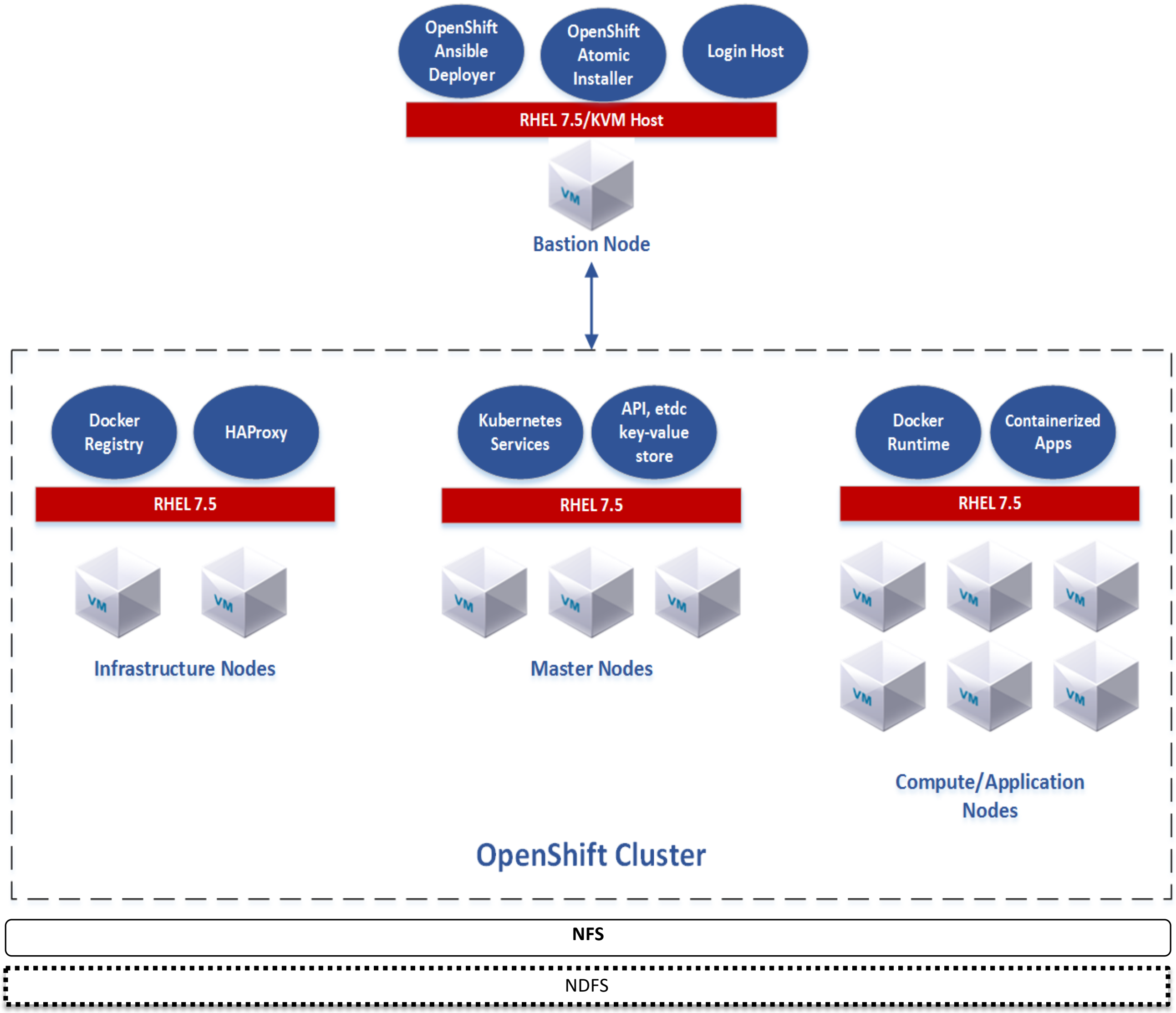


Figure 38. OpenShift Nodes

Bastion node

This is a dedicated node that serves as the main deployment and management server for the OpenShift cluster. This is used as the logon node for the cluster administrators to perform the system deployment and management operations, such as running the Ansible OpenShift deployment playbooks. The *Bastion* node runs RHEL 7.5 Server with the Linux KVM packages installed.

Master node

The OpenShift Container Platform *master* is a server that performs control functions for the whole cluster environment. It is responsible for the creation, scheduling, and management of all objects specific to OpenShift. It includes API, controller manager, and scheduler capabilities in one OpenShift binary. It is also a common practice to install an etcd key-value store on OpenShift *masters* to achieve a low-latency link between etcd and OpenShift *masters*. It is recommended that you run both OpenShift *masters* and etcd in highly available environments. This can be achieved by running multiple OpenShift *masters* in conjunction with an external active-passive load balancer and the clustering functions of etcd. The OpenShift *master* node runs RHEL 7.5 Server.

Infrastructure node

The OpenShift *infrastructure* node runs infrastructure-specific services such as the Docker Registry and the HAProxy router. The Docker Registry stores application images in the form of containers. The HAProxy router provides routing functions for OpenShift applications. It currently supports HTTP(S) traffic and TLS-enabled traffic via Server Name Indication (SNI). Additional applications and services can be deployed on OpenShift *infrastructure* nodes. The OpenShift *infrastructure* node runs RHEL Server 7.5.

Application node

The OpenShift *application* nodes run containerized applications created and deployed by developers. An OpenShift *application* node contains the OpenShift node components combined into a single binary, which can be used by OpenShift *masters* to schedule and control containers. An OpenShift *application* node runs RHEL Server 7.5.

7.2.2 OpenShift architecture

Kubernetes is an open source project developed by Google. The project gained popularity via its open and flexible architecture for managing containerized workloads at large scale. It provides APIs that can be easily integrated into other tools such as the Red Hat OpenShift Container platform. Kubernetes provides the orchestration capabilities for containers, including scheduling the container images to nodes in a cluster, managing the container life cycle, availability, replication, persistent and non-persistent storage for containers, policy, multi-tenancy, network virtualization, routing, hierarchical clusters via federation APIs, and so forth.

A detailed software description of the Kubernetes components is described on this website: docs.openshift.com/enterprise/3.0/architecture/infrastructure_components/kubernetes_infrastructure.html.

Figure 39 shows the OpenShift high-level architecture and components.

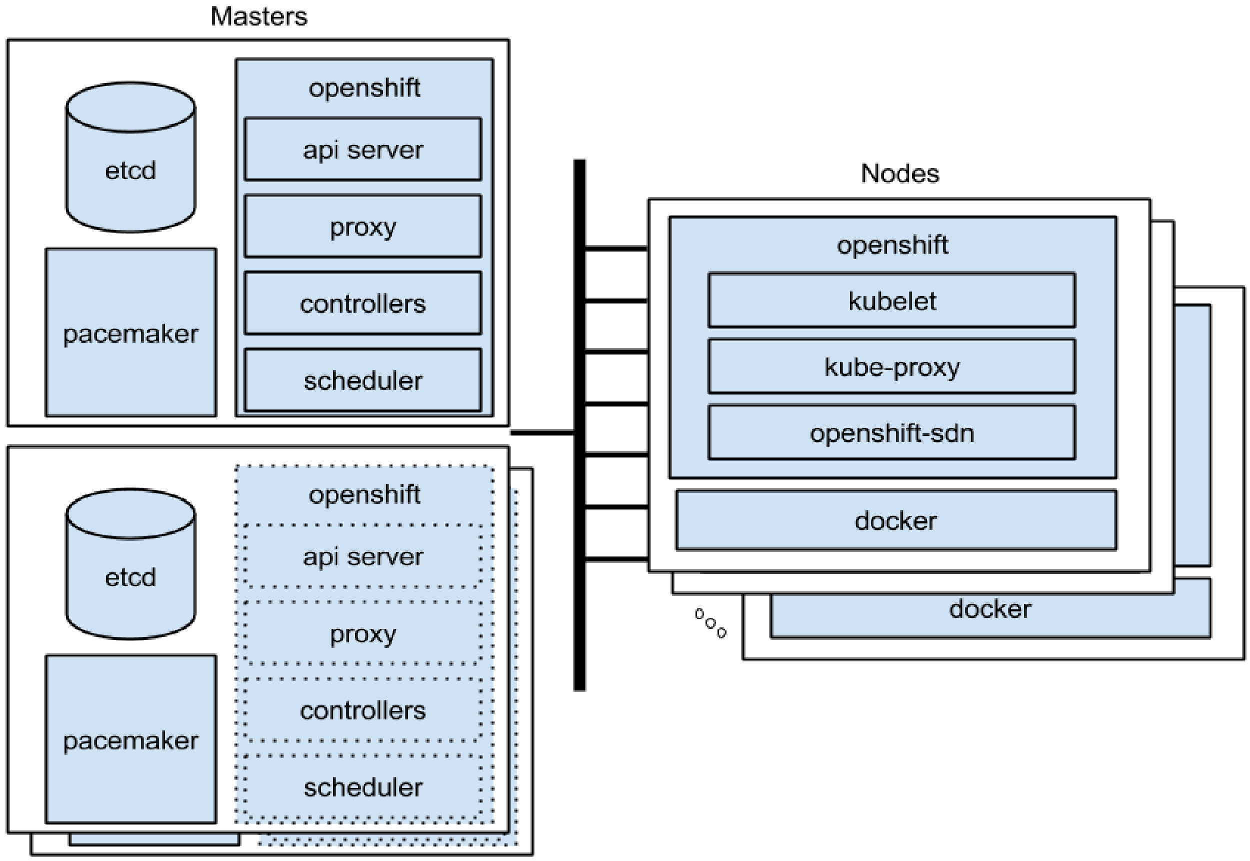


Figure 39. OpenShift component architecture

The master nodes, as described previously, are responsible for core services such as API interface, authentication/authorization, container scheduling, controller management, and configuration database. The master manages the state of the cluster and the lifecycle of the user container images. For redundancy and high availability, you can have multiple master nodes with frontend load-balancers such as HAproxy. The command line interface to the master nodes is implemented via the “oc” command.

The (worker) nodes are where users’ container images are executed. In OpenShift terminology the worker nodes run “pods”, each of which manages one or more running containers. Each node implements a “kubelet”, which is the node level controller that manages the pods and interacts with the OpenShift master.

In addition to the core OpenShift services, the Red Hat OpenShift platform also includes other features such as the Web based user self-service console, monitoring, an integrated container registry, storage management, authentication/authorization, automation via built-in Ansible playbooks, and other administrative tools for managing the container platform.

7.3 Operational model

The OpenShift Container Platform can be implemented in development/test, staging, and production settings. Each node role has its own dedicated servers or VMs for performance and availability. However, in a non-production environment, a minimal environment can be provided to test applications before moving them to a staging or production environment.

For a test/development environment, you can use minishift or ‘oc cluster up’ to deploy all on one server, or implement OpenShift with five servers for a more formal platform as shown in table below:

Node type	Quantity	Node role
Deployment	1	Deployment of the environment, Ansible playbooks, hardware management, etc.
Master	1	OpenShift API master, Kubernetes scheduler, etcd, other core services
Compute	3	Runs the application containers

For a production OpenShift deployment, all of the core services such as the API servers, Kubernetes scheduler, etcd, etc., need to be highly available. The table below shows the recommended configuration for a production deployment.

Node type	Quantity	Node role
Bastion	1	Deployment of the environment, Ansible playbooks, hardware management, etc.
Infrastructure	2	OpenShift HAProxy, container registry, routing, etcd.
Master	3	OpenShift API master, Kubernetes scheduler
Compute	3+	Runs the application containers
Storage	0	Using NFS shared storage in Nutanix cluster

There are performance and availability implications of running the Red Hat OpenShift Container Storage alongside the workload containers in a hyperconverged environment. For production environments, it is recommended to separate hyper-converged compute servers from storage-only servers, or to ensure that the servers have sufficient CPU, memory, and storage resources to avoid any performance bottlenecks.

7.3.1 Compute servers

The OpenShift Container Platform can be implemented on a small footprint of VMs on Lenovo ThinkAgile HX HX752x cluster. The VMs could be clustered together and scaled as the user workloads grow. Each HX server has the following configuration:

- 2 x Intel Xeon Scalable 8170 (26 cores @ 2.1 GHz) processors
- 768GB RAM
- 4 x 1.92TB SATA SSDs
- 20 x 2TB SATA HDDs
- ESXi 6.5 U1
- Set UEFI to “Performance Bias” – for more information, see lenovopress.com/lp0780

The Nutanix CVM is configured as follows:

- 12 vCPUs
- 96 GB RAM

7.3.2 Persistent storage for containerized workloads

There are two types of storage consumed by containerized applications – ephemeral (non-persistent) and persistent. As the names suggest, non-persistent storage is created and destroyed along with the container and is only used by applications during their lifetime as a container. Hence, non-persistent storage is used for temporary data. When implementing the OpenShift Container Platform, local disk space on the application nodes can be configured and used for the non-persistent storage volumes.

Persistent storage, on the other hand, is used for data that needs to be persisted across container instantiations. An example is a 2 or 3-tier application that has separate containers for the web and business logic tier and the database tier. The web and business logic tier can be scaled out using multiple containers for high availability. The database that is used in the database tier requires persistent storage that is not destroyed.

OpenShift uses a persistent volume framework that operates on two concepts – persistent storage and persistent volume claim. Persistent storage is the physical storage volumes that are created and managed by the OpenShift cluster administrator. When an application container requires persistent storage, it would create a persistent volume claim (PVC). The PVC is a unique pointer/handle to a persistent volume on the physical storage, except that PVC is not bound to a physical volume. When a container makes a PVC request, OpenShift would allocate the physical disk and binds it to the PVC. When the container image is destroyed, the volume bound to the PVC is not destroyed unless you explicitly destroy that volume. In addition, during the lifecycle of the container if it relocates to another physical server in the cluster, the PVC binding will still be maintained. After the container image is destroyed, the PVC is released, but the persisted storage volume is not deleted. The specific persistent storage policy for the volume will determine when the volume gets deleted.

For more detailed conceptual information on persistent volumes see: access.redhat.com/documentation/en-us/openshift_container_platform/3.9/html-single/architecture/#architecture-additional-concepts-storage

A variety of persistent storage options are available for OpenShift, choices including NFS, OpenStack Cinder, Ceph RBD, iSCSI, fiber channel SAN, hyperconverged storage using Red Hat OpenShift Container Storage, AWS elastic block storage (EBS), and others. For a complete list of these choices and the corresponding requirements, see the link below: access.redhat.com/documentation/en-us/openshift_container_platform/3.9/html-single/installation_and_configuration/#configuring-persistent-storage

In the OpenShift platform, persistent storage use NFS storage in the ThinkAgile HX cluster as backend. Lenovo recommends configuration the following persistent storage:

- 500 GB NFS storage for Docker registry
- 5 TB NFS storage for other applications

7.3.3 Networking

For OpenShift Container Platform deployment in Lenovo ThinkAgile HX752x cluster, VMware virtual infrastructure networking is recommended as the choice for all cluster-wide communication for the core OpenShift services, application container workloads, storage services access with NFS, as well as all east-west traffic and north-south traffic across the container workloads.

There are three logical networks:

- **External Network**: The external network is used for the public API, the OpenShift web interface, and exposed applications (services and routes).
- **Internal Network**: This is the primary, non-routable network used for cluster management and inter-node communication. The same network acts as the layer for server provisioning using PXE and HTTP. Domain Name Servers (DNS) and Dynamic Host Configuration Protocol (DHCP) services also reside on this network to provide the functionality necessary for the deployment process and the cluster to work. Communication with the Internet is provided by NAT configured on the *bastion* node.
- **Storage Network**: This is isolated network used for connection between OpenShift cluster and shared NFS storage created in Lenovo ThinkAgile HX752x cluster.

Figure 40 shows the Red Hat OpenShift servers and the recommended network architecture.

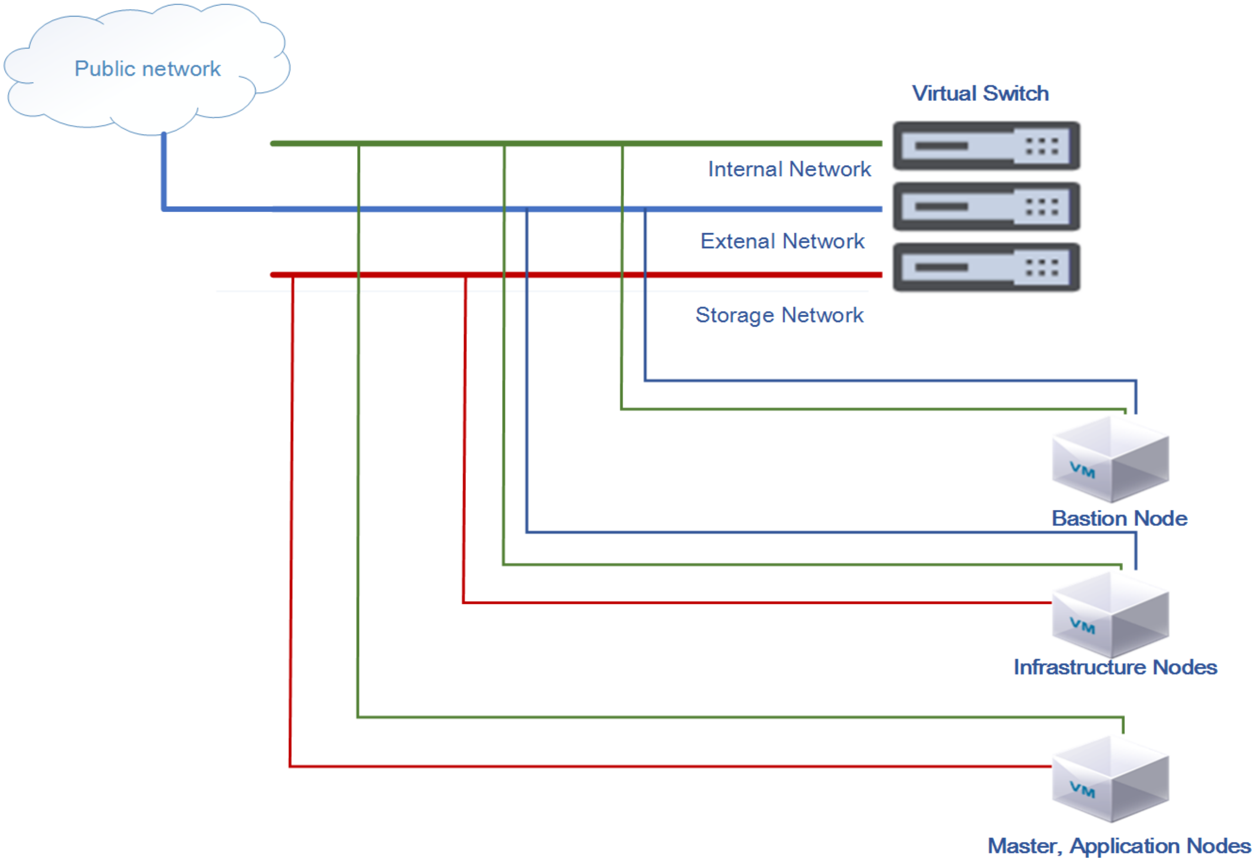


Figure 40. OpenShift Network Connectivity

All OpenShift nodes are connected via the internal network, where they can communicate with each other. Furthermore, Open vSwitch creates its own network for OpenShift pod-to-pod communication. Because of the multi-tenant plugin, Open vSwitch pods can communicate to each other only if they share the same project namespace. There is a virtual IP address managed by Keepalived on two *infrastructure* hosts for external access to the OpenShift web console and applications.

7.4 Best practices

This section describes recommended the best practices for implementing Red Hat OpenShift Container platform in the ThinkAgile HX Nutanix environment.

7.4.1 Shared NFS storage as persistent storage

A Nutanix storage container is a defined subset of available storage within a storage pool. Storage containers allow users to apply rules or transformations such as compression to a data set.

By default all Nutanix storage containers are thin provisioned which reduces unused capacity and automatically provisions additional storage capacity when needed. It is also very easy to add additional storage capacity for databases by simply adding nodes to the cluster. It is also possible to set a storage reservation amount to guarantee a minimum amount of storage capacity.

Data compression can be used to further increase data capacity especially for data that is less frequently accessed. Lenovo recommends disabling compression for the Red Hat OpenShift Container platform when supporting multiple types of applications.

A resiliency factor of 2 is the default. This provides a minimum level of data redundancy but a resiliency factor of 3 might be important in some environments. Using erasure coding saves significant storage capacity but it only recommended for archive data.

Nutanix virtual storage using shared NFS storage should be configured on each OpenShift VM node. Users can create volume manually on shared NFS and configure it in container configuration files. Users can also use third-party tools to utilize shared NFS by PVC (persistent volume claim).

Nutanix shared NFS storage can support most types of applications services. However, for some no-SQL databases such as MongoDB and Elastic Search, it is recommended to use other types of storage such as fiber channel SAN, Red Hat OpenShift Container Storage for the persistent storage backend.

7.4.2 Cluster high availability

The minimum number of nodes in each cluster is 3 and should be at least 4 to provide failover. The following high availability features are recommended for an ESXi-based ThinkAgile HX cluster:

- VMware vSphere high availability (HA) for failover
- VMware vSphere distributed resource scheduler (DRS) for load balancing

VMware vSphere HA pools VMs into a cluster to increase data resiliency. If a host fails, VMware HA moves the VMs to other hosts with spare capacity. Lenovo recommends enabling the “Admission Control Setting” and using the “Admission Control Policy” to set the percentage of cluster resources reserved as failover spare capacity.

VMware vSphere DRS can be used to group ESXi hosts into resource clusters to provide highly available resources and balance workloads. In order to keep the active working set for each OpenShift application VM to the ThinkAgile HX node, Lenovo recommends creating a host group for each node and a “should” rule that keeps each OpenShift application VM on a 1 to 1 ratio with the ESXi host. The hosts should be configured with a minimum of N+1 availability. Other VM nodes (bastion node, master node, infrastructure node) are recommended to run on non-DRS hosts to allow OpenShift to manage the load balancing for these nodes.

7.5 Deployment example

This deployment example was fully tested and verified by Lenovo. All of the deployment files are available from GitHub at the following location: https://github.com/lenovo/deployment-scripts-for-OpenShift-Platform/tree/master/OCP_HX.

The example uses 10 nodes as follows:

- 1 Bastion node
- 3 Master nodes
- 2 Infrastructure nodes
- 4 Application nodes with Red Hat OpenShift Container Storage to provide hyper-converged compute

This configuration represents a production grade OpenShift implementation that meets high-availability, redundancy, and scale requirements for enterprises. Additional Application nodes can be added to increase the available compute and storage capacity.

Table 25 provides a VM configuration summary for this example deployment on Lenovo ThinkAgile HX752x cluster.

Table 25. VM Configuration for each OpenShift Node

Node Type	CPU	Memory	Hard Disk	Network Adapter
Bastion	2 vCPU(s)	20 GB	100 GB	2 VMXNET3 NIC
Master	4 vCPU(s)	40 GB	200 GB	2 VMXNET3 NIC
Infrastructure	4 vCPU(s)	100 GB	300 GB	3 VMXNET3 NIC
Application	4 vCPU(s)	100 GB	300 GB	2 VMXNET3 NIC

7.5.1 Ansible inventory file

In order to install the OpenShift Container Platform cluster, an Ansible inventory file has to be created with the environment's description. The rest of this section explains the inventory file and provides the additional variables that are used for the automatic prerequisites and a Keepalived deployment.

The Ansible section below specifies the types of nodes that are used in an OpenShift Container Platform environment. Required groups are nodes, masters, and etcd. Optional groups are lb (for load balancing in multi-master clusters), local (which specifies the *Bastion* node).

```
[OSEv3:children]
masters
nodes
etcd
lb
local
```

The Ansible section below describes global cluster parameters. Parameter openshift_master_cluster_method specifies the load balancing method in a multi-master environment. With the native value, there will be a separated HAProxy load balancer installed on the specified host and configured for the whole environment.

The hostname for users and cluster components to access the cluster load balancer from external and internal networks is set in the **openshift_master_cluster_hostname** and **openshift_master_cluster_public_hostname** parameters. The parameter **openshift_master_identity_providers** configures the way for authentication of OpenShift users. In this example, this parameter is based on htpasswd files stored in the OpenShift configuration directory. However, many other authentication methods can be used such as LDAP, Keystone*, or GitHub* accounts. **os_sdn_network_plugin_name** specifies the SDN Open vSwitch plugin used in environment. For this example, **redhat/openshift-ovs-multitenant** provides isolation between OpenShift projects on the network level. In this solution, Docker Registry uses Nutanix shared NFS storage for Docker image storage.

```
[OSEv3:vars]
ansible_ssh_user=root
ansible_become=true
containerized=true
openshift_master_cluster_method=native
openshift_master_cluster_hostname=lb.ocp.example.local
openshift_master_cluster_public_hostname=openshift.ocp.example.com
openshift_master_default_subdomain=apps.ocp.example.com
openshift_master_cluster_ip=172.29.7.30
openshift_master_cluster_public_ip=10.240.39.98
openshift_master_portal_net=10.0.0.0/16
openshift_deployment_type=openshift-enterprise
openshift_release=v3.9
os_sdn_network_plugin_name='redhat/openshift-ovs-multitenant'

# Allow all auth
openshift_master_identity_providers=[{'name': 'allow_all', 'login': 'true',
'challenge': 'true', 'kind': 'AllowAllPasswordIdentityProvider'}]

rhel_subscription_user=??????
rhel_subscription_pass=******

openshift_enable_unsupported_configurations=True
openshift_hosted_registry_storage_kind=nfs
openshift_hosted_registry_storage_access_modes=['ReadWriteMany']
openshift_hosted_registry_storage_host=192.168.5.2
openshift_hosted_registry_storage_nfs_directory=/RedHat
openshift_hosted_registry_storage_volume_name=registry
openshift_hosted_registry_storage_volume_size=200Gi

external_interface=ens192
external_netmask=255.255.252.0
external_gateway=10.240.36.1
external_dns=10.240.0.10
```

```
internal_netmask=255.255.0.0
bastion_ip=172.29.7.1
internal_gateway={{ bastion_ip }}
internal_interface=ens224
local_dns={{ bastion_ip }}
dhcp_first_ip=172.29.7.100
dhcp_last_ip=172.29.7.150
root_password=Passw0rd

openshift_install_examples=true

openshift_master_api_port=8443
openshift_master_console_port=8443

openshift_disable_check=docker_image_availability
```

The Ansible section below describes which servers act as OpenShift masters. For this example, three OpenShift masters are implemented for control plane HA purposes. OpenShift master components can be installed with two methods: rpm-based or container-based. In this RA, all OpenShift components are implemented as containers, which is determined by the **`containerized=True`** parameter.

```
[masters]
master1.ocp.example.local containerized=true openshift_ip=172.29.7.2
openshift_hostname=master1.ocp.example.local
master2.ocp.example.local containerized=true openshift_ip=172.29.7.3
openshift_hostname=master2.ocp.example.local
master3.ocp.example.local containerized=true openshift_ip=172.29.7.4
openshift_hostname=master3.ocp.example.local
```

The Ansible section below describes which servers act as OpenShift nodes. For this example, seven OpenShift nodes are implemented. Two of them perform infrastructure functions, which is determined by the openshift_node_labels="{'region': 'infra'}" parameter. OpenShift node components are also installed on OpenShift master servers. User application could be deployed on these servers when configuring parameter openshift_schedulable=true. In this RA, all node components are implemented as containers, which is determined by the **containerized=true** parameter.

```
[nodes]
master1.ocp.example.local  containerized=true openshift_ip=172.29.7.2
openshift_hostname=master1.ocp.example.local
master2.ocp.example.local  containerized=true openshift_ip=172.29.7.3
openshift_hostname=master2.ocp.example.local
master3.ocp.example.local  containerized=true openshift_ip=172.29.7.4
openshift_hostname=master3.ocp.example.local
infra2.ocp.example.local openshift_node_labels="{'region': 'infra'}"
containerized=true  openshift_ip=172.29.7.6
openshift_hostname=infra2.ocp.example.local
```

```
infra1.ocp.example.local openshift_node_labels="{'region': 'infra'}"
containerized=true  openshift_ip=172.29.7.5
openshift_hostname=infra1.ocp.example.local
app1.ocp.example.local openshift_schedulable=true containerized=true
openshift_ip=172.29.7.7 openshift_hostname=app1.ocp.example.local
openshift_hostname_check=false
app2.ocp.example.local openshift_schedulable=true containerized=true
openshift_ip=172.29.7.8 openshift_hostname=app2.ocp.example.local
app3.ocp.example.local openshift_schedulable=true containerized=true
openshift_ip=172.29.7.9 openshift_hostname=app3.ocp.example.local
app4.ocp.example.local openshift_schedulable=true containerized=true
openshift_ip=172.29.7.10 openshift_hostname=app4.ocp.example.local
```

The Ansible section below describes hosts that will run etcd instances. For this example, three *etcd* instances are installed on three *master* servers to achieve low-latency traffic between them. When many etcd instances are specified in an inventory file, they are automatically clustered in order to provide a highly available key-value etcd store. An etcd cluster that consists of three etcd instances resists a failure of one etcd instance. It is also recommended to have an odd number of etcd instances in a cluster.

```
[etcd]
master1.ocp.example.local containerized=true openshift_ip=172.29.7.2
openshift_hostname=master1.ocp.example.local
master2.ocp.example.local containerized=true openshift_ip=172.29.7.3
openshift_hostname=master2.ocp.example.local
master3.ocp.example.local containerized=true openshift_ip=172.29.7.4
openshift_hostname=master3.ocp.example.local
```

When **openshift_master_cluster_method** is set to **native**, then the Ansible section below specifies a host on which HAProxy load balancer will be installed and configured. For this example, two HAProxy load balancers are installed on two infrastructure servers. They use one common virtual IP address that is managed by Keepalived software to achieve a highly available OpenShift Container Platform cluster.

```
[lb]
infra1.ocp.example.local openshift_hostname=infra1.ocp.example.local
openshift_hostname_check=false openshift_ip=172.29.7.5
openshift_public_ip=10.240.39.91
infra2.ocp.example.local openshift_hostname=infra2.ocp.example.local
openshift_hostname_check=false openshift_ip=172.29.7.6
openshift_public_ip=10.240.39.92
```

7.5.2 Software

For this example, the following software is needed:

- **OpenShift Container Platform**, which adds developer- and operation-centric tools to enable rapid application development, easy deployment, scaling, and long-term lifecycle maintenance for small and large teams and applications

In addition, the OpenShift Container Platform requires the following software packages:

- **Docker** to build, ship, and run containerized applications
- **Kubernetes** to orchestrate and manage containerized applications
- **Etcd***, which is a key-value store for the OpenShift Container Platform cluster
- **Open vSwitch*** to provide software-defined networking (SDN)-specific functions in the OpenShift Container Platform environment
- **Ansible®** for installation and management of the OpenShift Container Platform deployment
- **HAProxy*** for routing and load-balancing purposes
- **Keepalived*** for virtual IP management for HAProxy instances

Table 26 lists the software versions used for this example deployment

Table 26. Software versions

Component	**Version**
Red Hat Enterprise Linux	7.5
OpenShift Container Platform	3.9
Docker	1.13.1
Ansible	2.4.6.0
rhel7/etcd	latest
openshift3/openvswitch	v3.9.43
openshift3/ose	v3.9.43
openshift3/ose-pod	v3.9.43
openshift3/node	v3.9.43
openshift3/ose-haproxy-router	v3.9.43
openshift3/keepalived	1.0_ra

Each VM node is installed with RHEL 7.5 as the base operating system and the following resources:

- Minimum 40 GB hard disk space for the file system containing ***`/var/`***.
- Minimum 1 GB hard disk space for the file system containing ***`/usr/local/bin/`***.
- Minimum 1 GB hard disk space for the file system containing the system's temporary directory

7.5.3 Networking

The Lenovo ThinkAgile HX cluster provides virtual networks for user VMs. Table 27 shows the network configuration used in OpenShift cluster (as shown in Figure 40).

Table 27. Network configuration for OpenShift Cluster

Node Type	**External Network**	**Internal Network**	**Storage Network**
Bastion	VMXNET 3	VMXNET 3	
Master		VMXNET 3	VMXNET 3
Infrastructure	VMXNET 3	VMXNET 3	VMXNET 3
Application		VMXNET 3	VMXNET 3

7.5.4 Automatic prerequisites installation

After the operating system has been installed and configured, the nodes need to be prepared for OpenShift installation. Perform the following preliminary steps: prepare an *openshift* account and exchange SSH keys across all nodes, attach software licenses, install and configure the DNS service, install additional packages, and configure Docker Engine.

All the remaining tasks can be executed automatically using the Ansible playbooks available at https://github.com/lenovo/deployment-scripts-for-OpenShift-Platform/tree/master/OCP_HX.

Based on information from the operating system deployment, prepare an Ansible inventory file and place it in the location /etc/ansible/hosts on the *Bastion* node. The hosts file for this deployment example is named hosts.example and is provided in the GitHub repository.

After that, clone the git repository onto the *Bastion* node:

```
$ git clone https://github.com/lenovo/deployment-scripts-for-OpenShift-Platform.git
$ cd deployment-scripts-for-OpenShift-Platform/OCP_HX/src/prerequisites
```

In the inventory file, set up the following additional variables:

rhel_subscription_user:	Name of the user who will be used for registration
rhel_subscription_pass:	Password of the user who will be used for registration
ansible_ssh_user:	Insert root or other user with root privileges
ansible_become:	Set to True to run commands with sudo privileges
local_dns:	Type a proper IP address for your bastion node that runs the DNS service

Finally start the Ansible playbook by entering the following command:

```
$ ansible-playbook nodes_setup.yaml -k
```

7.5.5 Automatic Keepalived deployment

The OpenShift Container Platform delivers two flavors of HAProxy load balancing software. The first flavor, which is spawned as a daemon, distributes API calls between *master* servers. The second flavor, spawned as a Docker container, provides the *router* mechanism for exposing applications inside a cluster.

To achieve high availability (HA), maximum fault tolerance, and performance, this deployment example includes an additional package called Keepalived. It is open-source software distributed under the GPL license and is recognized by Red Hat as their recommended solution. Please see the following web site for more information: Red Hat Enterprise Linux documentation.

This example deployment uses both flavors of HAProxy instances, which are installed on both Infrastructure nodes. A single point of failure is eliminated is eliminated when used in conjunction with floating IP addresses provided by Keepalived,

Installation and configuration of this HA Solution can be performed manually or through a single command using an Ansible playbook. First the following variables must be defined in the Ansible inventory:

```
external_interface=enp6s0f0
external_netmask=255.255.250.128
external_gateway=10.240.202.129
```

```
external_vlan= 302
external_dns=114.114.114.114
internal_interface=bond0
internal_netmask=255.255.0.0
openshift_master_cluster_ip=172.30.4.30
openshift_master_cluster_public_ip=10.240.202.188
```

To deploy the Keepalived daemons using an Ansible playbook on Infrastructure nodes, enter following command inside the cloned GitHub repository:

```
$ su openshift
$ ansible-playbook \
    openshift-container-architecture/OCP_HX/src/keepalived-multimaster/keepalived.yaml
```

7.5.6 OpenShift container platform installation

When the inventory file with the environment description is prepared and all prerequisites are configured, the OpenShift Container Platform install can be performed from the *Bastion* node.

For this deployment example, a containerized version of the OpenShift Container Platform is installed on servers. This installer image provides the same functionality as the RPM-based installer, but it runs in a containerized environment that provides all of its dependencies rather than being installed directly on the node. The only requirement to use it is the ability to run a container, and atomic packages are installed. The installer image can be used as a system container. System containers are stored and run outside of the traditional Docker service. This enables running the installer image from one of the target hosts without concern for the install restarting Docker on the host.

This install process is straightforward, and requires three steps:

First, Create shared storage:

```
Create a NFS storage container (named Redhat and granted ReadWriteMany access modes)
in Nutanix cluster. And Create a directory (named registry and granted ReadWriteMany
access modes) in storage container.
```

Second, execute the prerequisites:

```
$ atomic install --system --set INVENTORY_FILE=/etc/ansible/hosts --storage=ostree
--set PLAYBOOK_FILE=/usr/share/ansible/openshift-ansible/playbooks/prerequisites.yml
--set OPTS="-v" registry.access.redhat.com/openshift3/ose-ansible:v3.9
```

Third, deploy OpenShift platform：

```
$ atomic install –-system --storage=ostree  --set INVENTORY_FILE=/etc/ansible/hosts
--set PLAYBOOK_FILE=/usr/share/ansible/openshift-ansible/playbooks/deploy_cluster.yml
--set OPTS="-v" registry.access.redhat.com/openshift3/ose-ansible:v3.9
```

After the installation process, the Ansible playbook should report no errors and the OpenShift Container Platform environment is set up. If needed, you can easily uninstall the environment with the following command:

```
$ ansible-playbook -i /etc/ansible/hosts /usr/share/ansible/openshift-
ansible/playbooks/adhoc/uninstall.yml
```

Next user credentials should be created using the following commands:

```
$ sudo yum install httpd-tools
$ touch users.htpasswd
$ htpasswd -n <user_name> >> users.htpasswd
```

Use the **htpasswd** command for each user account and propagate the users.htpasswd file to every OpenShift *master* node, into the **/etc/origin/master/** directory.

Lastly restart the API services on each *master* node using the command:

```
$ sudo systemctl restart atomic-openshift-master-api
```

7.5.7 Deployment validation

The deployment should be validated before it is used.

First, log on to one of the OpenShift *master* nodes and check that all nodes are connected to the cluster using the commands:

```
$ ssh master1.ocp.example.local
$ oc get nodes
```

Here is some example output from the command:

```
NAME                        STATUS  ROLES      AGE     VERSION
app1.ocp.example.local      Ready   compute    26d     v1.9.1+a0ce1bc657
app2.ocp.example.local      Ready   compute    26d     v1.9.1+a0ce1bc657
app3.ocp.example.local      Ready   compute    26d     v1.9.1+a0ce1bc657
app4.ocp.example.local      Ready   compute    26d     v1.9.1+a0ce1bc657
infra1.ocp.example.local    Ready   <none>     26d     v1.9.1+a0ce1bc657
infra2.ocp.example.local    Ready   <none>     26d     v1.9.1+a0ce1bc657
master1.ocp.example.local   Ready   master     26d     v1.9.1+a0ce1bc657
master2.ocp.example.local   Ready   master     26d     v1.9.1+a0ce1bc657
master3.ocp.example.local   Ready   master     26d     v1.9.1+a0ce1bc657
```

All cluster nodes should be listed and marked as **Ready**. If any node is in a **NotReady** state then it is not properly assigned to a cluster and should be inspected using the following command to verify the etcd state:

```
$ sudo etcdctl -C https://etcd1.ocp.example.local:2379 --ca-file=/etc/etcd/ca.crt --
cert-file=/etc/etcd/peer.crt --key-file=/etc/etcd/peer.key cluster-health
```

Here is some example output from the command:

```
member 5f0aab880290ddeb is healthy: got healthy result from
https://etcd1.ocp.example.local:2379
member c305190f3c57613c is healthy: got healthy result from
https://etcd2.ocp.example.local:2379
```

```
member c434590bbf158f3d is healthy: got healthy result from
https://etcd3.ocp.example.local:2379
```

All etcd members should be listed and marked as **healthy**. If any etcd member is in an **unhealthy** state then it is not properly assigned to an etcd cluster.

To further inspect and verify all the components of the OpenShift Container Platform cluster, use the command:

```
$ oc get all
```

All the listed items should have a **Running** status.

At a final verification step, log on to the OpenShift Container Platform web console using the following URL address: https://openshift.ocp.example.com:8443 and display the OpenShift container catalog. Figure 41 shows an example.

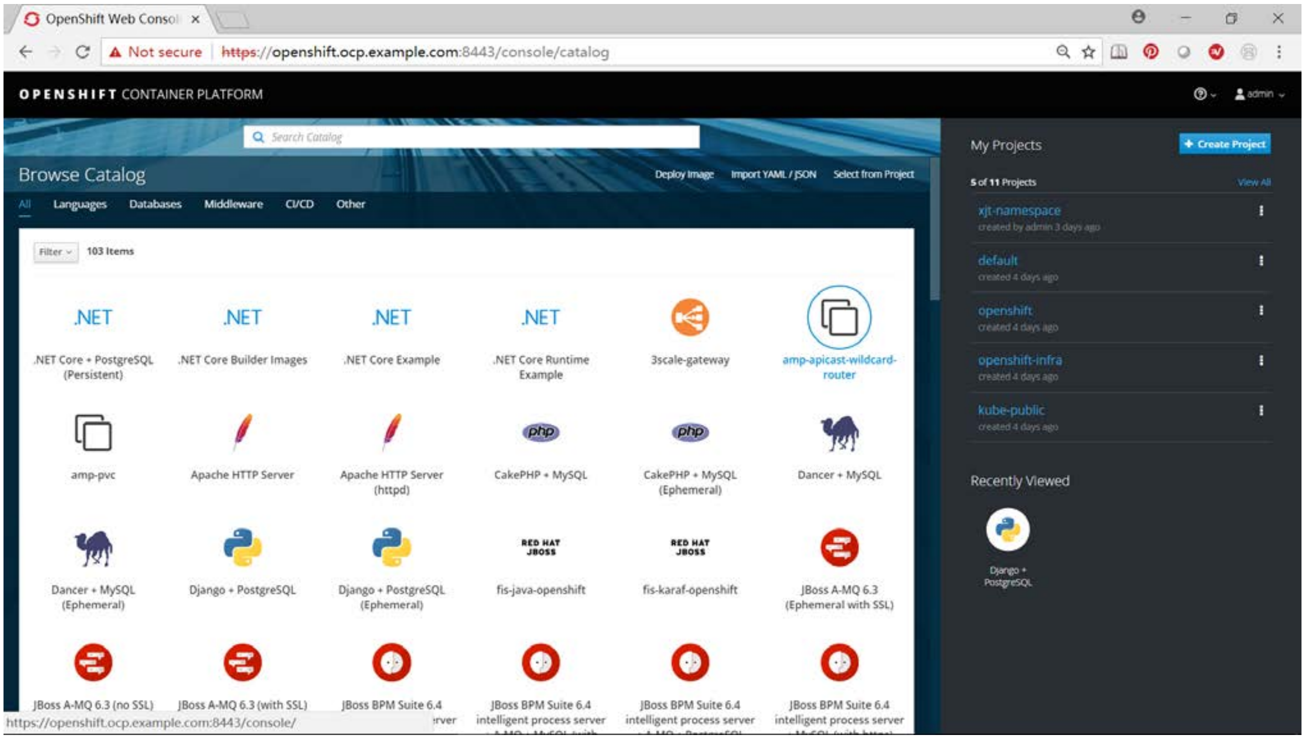
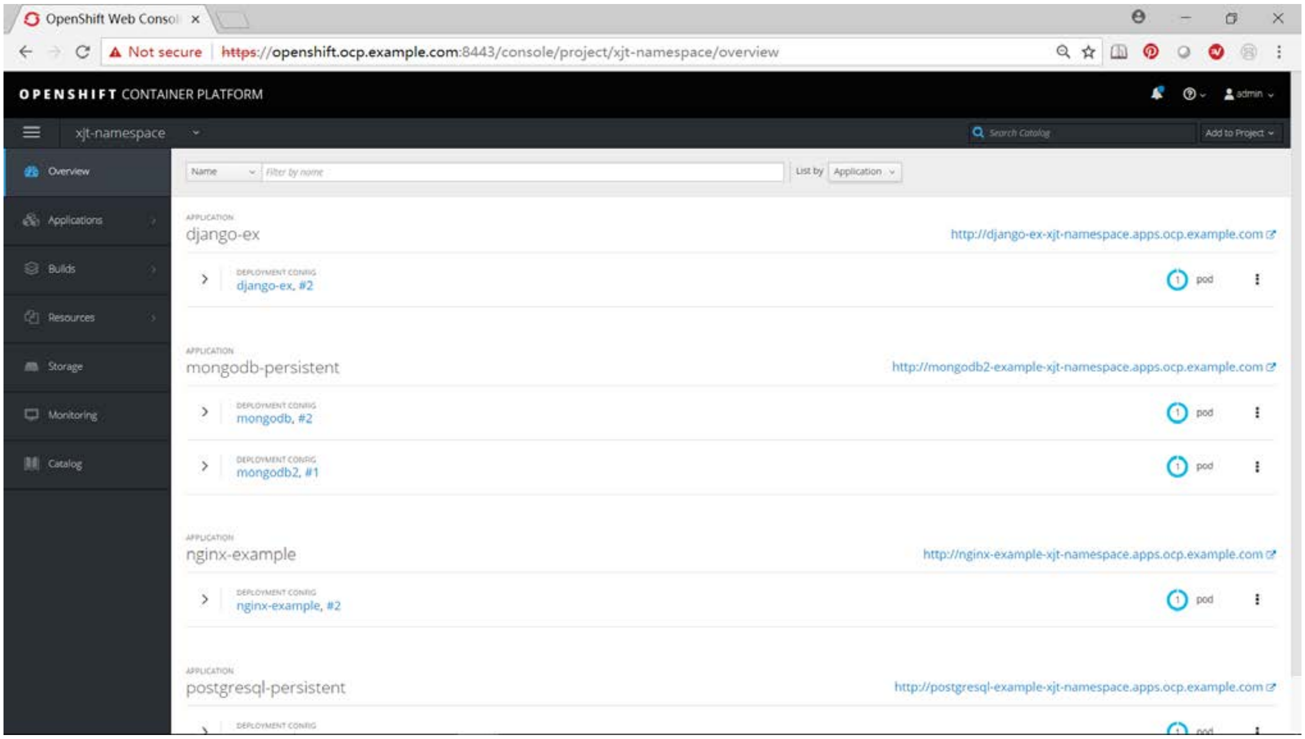


Figure 41 Example Catalog for OpenShift Container Platform

Figure 42 shows other OpenShift application components.



APPLICATION
django-ex
http://django-ex-xjt-namespace.apps.ocp.example.com

DEPLOYMENT CONFIG
django-ex, #2
1 pod

APPLICATION
mongodb-persistent
http://mongodb2-example-xjt-namespace.apps.ocp.example.com

DEPLOYMENT CONFIG
mongodb, #2
1 pod

DEPLOYMENT CONFIG
mongodb2, #1
1 pod

APPLICATION
nginx-example
http://nginx-example-xjt-namespace.apps.ocp.example.com

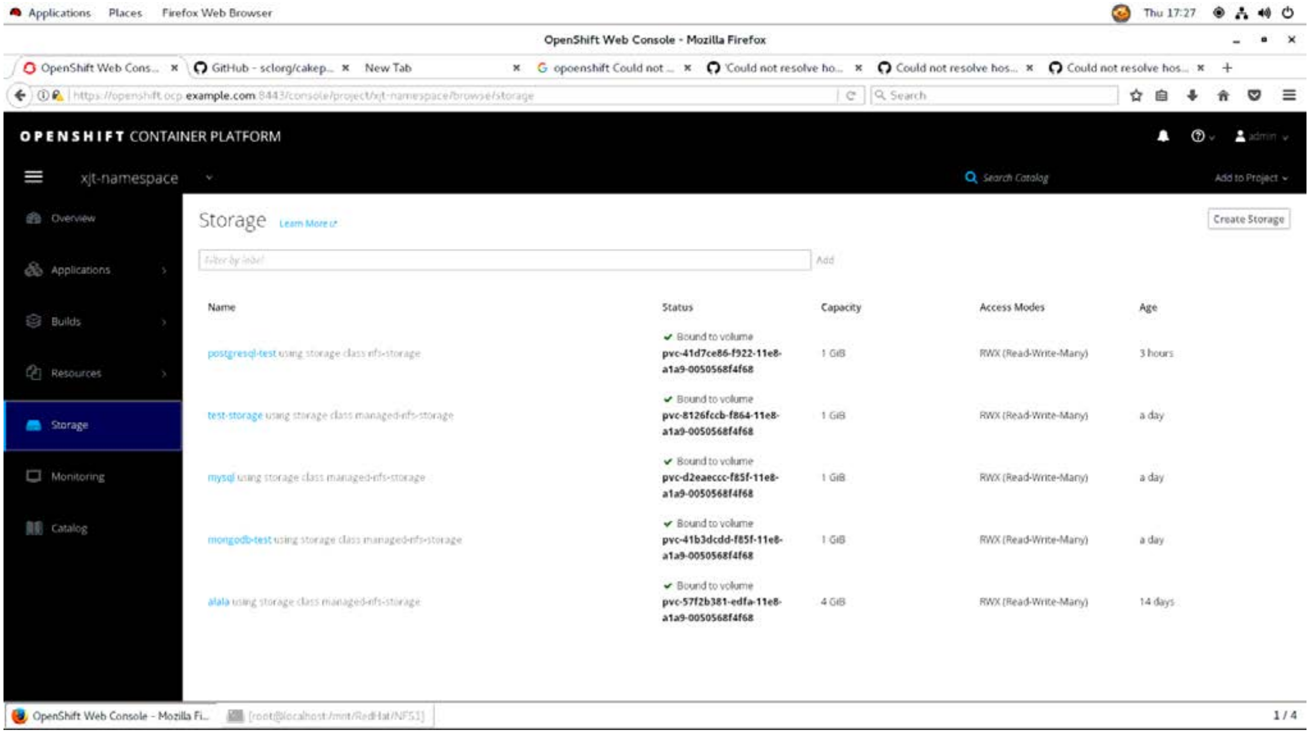
DEPLOYMENT CONFIG
nginx-example, #2
1 pod

APPLICATION
postgresql-persistent
http://postgresql-example-xjt-namespace.apps.ocp.example.com

DEPLOYMENT CONFIG

Figure 42 OpenShift Application Components

Figure 43 shows storage status.



Storage

Name	Status	Capacity	Access Modes	Age
postgresql-test using storage class nfs-storage	✓ Bound to volume pvc-41d7ce86-f922-11e8-a1a9-0050568f4f68	1 GiB	RWX (Read-Write-Many)	3 hours
test-storage using storage class managed-nfs-storage	✓ Bound to volume pvc-8126fccb-f864-11e8-a1a9-0050568f4f68	1 GiB	RWX (Read-Write-Many)	a day
mysql using storage class managed-nfs-storage	✓ Bound to volume pvc-d2eaeccc-f85f-11e8-a1a9-0050568f4f68	1 GiB	RWX (Read-Write-Many)	a day
mongodb-test using storage class managed-nfs-storage	✓ Bound to volume pvc-41b3dcdd-f85f-11e8-a1a9-0050568f4f68	1 GiB	RWX (Read-Write-Many)	a day
alala using storage class managed-nfs-storage	✓ Bound to volume pvc-57f2b381-edfa-11e8-a1a9-0050568f4f68	4 GiB	RWX (Read-Write-Many)	14 days

Figure 43 OpenShift Storage Status

8 VMware Horizon

Horizon View is a desktop virtualization product developed by VMware Inc. It provides remote desktop capabilities by using VMware virtualization technology and can deliver a consistent user experience across devices and locations while keeping corporate data secure and compliant. See this website for more details: vmware.com/products/horizon-view.

8.1 Solution overview

Figure 47 shows all of the main features of the Lenovo Hyper-converged Nutanix solution’s reference architecture with VMware Horizon 7.2 on VMware ESXi 6.7 U1 hypervisor. This chapter does not address the general issues of multi-site deployment and network management and limits the description to the components that are inside the customer’s intranet.

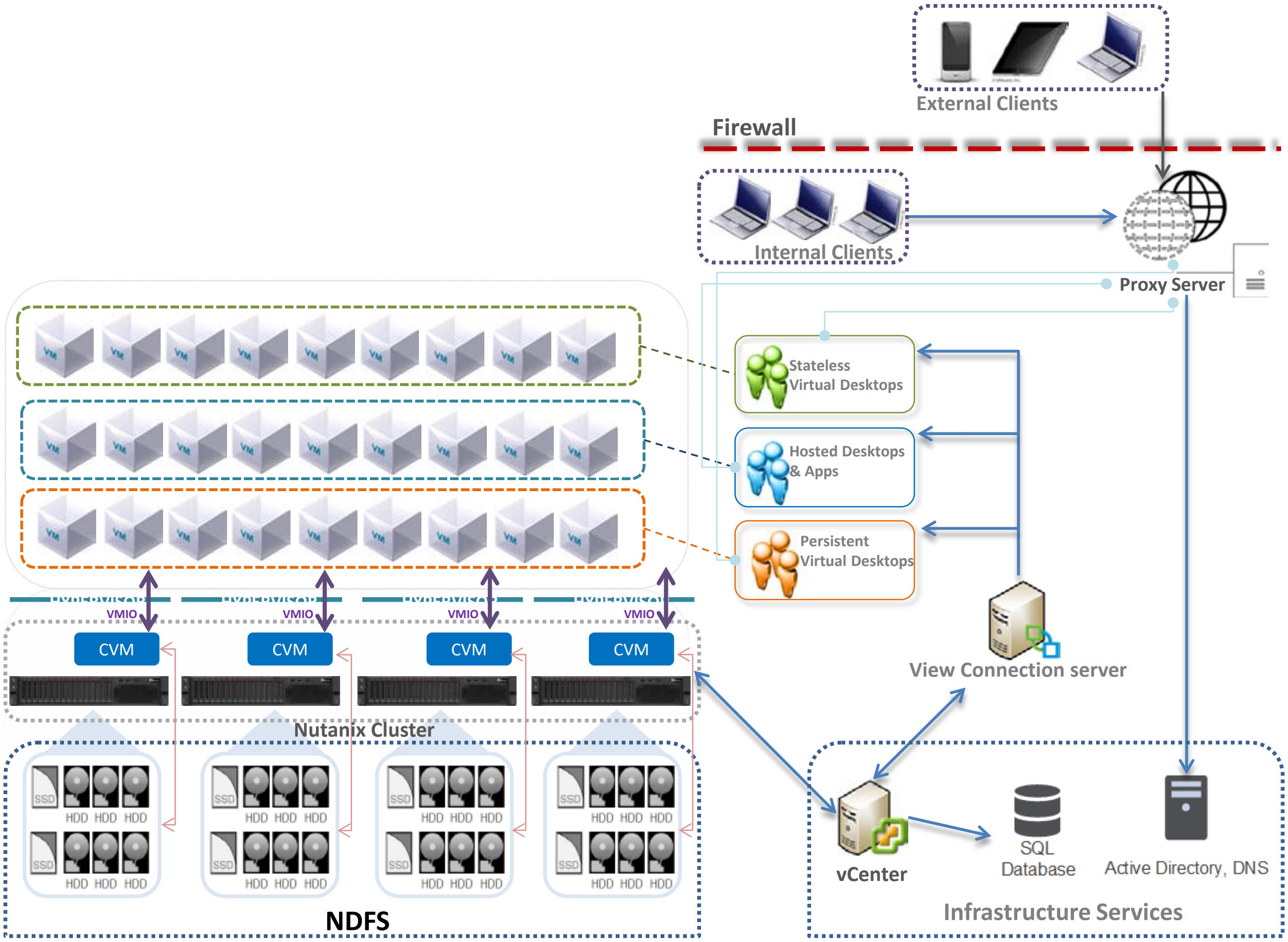


Figure 44: Lenovo ThinkAgile HX Series solution with VMware Horizon

8.2 Component model

Figure 48 is a layered component view for the VMware Horizon virtualization infrastructure.

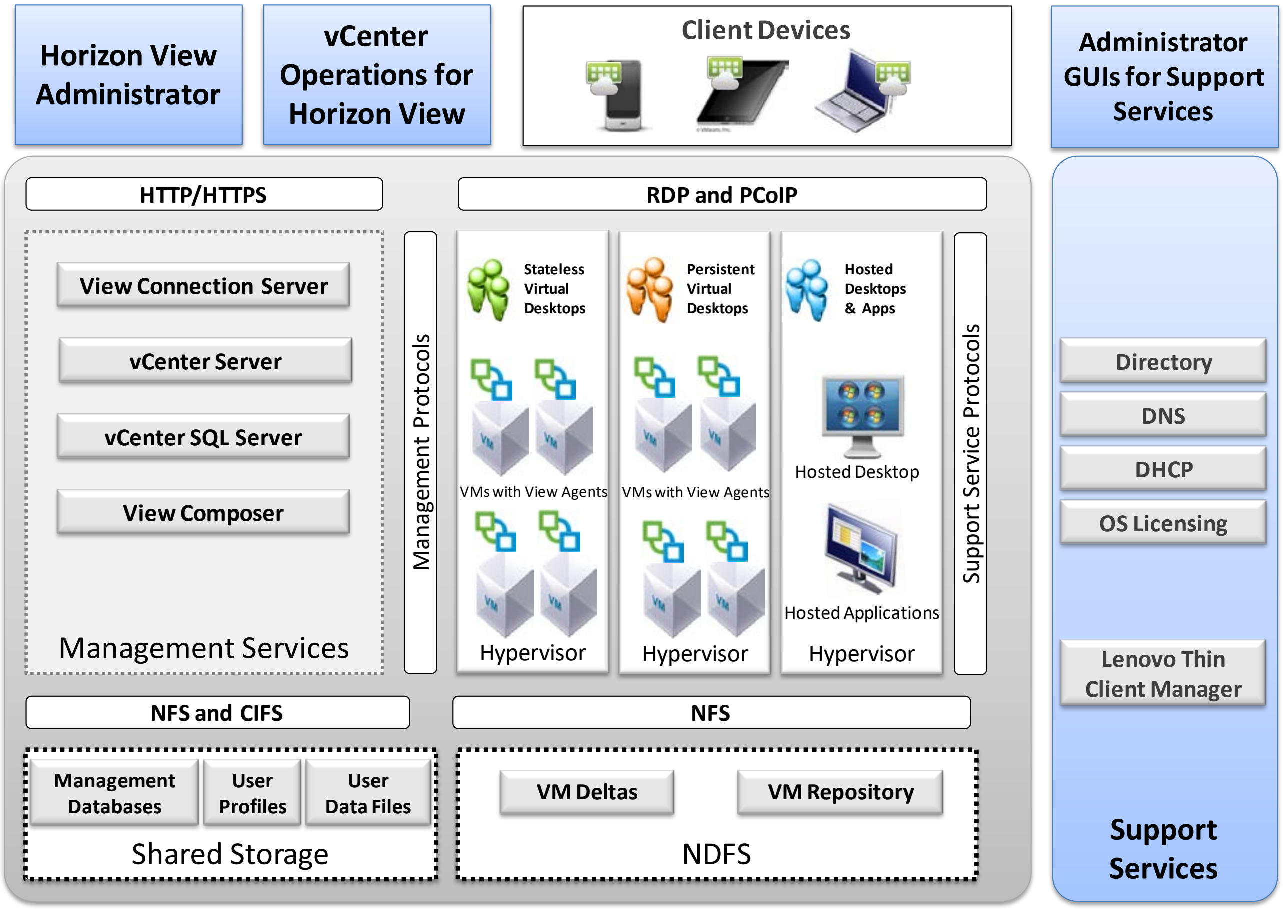


Figure 45: Component model with VMware Horizon

VMware Horizon with the VMware ESXi hypervisor features the following main components:

Horizon View Administrator By using this web-based application, administrators can configure ViewConnection Server, deploy and manage View desktops, control user authentication, and troubleshoot user issues. It is installed during the installation of ViewConnection Server instances and is not required to be installed on local (administrator) devices.

vCenter Operations for Horizon View This tool provides end-to-end visibility into the health, performance, and efficiency of the virtual desktop infrastructure (VDI) configuration. It enables administrators to proactively ensure the best user experience possible, avert incidents, and eliminate bottlenecks before they become larger issues.

View Connection Server The VMware Horizon Connection Server is the point of contact for client devices that are requesting virtual desktops. It authenticates users and directs the virtual desktop request to the appropriate virtual machine (VM) or desktop, which ensures that only valid users are allowed access. After the authentication is complete, users are directed to their assigned VM or desktop.

If a virtual desktop is unavailable, the View Connection Server works with the management and the provisioning layer to have the VM ready and available.

View Composer In a VMware vCenter Server instance, View Composer is installed. View Composer is required when linked clones are created from a parent VM.

vCenter Server By using a single console, vCenter Server provides centralized management of the virtual machines (VMs) for the VMware ESXi hypervisor. VMware vCenter can be used to perform live migration (called VMware vMotion), which allows a running VM to be moved from one physical server to another without downtime.

Redundancy for vCenter Server is achieved through VMware high availability (HA). The vCenter Server also contains a licensing server for VMware ESXi.

vCenter SQL Server vCenter Server for VMware ESXi hypervisor requires an SQL database. The vCenter SQL server might be Microsoft® Data Engine (MSDE), Oracle, or SQL Server. Because the vCenter SQL server is a critical component, redundant servers must be available to provide fault tolerance. Customer SQL databases (including respective redundancy) can be used.

View Event database VMware Horizon can be configured to record events and their details into a Microsoft SQL Server or Oracle database. Business intelligence (BI) reporting engines can be used to analyse this database.

Clients VMware Horizon supports a broad set of devices and all major device operating platforms, including Apple iOS, Google Android, and Google ChromeOS. Each client device has a VMware View Client, which acts as the agent to communicate with the virtual desktop.

Thin-client Manager The Lenovo Thin-client Manager (LTM) is used to manage and support Lenovo thin-client devices individually or in groups.

RDP, PCoIP The virtual desktop image is streamed to the user access device by using the display protocol. Depending on the solution, the choice of protocols available are Remote Desktop Protocol (RDP) and PC over IP (PCoIP).

For more information, see the Lenovo Client Virtualization base reference architecture document that is available at this website: lenovopress.com/lp0756.

8.3 VMware Horizon provisioning

VMware Horizon supports stateless and persistent virtual desktop models. Provisioning for VMware Horizon is a function of vCenter server and View Composer for linked clones.

vCenter Server allows for manually created pools and automatic pools. It allows for provisioning full clones and linked clones of a parent image for dedicated and stateless virtual desktops.

Because persistent virtual desktops use large amounts of storage, linked clones can be used to reduce the storage requirements. Linked clones are created from a snapshot (replica) that is taken from a golden master image. One pool can contain up to 2000 linked clones.

This document describes the use of automated pools (with linked clones) for dedicated and stateless virtual desktops. The deployment requirements for full clones using Nutanix de-duplication functionality is beyond the scope of this document.

8.4 Management VMs

A key part of the VMware Horizon environment is the various management VMs used to manage the VDI infrastructure and user VMs. Table 34 lists the VM requirements and performance characteristics of each management service.

Table 28: Characteristics of VMware Horizon management services

Management service VM	Virtual processors	System memory	Storage	Windows OS	HA needed	Performance characteristic
vCenter Server	8	12 GB	60 GB	2012 R2	No	Up to 2000 VMs.
vCenter SQL Server	4	8 GB	200 GB	2012 R2	Yes	Double the virtual processors and memory for more than 2500 users.
View Connection Server	4	16 GB	60 GB	2012 R2	Yes	Up to 2000 connections.

These management VMs can be run on separate servers from the HX series cluster or within the cluster itself. Separating out the VMs means that the management VMs can be separately managed and sized to the requirements and dedicated servers used for the user VMs. Putting all of the VMs together in one cluster means that the compute servers will execute less user VMs and need to leave enough resources for the much larger and more granular management VMs. Lenovo recommends that the management and user VMs are separated for all but the smallest deployments (i.e. less than 600 users).

Table 35 lists the number of management VMs for each size of users following the requirements for high-availability and performance. The number of vCenter servers is half of the number of vCenter clusters because each vCenter server can handle two clusters of up to 1000 desktops.

Table 29: Management VMs needed

Horizon management service VM	300 users	600 users	1200 users	3000 users
vCenter servers	1	1	2	2
vCenter SQL servers	2 (1+1)	2 (1+1)	2 (1+1)	2 (1+1)
View Connection Server	2 (1+1)	2 (1+1)	2 (1+1)	2 (1+1)

It is assumed that common services, such as Microsoft Active Directory, Dynamic Host Configuration Protocol (DHCP), domain name server (DNS), and Microsoft licensing servers exist in the customer environment.

8.5 Graphics acceleration

This section is specific to the Lenovo ThinkAgile HX3520-G and ThinkAgile HX3521-G that support GPU acceleration. The VMware ESXi hypervisor supports the following options for graphics acceleration:

- Dedicated GPU with one GPU per user, which is called virtual dedicated graphics acceleration (vDGA) mode.
- GPU hardware virtualization (vGPU) that partitions each GPU for 1 - 8 users.
- Shared GPU with users sharing a GPU, which is called virtual shared graphics acceleration (vSGA) mode and is not recommended because of user contention for shared use of the GPU.

The vDGA option has a low user density as it restricts a single user to access each very powerful GPU. This option is not flexible and is no longer cost effective even for high-end power users. Therefore vDGA is no longer recommended especially given that the performance of the equivalent vGPU mode is similar.

When using the vGPU option with ESXi 6.5 and the latest drivers from NVidia, it is necessary to change the default GPU mode from “Shared” (vSGA) to “Shared Direct” (vGPU) for each GPU using VMware vCenter. This enables the correct GPU support for the VMs which would otherwise result in the VM not powering on correctly and the standard “graphics resources not available” error message. The host needs to be rebooted for the changes to take effect.

The performance of graphics acceleration was tested using the Lenovo ThinkSystem SR650 servers. Each server supports up to two GPU adapters. The Heaven benchmark is used to measure the per user frame rate for different GPUs, resolutions, and image quality. This benchmark is graphics-heavy and is fairly realistic for designers and engineers. Power users or knowledge workers usually have less intense graphics workloads and can achieve higher frame rates. *Table 36* lists the results of the Heaven benchmark as FPS that are available to each user with the GRID 2,0 M60 adapter by using vGPU mode with DirectX 11.

Table 30: Performance of GRID 2.0 M60 vGPU modes with DirectX 11

Quality	Tessellation	Anti-Aliasing	Resolution	M60-8Q	M60-4Q	M60-2Q	M60-4A	M60-2A
High	Normal	0	1280x1024	Untested	Untested	32.03	59.81	32.98
High	Normal	0	1680x1050	Untested	49.97	25.41	N/A	N/A
High	Normal	0	1920x1200	Untested	41.36	21.03	N/A	N/A
Ultra	Extreme	8	1280x1024	Untested	Untested	18.02	37.01	18.68
Ultra	Extreme	8	1680x1050	56.01	29.67	14.18	N/A	N/A
Ultra	Extreme	8	1920x1080	50.38	25.69	12.76	N/A	N/A
Ultra	Extreme	8	1920x1200	46.01	22.79	Untested	N/A	N/A
Ultra	Extreme	8hai	2560x1600	27.42	14.16	Untested	N/A	N/A

Lenovo recommends that a medium to high powered CPU, such as the Xeon Scalable 6130, is used for accelerated graphics applications tend to also require extra load on the processor. For vGPU mode, Lenovo recommends at least 384GB of server memory. Because there are many variables when graphics acceleration is used, Lenovo recommends that testing is done in the customer environment to verify the performance for the required user workloads.

8.6 Performance testing

This section describes the performance benchmarking tool and the results obtained for different configurations of a cluster of 4 Lenovo ThinkAgile HX3320 appliances. This chapter applies equally to the ThinkAgile HX3321 certified nodes.

8.6.1 Login VSI benchmarking tool

Login VSI is a vendor-independent benchmarking tool that is used to objectively test and measure the performance and scalability of server-based Windows desktop environments. Leading IT analysts recognize and recommend Login VSI as an industry-standard benchmarking tool for client virtualization and can be used by user organizations, system integrators, hosting providers, and testing companies.

Login VSI provides multiple workloads to simulate real user work and suitable in performing load test, benchmarking and capacity planning for VDI environments. Table 37 lists the characteristics of the Login VSI 4.1 workloads that are used in the Lenovo testing.

Table 31. Login VSI Workload Comparison

Workload Name	Login VSI Version	Apps Open	CPU Usage	Disk Reads	Disk Writes	IOPS	Memory	vCPU
Office worker	4.1	5-8	82%	90%	101%	8.1	2GB	1vCPU
Knowledge worker	4.1	5-9	100%	100%	100%	8.5	2GB	2vCPU
Power worker	4.1	8-12	119%	133%	123%	10.8	3GB	3vCPU

The VSImax score parameter (the number indicates user density) is used to determine the performance of a particular system configuration. The following parameters and rules are used for Login VSI tests:

- User login interval: 30 seconds per node
- Workload: Office Worker, Knowledge Worker, or Power User
- All virtual desktops were pre-booted before the tests
- The number of powered-on VMs was adjusted to stay within a 10% margin of VSImax to avoid unreasonable overhead by “idling” virtual machines
- VSImax score is derived using the “classic model” calculation

8.6.2 Performance results for virtual desktops

This section shows the virtual desktop performance results for Lenovo ThinkAgile HX3320 appliances each configured with dual Xeon Scalable 6230 processors, 768 GB of memory, two 1.92 TB SATA SSDs, and six 2 TB SATA disk drives.

The recommended configuration of the Nutanix CVM is as follows:

vCPU	12
CPU Reservation	10000 MHz
Memory	24GB
Memory Reservation	24GB
NUMA	No Affinity
Advance CPU – Scheduling Affinity	No Affinity

Table 38 lists the Login VSI performance results of a HX Series appliance 4 node cluster using VMware ESXi 6.7 U1 and Windows 10 with the latest security fixes and the new Xeon Scalable Processors generation 2.

Table 32: Login VSI Performance

Processor	Workload	Stateless	Dedicated
Two Scalable 6230 processors 2.10 GHz, 20C 125W	Office worker	897 users	884 users
Two Scalable 6230 processors 2.10 GHz, 20C 125W	Knowledge worker	693 users	697 users
Two Scalable 6230 processors 2.10 GHz, 20C 125W	Power worker	587 users	600 users

8.6.3 Performance results from boot storm testing

A boot storm occurs when a substantial number of VMs are all booted within a short period of time. Booting a large number of VMs simultaneously requires large IOPS otherwise the VMs become slow and unresponsive.

Different numbers of VMs were booted on a cluster of 4 HX3320 hybrid appliances. The VMs were unpowered in vCenter and the boot storm created by powering on all of the VMs simultaneously. The time for all of the VMs to become visible in VMware Horizon was measured.

Figure 49 shows a comparison of the boot times for a variety of VMs. With an even spread of VMs on each node, the boot time for the VMs on each node was similar to the overall cluster boot time.

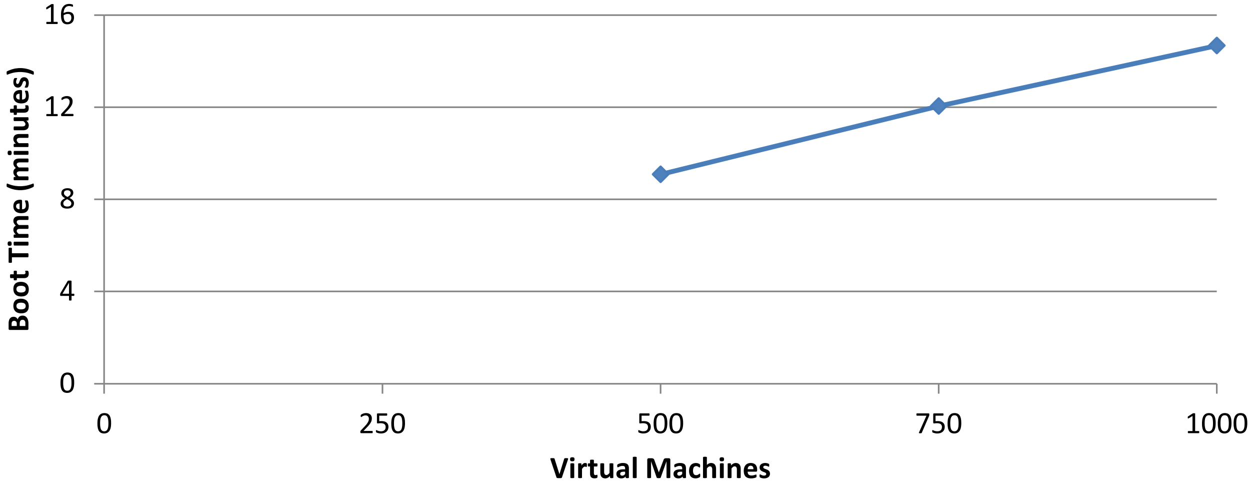


Figure 46: Boot storm comparison

8.7 Performance recommendations

This section provides sizing performance recommendations and best practices.

8.7.1 Sizing recommendations for virtual desktops

The default recommendation is two Xeon Scalable 6230 processors and 768 GB of system memory because this configuration provides the best coverage and density for a range of users. Assuming there is enough storage configured for the VMs, this configuration is recommended for any of the HX 3000 Series appliances.

For an office worker, Lenovo testing shows that 150 users per server is a good baseline and has an average of 82% usage of the processors in the server. If a server goes down, users on that server must be transferred to the remaining servers. For this degraded failover case, Lenovo testing shows that 180 users per server have an average of 88% usage of the processors. It is important to keep this 25% headroom on servers to cope with possible failover scenarios. Lenovo recommends a general failover ratio of 5:1. By using a target of 150 users per server, the maximum number of office workers is 9,600 in a 64 node cluster.

For a knowledge worker, Lenovo testing shows that 125 users per server is a good baseline and has an average of 76% usage of the processors in the server. For the degraded failover case, Lenovo testing shows that 150 users per server have an average of 85% usage of the processors. By using a target of 125 users per server, the maximum number of knowledge workers is 8,000 in a 64 node cluster.

For a power worker, Lenovo testing shows that 100 users per server is good baseline and has an average of 77% usage of the processors in the server. For the degraded failover case, Lenovo testing shows that 125 users per server have an average of 85% usage of the processors. By using a target of 125 users per server, the maximum number of power workers is 6,400 in a 64 node cluster.

Table 39 summarizes the processor usage with ESXi for the recommended user counts for normal mode and failover mode.

Table 33: Processor usage

Processor	Workload	Users per Server	CPU Utilization
Two 6230	Office worker	150 users – Normal Mode	82%
Two 6230	Office worker	180 users – Failover Mode	88%
Two 6230	Knowledge worker	125 users – Normal Mode	76%
Two 6230	Knowledge worker	150 users – Failover Mode	85%
Two 6230	Power worker	100 users – Normal Mode	77%
Two 6230	Power worker	125 users – Failover Mode	85%

Table 40 lists the recommended number of virtual desktops per server for different workload types and VM memory sizes. The number of users is reduced in some cases to fit within the available memory and still maintain a reasonably balanced system of compute and memory.

Table 34: Recommended number of virtual desktops per server

Workload	Office worker	Knowledge worker	Power worker
Processor	Two 6230	Two 6230	Two 6230
VM memory size	3 GB	4 GB	5 GB
System memory	768 GB	768 GB	768 GB
Memory overhead of CVM	24 GB	24 GB	24 GB
Desktops per server (normal mode)	150	125	100
Desktops per server (failover mode)	180	150	125

Table 41 lists the approximate number of compute servers that are needed for different numbers of users and Office worker workloads.

Table 35: Compute servers needed for Office workers and different numbers of users

Office workers	300 users	600 users	1200 users	3000 users
Compute servers @150 users (normal)	3	5	8	20
Compute servers @180 users (failover)	2	4	7	17

Table 42 lists the approximate number of compute servers that are needed for different numbers of users and Knowledge worker workloads.

Table 36: Compute servers needed for Knowledge workers and different numbers of users

Knowledge workers	300 users	600 users	1200 users	3000 users
Compute servers @125 users (normal)	3	5	10	24
Compute servers @150 users (failover)	2	4	8	20

Table 43 lists the approximate number of compute servers that are needed for different numbers of users and power worker workloads.

Table 37: Compute servers needed for Power workers and different numbers of users

Power workers	300 users	600 users	1200 users	3000 users
Compute servers @100 users (normal)	3	6	12	30
Compute servers @125 users (failover)	3	5	10	24

8.7.2 Best practices

The number of desktops that can be run on a specific server depends upon the available system memory, compute power of the processors, and number of logons per second during a logon storm. For a cost-effective solution, the maximum number of users should be put on each server to balance processor, memory, storage I/O, and networking. Lenovo recommends using all flash appliances for situations where the user logon rate is high or time to reboot all the VMs on a node must be less than 10 minutes.

Another important consideration for compute servers is system memory. For stateless users, the typical range of memory that is required for each desktop is 2 GB - 4 GB. For dedicated users, the range of memory for each desktop is 2 GB - 6 GB. In general, power users that require larger memory sizes also require more virtual processors. This reference architecture standardizes on 2 GB per desktop as the minimum requirement of a Windows 10 desktop. The virtual desktop memory should be large enough so that swapping is not needed and vSwap can be disabled.

It is a best practice not to overcommit on memory as swapping to disk can have a severe effect on performance; a better strategy is to give each desktop more memory. Alternatively, a monitoring tool can be run to gather information about existing desktops. The desktop memory size that is required does not necessarily have to match the memory supplied in a desktop machine; it can be larger or smaller.

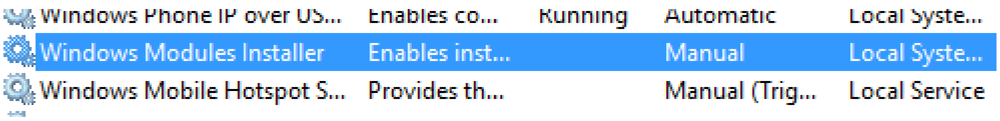
Lenovo recommends the use of VLANs to partition the network traffic. The following VLANs should be used:

- User (for web protocols, display protocols, and support service protocols)
- Management (for management protocols)
- Storage (for NDFS)

Lenovo recommends to always perform user virtualization, even if users have dedicated desktops. This separation of user-specific data makes it much easier to manage and perform upgrades.

Windows 10 was used for all of the performance testing. In general Windows 10 requires 10% to 20% more compute power than Windows 7. The following optimizations are recommended for the Windows 10 base image:

- Applied #VDILIKEAPRO Tuning Template(developed by loginVSI) – see the following for more details: loginvsi.com/blog/520-the-ultimate-windows-10-tuning-template-for-any-vdi-environment
- Set Adobe acrobat as a default app for PDF files using steps in following webpage: adobe.com/devnet-docs/acrobatetk/tools/AdminGuide/pdfviewer.html
- Disabled **Windows Modules installer** service on the base image because the CPU utilization can remain high after rebooting all the VMs. By default this service is set to manual rather than disabled.



Windows Phone IP over US...	Enables co...	Running	Automatic	Local Syste...
Windows Modules Installer	Enables inst...		Manual	Local Syste...
Windows Mobile Hotspot S...	Provides th...		Manual (Trig...	Local Service

Please refer below links for best practices and optimizations recommended by VMware:

- View Architecture Planning – VMware Horizon 6.0: pubs.vmware.com/horizon-view-60/topic/com.vmware.ICbase/PDF/horizon-view-60-architecture-planning.pdf
- VMware Horizon 6 with View Performance and Best Practices: vmware.com/files/pdf/view/vmware-horizon-view-best-practices-performance-study.pdf

8.8 Deployment ready solutions

There are 3 deployment ready solutions for VMware Horizon in the Lenovo Data Center Solution Configurator (DCSC): https://dcsc.lenovo.com/#/reference-solution.

Solution Identifier	SID0000116	SID0000118	SID0000120
ThinkAgile HX node	HX3720 or HX3721	HX3320 or HX3321	HX3520-G or HX3521-G
Office Workers (3GB)	180 users/node	180 users/node	N/A
Knowledge workers (4GB)	150 users/node	150 users/node	N/A
Power users (5GB)	N/A	125 users/node	N/A
Power user with GPU (6GB)	N/A	N/A	48 users/node
Light Cad with GPU (8GB)	N/A	N/A	16 users/node
Heavy Cad with GPU (16GB)	N/A	N/A	8 users/node

9 VMware vCloud Suite

VMware vCloud Suite is an integrated offering that brings together VMware's industry-leading vSphere hypervisor and VMware vRealize Suite multi-vendor hybrid cloud management platform. This chapter covers the following VMware products:

- vSphere 6.7 (and vSphere 6.5 U2), which provides compute virtualization
- vCloud Suite 7.4, which provides a VMware vSphere-based private cloud using vRealize Suite products and additional products to support vSphere Data Protection and Availability
- vRealize Suite 7.4, which provides cloud management capabilities for private, public and hybrid clouds with support for multiple hypervisors
- AWS Server Migration Service Connector 1.0.12.50 which supports migration of virtual machines from on-premise vSphere cloud to AWS pubic cloud.

VMware NSX 6.4.1 provides network virtualization by using software defined networking (SDN) and supports integration with hardware layer 2 gateways. The use of NSX with Lenovo ThinkAgile HX Series appliances and certified nodes was verified but a full description is outside the scope of this Reference Architecture. For more details see: www.nutanix.com/go/vmware-nsx-for-vsphere.php.

9.1 Solution Overview

This section gives an architectural overview of vCloud Suite products. Figure 50 gives an overview of how those products are deployed into shared edge and compute, management, and additional compute clusters.

This separation of function into these clusters allows for scaling in larger environments.

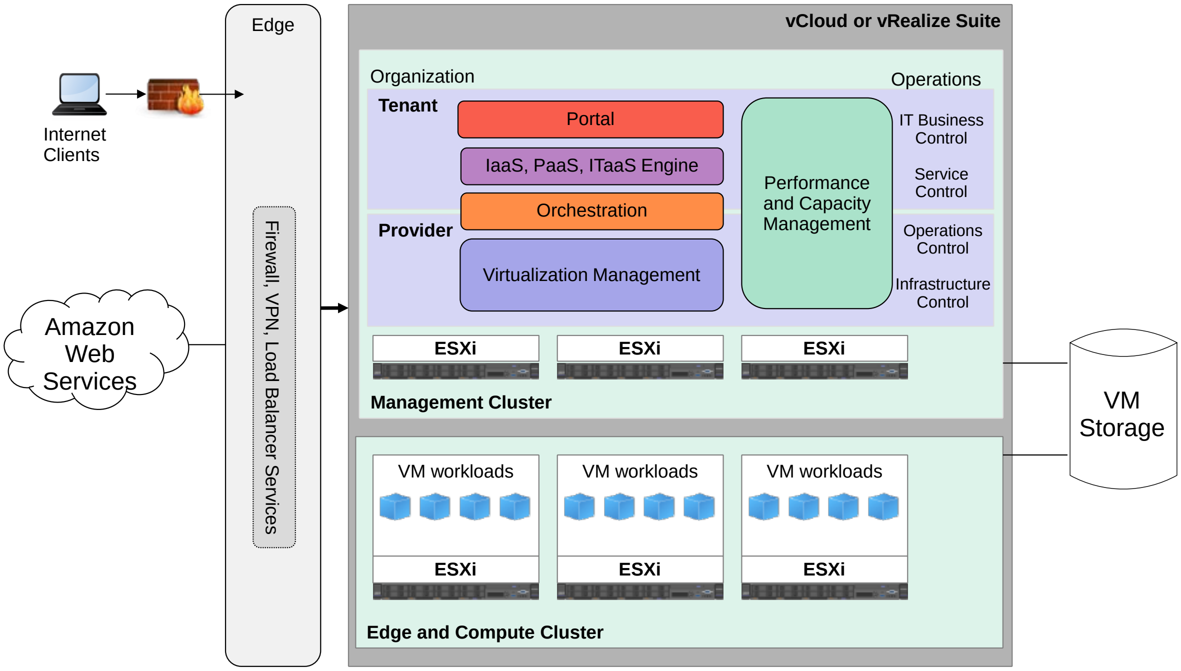


Figure 47: Conceptual design of vCloud Suite

The management cluster runs the components required to support vCloud Suite and is used for management, monitoring, and infrastructure services. A management cluster provides resource isolation which helps these services to operate at their best possible performance level. A separate cluster can satisfy an organization's policy to have physical isolation between management and production hardware and a single management cluster is required for each physical location.

The shared edge and compute cluster supports virtualized infrastructure services as well as network devices that provide interconnectivity between environments. It provides protected capacity by which internal data center networks connect via gateways to external networks. Networking edge services and network traffic management occur in this cluster and all external facing network connectivity ends in this cluster. The shared edge and compute cluster also supports the delivery of all other (non-edge) customer workloads and there can be one or more compute clusters, depending on the customer environment. Multiple compute clusters can be for different organizations or tenants, different workload types, or to spread the load in a large enterprise.

9.2 Component model

This section describes the component model for VMware vCloud Suite and optionally extending it into public clouds with hybrid cloud connections.

Figure 51 shows an overview of the major components of the VMware vCloud Suite.

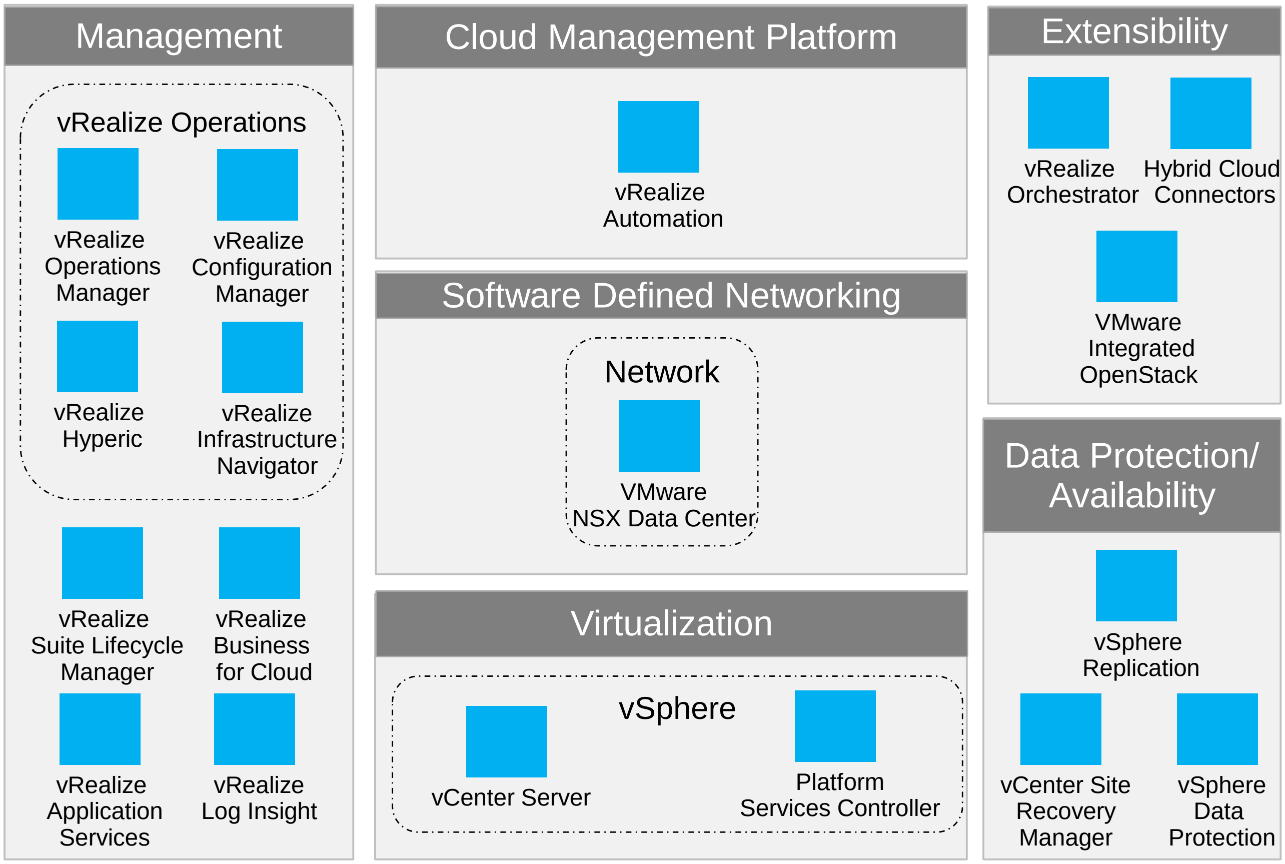

Management
vRealize Operations
vRealize Operations Manager
vRealize Configuration Manager
vRealize Hyperic
vRealize Infrastructure Navigator
vRealize Suite Lifecycle Manager
vRealize Business for Cloud
vRealize Application Services
vRealize Log Insight
Cloud Management Platform
vRealize Automation
Software Defined Networking
Network
VMware NSX Data Center
Virtualization
vSphere
vCenter Server
Platform Services Controller
Extensibility
vRealize Orchestrator
Hybrid Cloud Connectors
VMware Integrated OpenStack
Data Protection/ Availability
vSphere Replication
vCenter Site Recovery Manager
vSphere Data Protection

Figure 48: vCloud Suite components

The VMware vCloud Suite features the following components:

ESXi hypervisor	Provides bare-metal virtualization of servers so you can consolidate your applications on less hardware.

vCenter Server	Provides a centralized platform for managing vSphere environments and includes vSphere replication and vSphere data protection.
Platform Services Controller (PSC)	Provides a set of common infrastructure services that encompasses single sign-on (SSO), licensing, and a certificate authority (CA).
vRealize Suite Lifecycle Manager	Provides deployment options such as install, configure, import, and upgrade vRealize Suite environments and perform drift analysis and view the health of those environments
vRealize Automation	Provides a self-service, policy-enabled IT and application services catalog for deploying and provisioning of business-relevant cloud services across private and public clouds, physical infrastructure, hypervisors, and public cloud providers.
vRealize Operations	Provides a set of components for automation of operations including infrastructure health, configurations and compliance, application discovery, and monitoring of hardware and software.
• vRealize Operations Manager	Provides comprehensive visibility and insights into the performance, capacity and health of your infrastructure.
• vRealize Configuration Manager	Provides automation of configuration and compliance management across your virtual, physical, and cloud environments, which assesses them for operational and security compliance.
• vRealize Infrastructure Navigator	Provides automated discovery of application services, visualizes relationships, and maps dependencies of applications on virtualized compute, storage, and network resources.
• vRealize Hyperic	Provides monitoring of operating systems, middleware, and applications that are running in physical, virtual, and cloud environments.
vRealize Business for Cloud	Provides transparency and control over the costs and quality of IT services that are critical for private (vCloud Suite) or hybrid cloud (vRealize Suite) success.
vRealize Log Insight	Provides analytics capabilities to unstructured data and log management, which gives operational intelligence and deep, enterprise-wide visibility across all tiers of the IT infrastructure and applications. Standard for vRealize Suite.
vCenter Site Recovery Manager (SRM)	Provides disaster recovery capability with which you can perform automated orchestration and non-disruptive testing for virtualized applications by using ESXi hypervisor only. SRM is standard for vCloud Suite and optional for vRealize Suite.

vRealize Orchestrator

Provides the capability to create workflows that automate activities, such as provisioning VM, performing scheduled maintenance, and starting backups.

NSX

NSX provides virtualization of networking and is optional for vCloud Suite deployments.

VMware Integrated OpenStack (VIO)

Provides a VMware-supported OpenStack distribution (distro) that makes it easier for IT to run a production-grade, OpenStack-based deployment on top of their VMware infrastructure. For more information, see this website: vmware.com/products/openstack.

Hybrid Cloud Connectors

Allows an administrator to provide hybridization using public cloud providers such as Amazon AWS. See the next section for more information.

The vCloud Suite products also have dependencies on the following external components:

Identity source

Identity sources (Active Directory, OpenLDAP, or Local OS) or similar is required to implement and operate the vCloud Suite or vRealize Suite infrastructure.

DNS

DNS must be configured for connectivity between vCenter Server, Active Directory, ESXi hosts, and the VMs

DHCP/TFTP

PXE boot is required for vSphere Auto Deploy functionality.

Time synchronization

Accurate time keeping and time synchronization is critical for a healthy infrastructure. All components (including ESXi hosts, vCenter Server, the SAN, physical network infrastructure, and VM guest operating systems) must have accurate time keeping.

Microsoft SQL Server or Oracle database

Many of the vCloud Suite components come with embedded databases or they can use external databases such as Microsoft SQL Server or Oracle, depending on the component and the intended environment.

Other software components such as Lenovo XClarity Administrator are not shown. As well as providing management of Lenovo hardware, XClarity Administrator also has plugins for VMware vCenter, VMware vRealize Orchestrator, and VMware vRealize Log Insight which are further described in “Systems management” on page 109.

9.2.1 Hybrid Clouds

On-premise VMware vCloud suite connects to public clouds such as Amazon Web Services (AWS), IBM Public Cloud and OVH Public Cloud. This document only discusses AWS support.

The Amazon Elastic Compute Cloud (EC2) provides scalable computing capacity in the Amazon Web Services (AWS) public cloud by offering compute, storage, networking, software, and development tools. AWS provides Virtual Private Cloud and Dedicated Hosts for compute and different services. It supports a hybrid architecture by integrating networking, security and access control, automated workload migrations and controlling AWS from an on-premise infrastructure management tool.

AWS Server Migration Service (AWS SMS) is an agentless service to migrate on-premise workloads from vCenter to AWS. It supports incremental replications and migration can be performed faster while minimizing network bandwidth consumption and reducing server downtime. Each server volume replicated is saved as a new Amazon Machine Image (AMI) which can be launched as an EC2 instance (VM) in the AWS cloud. The AWS Server Migration Service replicates server volumes from on-premises environment to S3 temporarily and purges them from S3 immediately after creating the Elastic Block Store (EBS) snapshots.

Table 44 describes the features supported by the AWS SMS connector. Some best practices for deployment are described in "Hybrid networking to public clouds" on page 44.

Table 38: Hybrid Cloud Features

Feature	**AWS SMS Connector**
Bi-Directional Migration	No
Integration with vCenter	Yes
vCenter→Connector Association	Many-Many
Connector→Public Cloud Association	1-1
vCenter linked Mode support	No
Integration with vSphere client	Use AWS Management Console
Integration with vSphere web client	Use AWS Management Console
Integration with vSphere Replication	No
Integration with vRealize Automation	No
Multi-tenant support	No
VM Management public cloud	Use AWS Management Console
VM Management on-premise	No
Migration to all public cloud regions	Limited Currently
Copy/Migrate Templates to public cloud	Yes (AMI)
Deploy VM from Template to public cloud	Yes (AMI)
Live Migration to public cloud	Yes
Cold Migration to public cloud	Yes
Bulk Migration to public cloud	Yes
Layer 2 Extension	No
Migrate VM to on-premise	No (OVA download)
Offline Data Transfer	No
Common Content Library	No
Number of Concurrent Migration	50 per account
License	90 days Free
vSphere Standard Switch Support	Yes
vSphere Distributed Switch Support	Yes
Network(NSX) Policy Migration	No

9.2.2 VMware Licensing

The licensing for vSphere is based on a CPU metric and licensing for other products is based on the number of OS instances. The vCloud Suite license is a single perpetual license that includes vSphere Enterprise Plus and vRealize Suite. For vSphere environments, the license can be purchased separately for vSphere Enterprise Plus and vRealize Suite. Other components have their own separate licenses and are optional add-ons.

Table 45 lists the standard and optional components that are provided with a vCloud Suite License or vRealize Suite License.

Table 39: VMware Licensing

License	Component	vCloud Standard	vCloud Advanced	vCloud Enterprise	vRealize Standard	vRealize Advanced	vRealize Enterprise
Base	vSphere	Enterprise Plus	Enterprise Plus	Enterprise Plus			
	vRealize Suite Lifecycle Mgr	Included	Included	Included	Included	Included	Included
	vRealize Automation	N/A	Advanced	Enterprise	N/A	Advanced	Enterprise
	vRealize Operations	Advanced	Advanced	Enterprise	Advanced	Advanced	Enterprise
	• vRealize Operations Mgr • vRealize Configuration Mgr • vRealize Infrastructure Nav						
	• vRealize Hyperic						
	vRealize Business for Cloud	Standard	Advanced	Advanced	Standard	Advanced	Advanced
	vSphere Replication	Included	Included	Included	N/A	N/A	N/A
	vSphere Data Protection	Included	Included	Included	N/A	N/A	N/A
	vSphere Big Data Extensions	Included	Included	Included	N/A	N/A	N/A
	vRealize Log Insight	Included	Included	Included	Included	Included	Included
	vCenter Site Recovery Mgr	Add On	Add On	Add On	Add On	Add On	Add On
Other licenses	vRealize Orchestrator	Included	Included	Included	Add On	Add On	Add On
	vRealize Automation Public Cloud Extension	N/A	Add On	Add On	Add On	Add On	Add On
	vRealize Operations Public Cloud Extension	N/A	Add On	Add On	Add On	Add On	Add On
	NSX Data Center for vSphere	Included	Included	Included	Add On	Add On	Add On
	VMware Integrated OpenStack	Add On	Add On	Add On	Add On	Add On	Add On

9.3 Shared edge and compute cluster

The shared edge and compute cluster uses its own dedicated vCenter server.

9.3.1 Edge and Infrastructure Services VMs

The VMs used for infrastructure services such as Active Directory, DNS/DHCP, firewalls, proxy and anti-virus are deployed in the shared edge and compute cluster. Table 46 lists each infrastructure service VM with the recommended sizes in terms of virtual CPUs, RAM, storage, and networking.

Table 40: Infrastructure services VMs

VM description	CPU (vCPUs)	Memory (GB)	Storage (GB)	Network bandwidth	High availability
AD, DHCP, DNS server	2	4	70	1 GbE	clustered
http proxy server	2	4	30	1 GbE	clustered

9.3.2 Hybrid cloud VMs

Table 47 lists the cloud connectivity VMs with the recommended sizes in terms of virtual CPUs, RAM, storage, networking, and location. Note that these VMs do not have options for high availability.

Table 41: Cloud connectivity VMs

VM description	CPU (vCPUs)	Memory (GB)	Storage (GB)	Network bandwidth	Location
AWS SMS Connector for vCenter	2	4	300	1 GbE	On-Premise

9.3.3 Server configuration

Since the shared cluster hosts compute workloads and edge services, the servers need to be sized appropriately. See other chapters in this Reference Architecture for specific workloads.

9.3.4 Load balancing and protection

An essential part of the infrastructure is load balancing of the server VMs and recognizing when a server is down and failing over to a second server.

For the shared edge and compute cluster connected to the Internet, it is also important to provide a firewall and protection against external threats. There are many ways to solve these problems such as using a F5 Big-IP edge gateway device or virtual machine. Using F5 protection and load balancing is outside the scope of this document.

9.4 Management cluster

The number of VMware vCloud Suite components in the management cluster increases as capabilities are added. This section addresses the management components that could be used. Third party add-ons must be sized separately.

9.4.1 Management cluster VMs

There are several considerations that contribute to an end-to-end sizing of an entire VMware vCloud environment including Lenovo software for systems management. This section is intended to provide some high-level guidance for management cluster configuration sizing. The recommended number of virtual CPUs, memory size, storage size, and network bandwidth is given for each VM and the VMs are grouped by each major component or appliance.

An essential part of the infrastructure is load balancing of the server VMs and recognizing when a server is down and failing over to another server. The following cases are available for VMs in the management cluster:

- vSphere HA: vCenter automatically restarts the VM on another server, but there is some downtime while the VM starts up.
- Microsoft SQL server clustering: The SQL server cluster automatically handles failover.
- Clustering within component to provide built-in high availability.
- Load balancing: An external load balancer such as a Big-IP switch from F5

Table 48 lists each management cluster VM for vSphere with its recommended size in terms of virtual CPUs, RAM, storage, and networking.

Table 42: Management cluster VMs for vSphere

VM description	CPU (vCPUs)	Memory (GB)	Storage (GB)	Network bandwidth	High availability
vCenter Server(1) Management Cluster	8	24	50	1 GbE	load balancer
vCenter Server(2) Edge and Compute Cluster	8	24	50	1 GbE	load balancer
vCenter Server Database (MS SQL)	4	8	200	1 GbE	SQL AlwaysOn Availability Group
Platform Service Controller (1) Management Cluster	2	4	50	1 GbE	load balancer
Platform Service Controller (2) Edge and Compute Cluster	2	4	50	1 GbE	load balancer
vSphere Replication	2	4	20	1 GbE	not required
vSphere Data Protection	4	4	1600	1 GbE	not required
vRealize Orchestrator Appliance	2	3	12	1 GbE	Clustered

Table 49 lists each management cluster VM for vRealize Automation with its size in terms of virtual CPUs, RAM, storage, and networking.

Table 43: Management cluster VMs for vRealize Automation

VM description	CPU (vCPUs)	Memory (GB)	Storage (GB)	Network bandwidth	High availability
vRealize Suite Lifecycle Manager	4	16	135	1 GbE	N/A
vRealize Automation Appliance	4	16	30	1 GbE	load balancer
IaaS Database (MS SQL)	8	16	100	1 GbE	SQL AlwaysOn Availability Group
Infrastructure Web Server	2	4	40	1 GbE	load balancer
Infrastructure Manager Server	2	4	40	1 GbE	load balancer
Distributed Execution Manager (DEM)	2	6	40	1 GbE	load balancer
vSphere Proxy Agent	2	4	40	1 GbE	load balancer
vRealize Application Services	8	16	50	1 GbE	vSphere HA

Table 50 lists each management cluster VM for vRealize Operations Manager with its size in terms of virtual CPUs, RAM, storage, and networking.

Table 44: Management cluster VMs for vRealize Operations Manager

VM description	CPU (vCPUs)	Memory (GB)	Storage (GB)	Network bandwidth	High availability
vRealize Operations Manager – Master	4	16	500	1 GbE	clustered
vRealize Operations Manager – Data	4	16	500	1 GbE	not required
vRealize Configuration Manager – Collector	4	16	72	1 GbE	not required
vRealize Configuration Manager Database (MS SQL)	4	16	1000	1 GbE	SQL AlwaysOn Availability Group
vRealize Hyperic Server	8	12	16	1 GbE	load balancer
vRealize Hyperic Server - Postgres DB	8	12	75	1 GbE	load balancer
vRealize Infrastructure Navigator	2	4	24	1 GbE	not required

Table 51 lists each of the remaining management cluster VMs.

Table 45: Other Management cluster VMs

VM description	CPU (vCPUs)	Memory (GB)	Storage (GB)	Network bandwidth	High Availability
vRealize Business Standard	2	4	50	1 GbE	vSphere HA
Site Recovery Manager	4	4	20	1 GbE	not required
Site Recovery Manager Database (MS SQL)	2	4	100	1 GbE	SQL AlwaysOn Availability Group
vRealize Log Insight	8	16	100	1 GbE	Cluster of 3 nodes

Table 52 lists the VMs that are needed for Lenovo software for systems management.

Table 46: Lenovo System Management VMs

VM description	CPU (vCPUs)	Memory (GB)	Storage (GB)	Network bandwidth	High availability
Lenovo XClarity Administrator	2	4	64	1 GbE	not required
Lenovo XClarity Integrator (Windows OS)	1	2	30	1 GbE	not required

9.4.2 Server configuration

The management cluster should have a minimum of four hosts for high availability. Because of the large number of management VMs that can be used in the management cluster, the following configuration is recommended for each server:

- Lenovo ThinkAgile HX332x or HX372x
- 2 x Intel® Xeon® Gold 6130 Processor (2.10 GHz 16 cores)
- 384 GB of system memory
- Dual M.2 boot drives with ESXi 6.5 U1

9.5 Hybrid networking to public clouds

This section contains deployment considerations for hybrid networking from an on-premise cloud to public clouds such as AWS.

9.5.1 AWS networking

Table 53 shows the required connectivity for each component used for AWS. . For more details, see the Shared Edge and Compute cluster underlay for the deployment example on page 59.

Table 47: AWS component connectivity

Virtual Machine	DvSwitch	http proxy support
AWS SMS Connector for vCenter	Edge-Compute	Yes

9.5.2 Best Practices

Amazon Web Services (AWS) provide internet and direct connect accessibility options to connect from the vSphere private cloud environments. The AWS connectors leverage an http proxy to establish connectivity to the AWS cloud.

The server hardware for an on-premise vSphere cloud may not use the same processor family. In this case Lenovo recommends enabling Enhanced vMotion Compatibility (EVC) on the cluster to avoid compatibility errors during vMotion.

AWS Connectors are deployed in an on-premise vSphere environment and they do not provide capabilities for bi-directional migration. The on-premise virtual machines can be migrated to AWS cloud but there is no option in the connectors to migrate a virtual machine back to vSphere environment. Instead the VMs need to be exported to OVA format and stored in S3. Then the image can be imported or downloaded to the on-premise vSphere environment.

9.6 Systems management

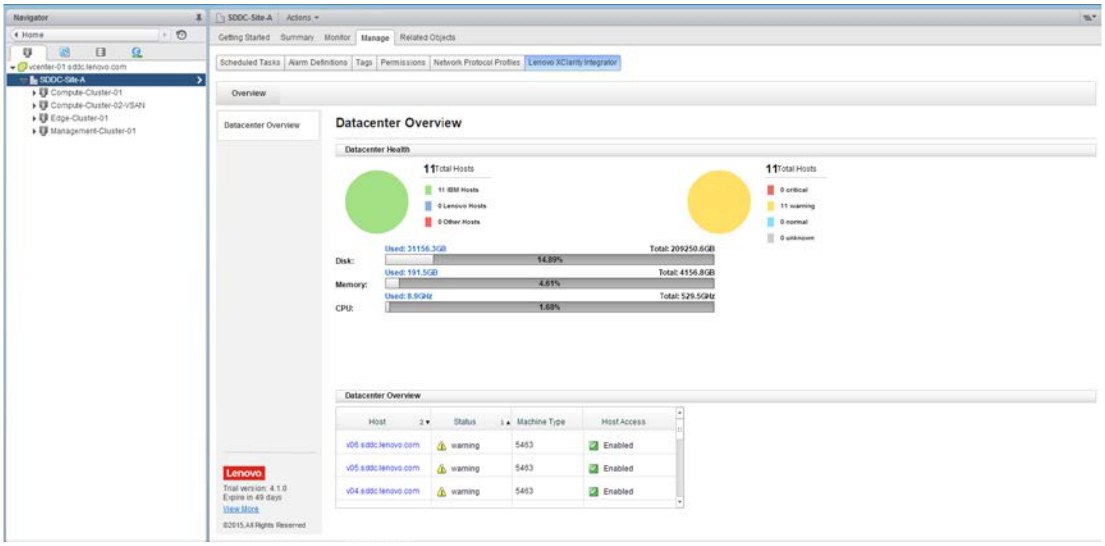
Lenovo XClarity™ Administrator is a centralized resource management solution that reduces complexity, speeds up response, and enhances the availability of Lenovo® server systems and solutions. See section 2.2.2 on page 6 for more details.

In addition Lenovo provides a number of plugins for VMware vCloud Suite components which are described below.

9.6.1 Lenovo XClarity integration

Lenovo also provides XClarity integration modules for VMware vCenter, VMware vRealize Orchestrator and VMware vRealize Log Insight. For more information, see this website: http://www3.lenovo.com/us/en/data-center/software/systems-management/c/systems-management.

By using the Lenovo XClarity Integrator for VMware vCenter, administrators can consolidate physical resource management in VMware vCenter, which reduces the time that is required for routine system administration.



Trial version: 4.1.0
Expire in 49 days
View More
©2015 All Rights Reserved

Datacenter Overview

Datacenter Health

11 Total Hosts
- 11 IBM Hosts
- 0 Lenovo Hosts
- 0 Other Hosts

11 Total Hosts
- 0 critical
- 11 warning
- 0 normal
- 0 unknown

Disk: Used: 31156.3GB 14.89% Total: 209250.6GB
Memory: Used: 191.5GB 4.61% Total: 4156.8GB
CPU: Used: 8.9GHz 1.68% Total: 529.5GHz

Datacenter Overview

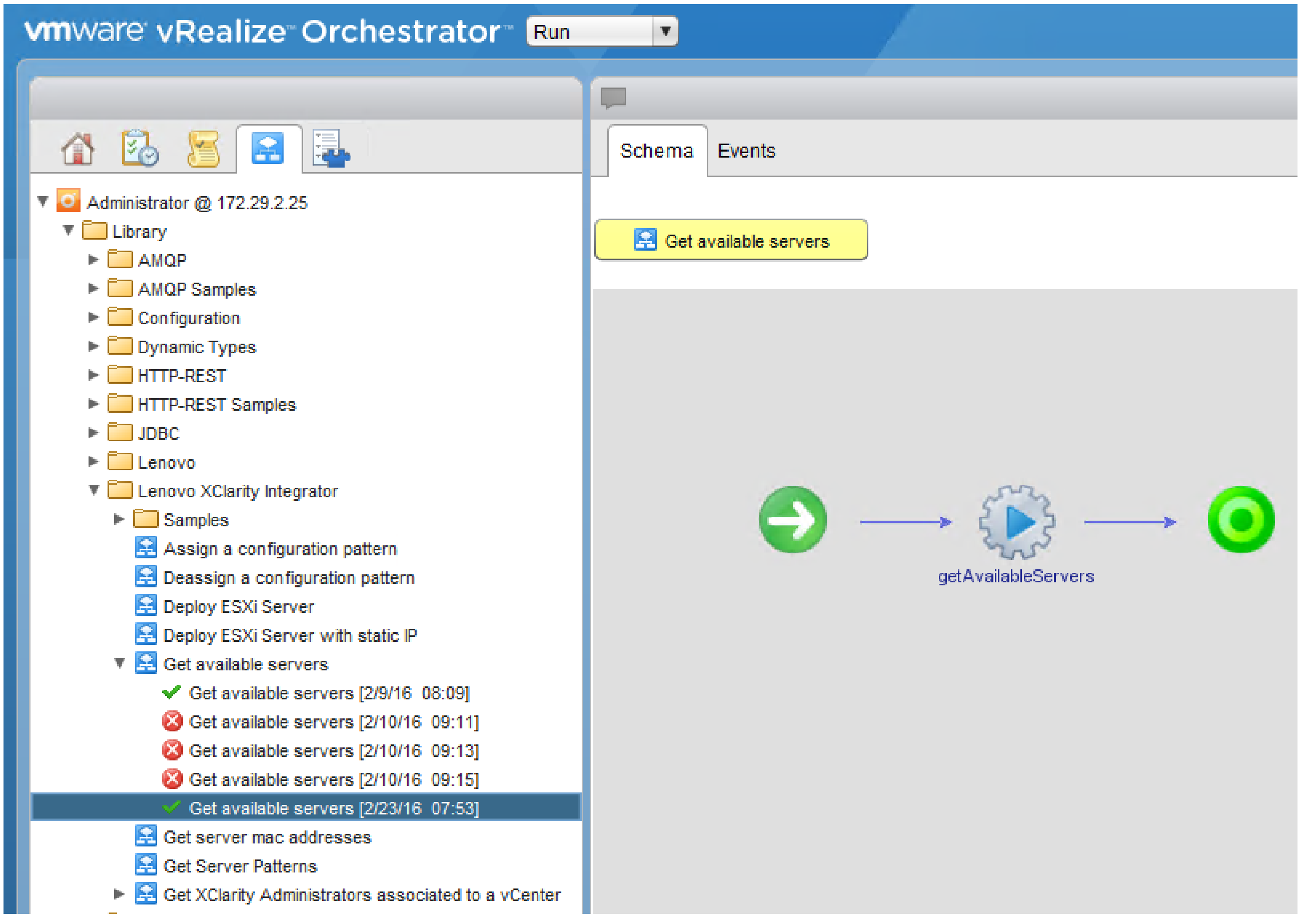
Host 2 ▾	Status 1 ▴	Machine Type	Host Access
v06.sddc.lenovo.com	⚠ warning	5463	✓ Enabled
v05.sddc.lenovo.com	⚠ warning	5463	✓ Enabled
v04.sddc.lenovo.com	⚠ warning	5463	✓ Enabled

Figure 49: Lenovo XClarity Integrator for VMware vCenter

The Lenovo XClarity Integrator for VMware vCenter provides the following features and benefits:

- Extends Lenovo XClarity Administrator features to the virtualization management console
- Enables management of legacy infrastructure from the virtualization management console
- Reduces workload downtime by dynamically triggering workload migration in clustered environments during rolling server reboots or firmware updates, and predicted hardware failures

The Lenovo XClarity Integrator for VMware vRealize Orchestrator provides IT administrators with the ability to coordinate physical server provisioning features of Lenovo XClarity Pro with broader vRealize Orchestrator workflows. Lenovo XClarity Integrator for VMware vRealize Orchestrator provides a library of simple yet robust and customizable workflow routines and actions designed to automate complex, repetitive IT infrastructure tasks such as system discovery and configuration, hypervisor installation, and addition of new hosts to vCenter.



- Administrator @ 172.29.2.25
  - Library
    - AMQP
    - AMQP Samples
    - Configuration
    - Dynamic Types
    - HTTP-REST
    - HTTP-REST Samples
    - JDBC
    - Lenovo
    - Lenovo XClarity Integrator
      - Samples
      - Assign a configuration pattern
      - Deassign a configuration pattern
      - Deploy ESXi Server
      - Deploy ESXi Server with static IP
      - Get available servers
        - ✓ Get available servers [2/9/16 08:09]
        - ☒ Get available servers [2/10/16 09:11]
        - ☒ Get available servers [2/10/16 09:13]
        - ☒ Get available servers [2/10/16 09:15]
        - ✓ Get available servers [2/23/16 07:53]
      - Get server mac addresses
      - Get Server Patterns
      - Get XClarity Administrators associated to a vCenter

Figure 50: Lenovo XClarity Integrator for VMware vRealize Orchestrator interface

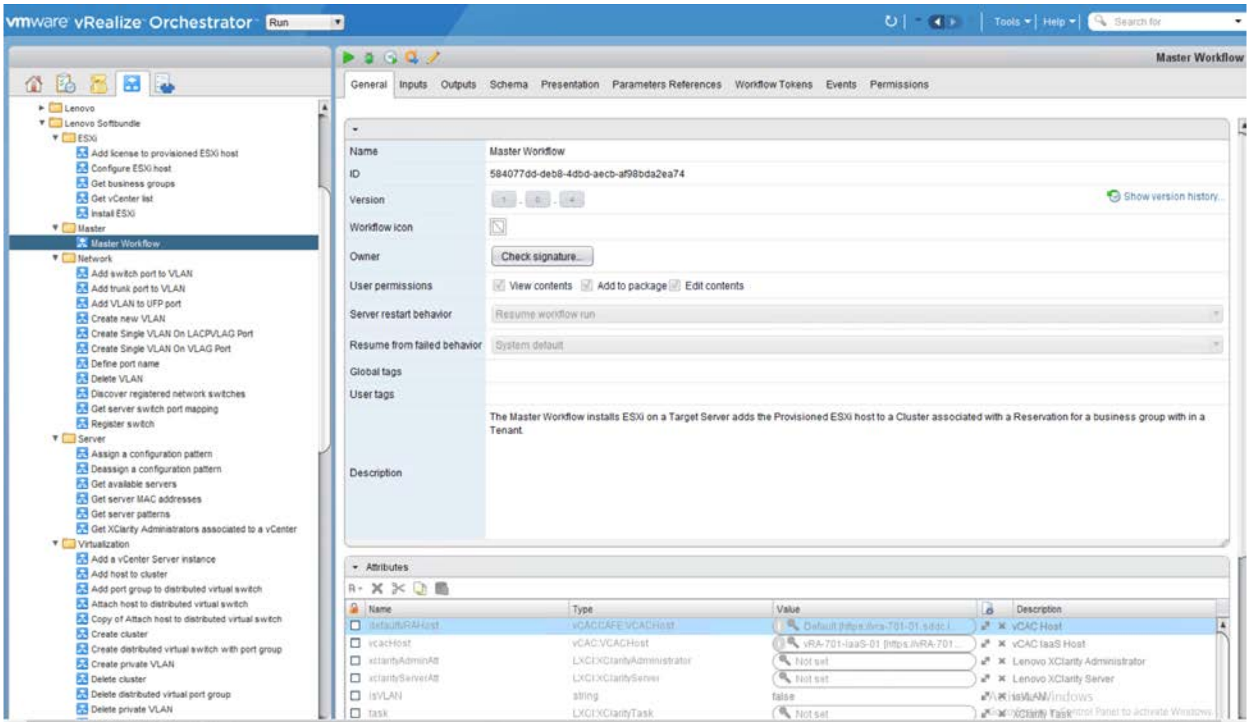
The Lenovo XClarity Administrator Content Pack for VMware vRealize Log Insight simplifies the collection and forwarding of Lenovo XClarity Administrator logs to VMware vRealize Log Insight for powerful processing and analytics, and displaying insightful information in an intuitive format.

The VMs for VMware vCenter, vRealize Orchestrator, Lenovo XClarity Administrator and Lenovo XClarity Administrator Integrator should have access to the management network used for managing servers, storage and networking.

Lenovo XClarity Integrator for vRealize Automation provides a set of blueprints to provision infrastructure services based on Lenovo servers, network switches and vSphere. This eases provisioning a new Lenovo server with vSphere installed, network isolation parameters configured on the Lenovo switches, apply vSphere distributed switch configurations and adding the server to the existing or new vSphere Cluster. These services leverage the workflows defined in the Lenovo vRealize SoftBundle for vRealize Orchestrator, Lenovo XClarity Integrator for vCenter, Lenovo XClarity Integrator for vRealize Orchestrator, and Lenovo Networking Integration plugin for vRealize Orchestrator.

The Lenovo vRealize SoftBundle package for vRealize Automation needs to imported into vRealize Orchestrator and then the Blueprints package is imported using the vRealize Cloud Client command line utility by Tenant Administrators and it creates catalog items automatically. The catalog items are created under Lenovo Servers, Lenovo Network, and Lenovo Virtualization services.

Figure 54 shows the workflows available in the Lenovo SoftBundle for vRealize Orchestrator.



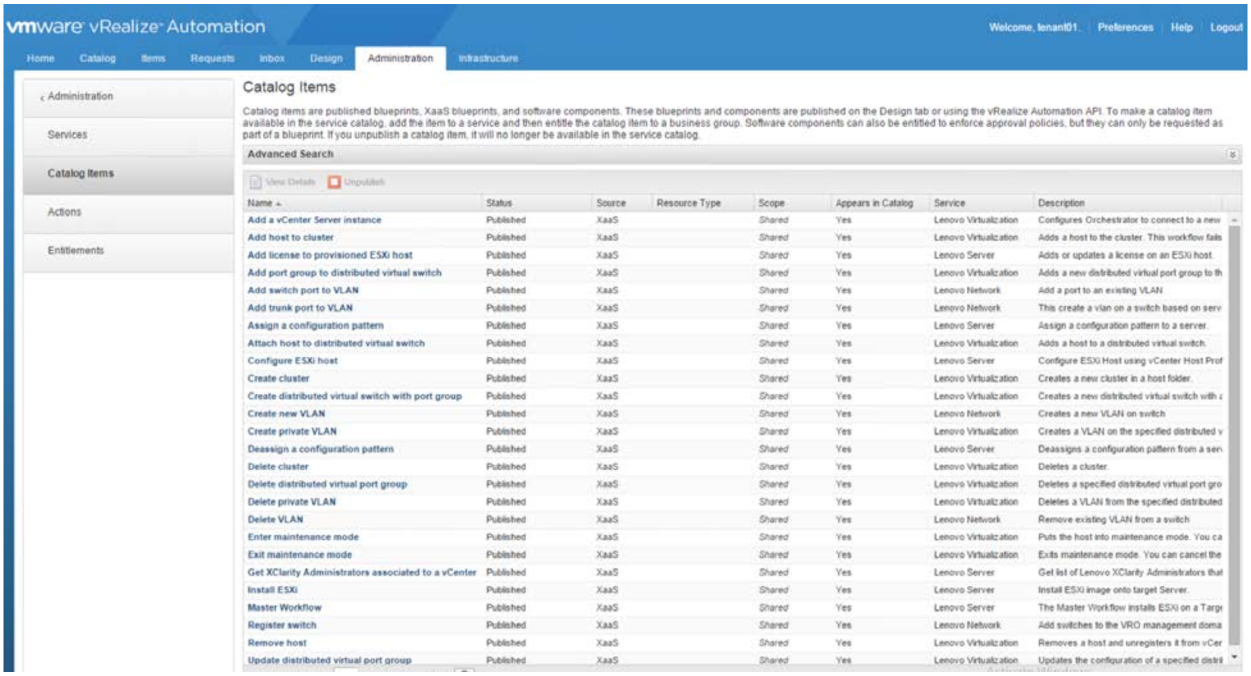
- Lenovo
- Lenovo Softbundle
  - ESXi
    - Add license to provisioned ESXi host
    - Configure ESXi host
    - Get business groups
    - Get vCenter list
    - Install ESXi
  - Master
    - Master Workflow
  - Network
    - Add switch port to VLAN
    - Add trunk port to VLAN
    - Add VLAN to UFP port
    - Create new VLAN
    - Create Single VLAN On LACPVLAG Port
    - Create Single VLAN On VLAG Port
    - Define port name
    - Delete VLAN
    - Discover registered network switches
    - Get server switch port mapping
    - Register switch
  - Server
    - Assign a configuration pattern
    - Deassign a configuration pattern
    - Get available servers
    - Get server MAC addresses
    - Get server patterns
    - Get XClarity Administrators associated to a vCenter
  - Virtualization
    - Add a vCenter Server instance
    - Add host to cluster
    - Add port group to distributed virtual switch
    - Attach host to distributed virtual switch
    - Copy of Attach host to distributed virtual switch
    - Create cluster
    - Create distributed virtual switch with port group
    - Create private VLAN
    - Delete cluster
    - Delete distributed virtual port group
    - Delete private VLAN

Name	Master Workflow
ID	584077dd-deb8-4dbd-aecb-af98bda2ea74
Version	1 . 0 . 4
Workflow icon	
Owner	
User permissions	☑ View contents ☑ Add to package ☑ Edit contents
Server restart behavior	Resume workflow run
Resume from failed behavior	System default
Global tags	
User tags	
Description	The Master Workflow installs ESXi on a Target Server adds the Provisioned ESXi host to a Cluster associated with a Reservation for a business group with in a Tenant.

	Name	Type	Value		Description
□	defaultVRAHost	vCACCAFE:VCACHost	Default [https://vra-701-01.sddc.l...		vCAC Host
□	vcacHost	vCAC:VCACHost	vRA-701-IaaS-01 [https://vRA-701...		vCAC IaaS Host
□	xclarityAdminAtt	LXCI:XClarityAdministrator	Not set		Lenovo XClarity Administrator
□	xclarityServerAtt	LXCI:XClarityServer	Not set		Lenovo XClarity Server
□	isVLAN	string	false		isVLAN
□	task	LXCI:XClarityTask	Not set		XClarity Task

Figure 51: Lenovo SoftBundle workflows for vRealize Orchestrator

Figure 55 shows Lenovo XClarity Integrator catalog items for vRealize Automation.



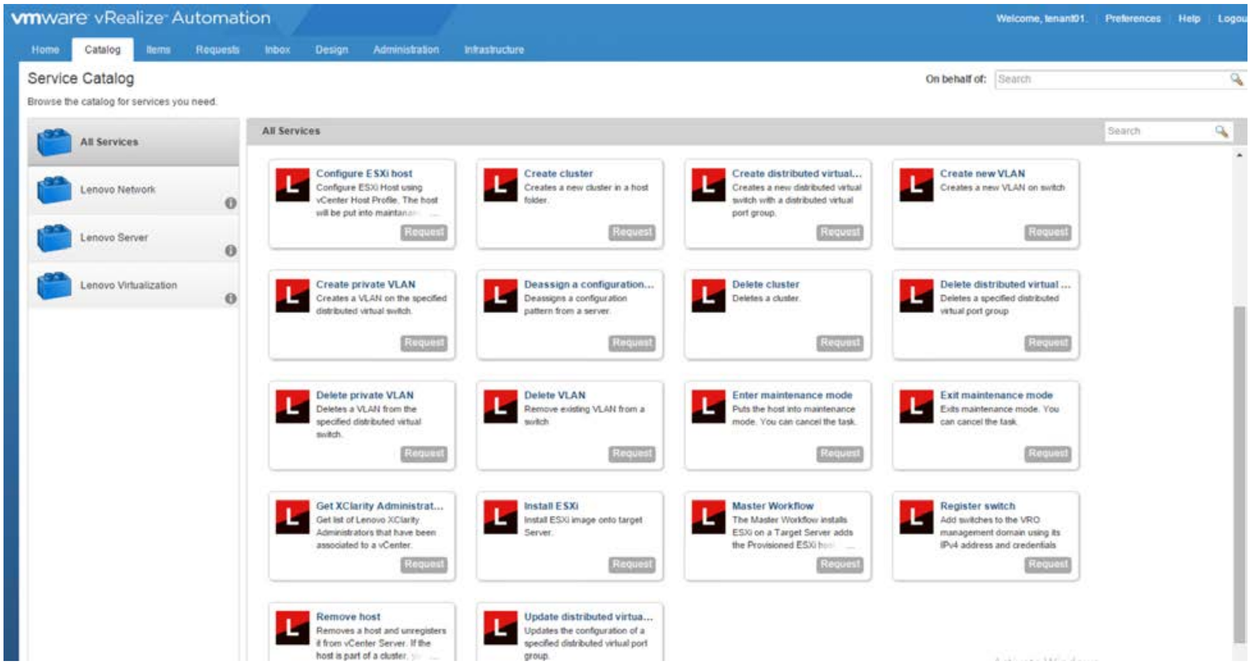
Catalog Items

Catalog items are published blueprints, XaaS blueprints, and software components. These blueprints and components are published on the Design tab or using the vRealize Automation API. To make a catalog item available in the service catalog, add the item to a service and then entitle the catalog item to a business group. Software components can also be entitled to enforce approval policies, but they can only be requested as part of a blueprint. If you unpublish a catalog item, it will no longer be available in the service catalog.

Name ▲	Status	Source	Resource Type	Scope	Appears in Catalog	Service	Description
Add a vCenter Server instance	Published	XaaS		Shared	Yes	Lenovo Virtualization	Configures Orchestrator to connect to a new
Add host to cluster	Published	XaaS		Shared	Yes	Lenovo Virtualization	Adds a host to the cluster. This workflow fails
Add license to provisioned ESXi host	Published	XaaS		Shared	Yes	Lenovo Server	Adds or updates a license on an ESXi host.
Add port group to distributed virtual switch	Published	XaaS		Shared	Yes	Lenovo Virtualization	Adds a new distributed virtual port group to th
Add switch port to VLAN	Published	XaaS		Shared	Yes	Lenovo Network	Add a port to an existing VLAN
Add trunk port to VLAN	Published	XaaS		Shared	Yes	Lenovo Network	This create a vlan on a switch based on serv
Assign a configuration pattern	Published	XaaS		Shared	Yes	Lenovo Server	Assign a configuration pattern to a server.
Attach host to distributed virtual switch	Published	XaaS		Shared	Yes	Lenovo Virtualization	Adds a host to a distributed virtual switch.
Configure ESXi host	Published	XaaS		Shared	Yes	Lenovo Server	Configure ESXi Host using vCenter Host Prof
Create cluster	Published	XaaS		Shared	Yes	Lenovo Virtualization	Creates a new cluster in a host folder.
Create distributed virtual switch with port group	Published	XaaS		Shared	Yes	Lenovo Virtualization	Creates a new distributed virtual switch with a
Create new VLAN	Published	XaaS		Shared	Yes	Lenovo Network	Creates a new VLAN on switch
Create private VLAN	Published	XaaS		Shared	Yes	Lenovo Virtualization	Creates a VLAN on the specified distributed v
Deassign a configuration pattern	Published	XaaS		Shared	Yes	Lenovo Server	Deassigns a configuration pattern from a ser
Delete cluster	Published	XaaS		Shared	Yes	Lenovo Virtualization	Deletes a cluster.
Delete distributed virtual port group	Published	XaaS		Shared	Yes	Lenovo Virtualization	Deletes a specified distributed virtual port gro
Delete private VLAN	Published	XaaS		Shared	Yes	Lenovo Virtualization	Deletes a VLAN from the specified distributed
Delete VLAN	Published	XaaS		Shared	Yes	Lenovo Network	Remove existing VLAN from a switch
Enter maintenance mode	Published	XaaS		Shared	Yes	Lenovo Virtualization	Puts the host into maintenance mode. You ca
Exit maintenance mode	Published	XaaS		Shared	Yes	Lenovo Virtualization	Exits maintenance mode. You can cancel the
Get XClarity Administrators associated to a vCenter	Published	XaaS		Shared	Yes	Lenovo Server	Get list of Lenovo XClarity Administrators that
Install ESXi	Published	XaaS		Shared	Yes	Lenovo Server	Install ESXi image onto target Server.
Master Workflow	Published	XaaS		Shared	Yes	Lenovo Server	The Master Workflow installs ESXi on a Targe
Register switch	Published	XaaS		Shared	Yes	Lenovo Network	Add switches to the VRO management doma
Remove host	Published	XaaS		Shared	Yes	Lenovo Virtualization	Removes a host and unregisters it from vCer
Update distributed virtual port group	Published	XaaS		Shared	Yes	Lenovo Virtualization	Updates the configuration of a specified distri

Figure 52: Lenovo XClarity Integrator for vRealize Automation Catalog Items

Figure 56 shows Lenovo XClarity Integrator services available for vRealize Automation.



Service Catalog

Browse the catalog for services you need.

Configure ESXi host
Configure ESXi Host using vCenter Host Profile. The host will be put into maintanan...

Create cluster
Creates a new cluster in a host folder.

Create distributed virtual...
Creates a new distributed virtual switch with a distributed virtual port group.

Create new VLAN
Creates a new VLAN on switch

Create private VLAN
Creates a VLAN on the specified distributed virtual switch.

Deassign a configuration...
Deassigns a configuration pattern from a server.

Delete cluster
Deletes a cluster.

Delete distributed virtual ...
Deletes a specified distributed virtual port group

Delete private VLAN
Deletes a VLAN from the specified distributed virtual switch.

Delete VLAN
Remove existing VLAN from a switch

Enter maintenance mode
Puts the host into maintenance mode. You can cancel the task.

Exit maintenance mode
Exits maintenance mode. You can cancel the task.

Get XClarity Administrat...
Get list of Lenovo XClarity Administrators that have been associated to a vCenter.

Install ESXi
Install ESXi image onto target Server.

Master Workflow
The Master Workflow installs ESXi on a Target Server adds the Provisioned ESXi ho...

Register switch
Add switches to the VRO management domain using its IPv4 address and credentials

Remove host
Removes a host and unregisters it from vCenter Server. If the host is part of a cluster, y...

Update distributed virtua...
Updates the configuration of a specified distributed virtual port group.

Figure 53: Lenovo XClarity Integrator for vRealize Automation Services Catalog

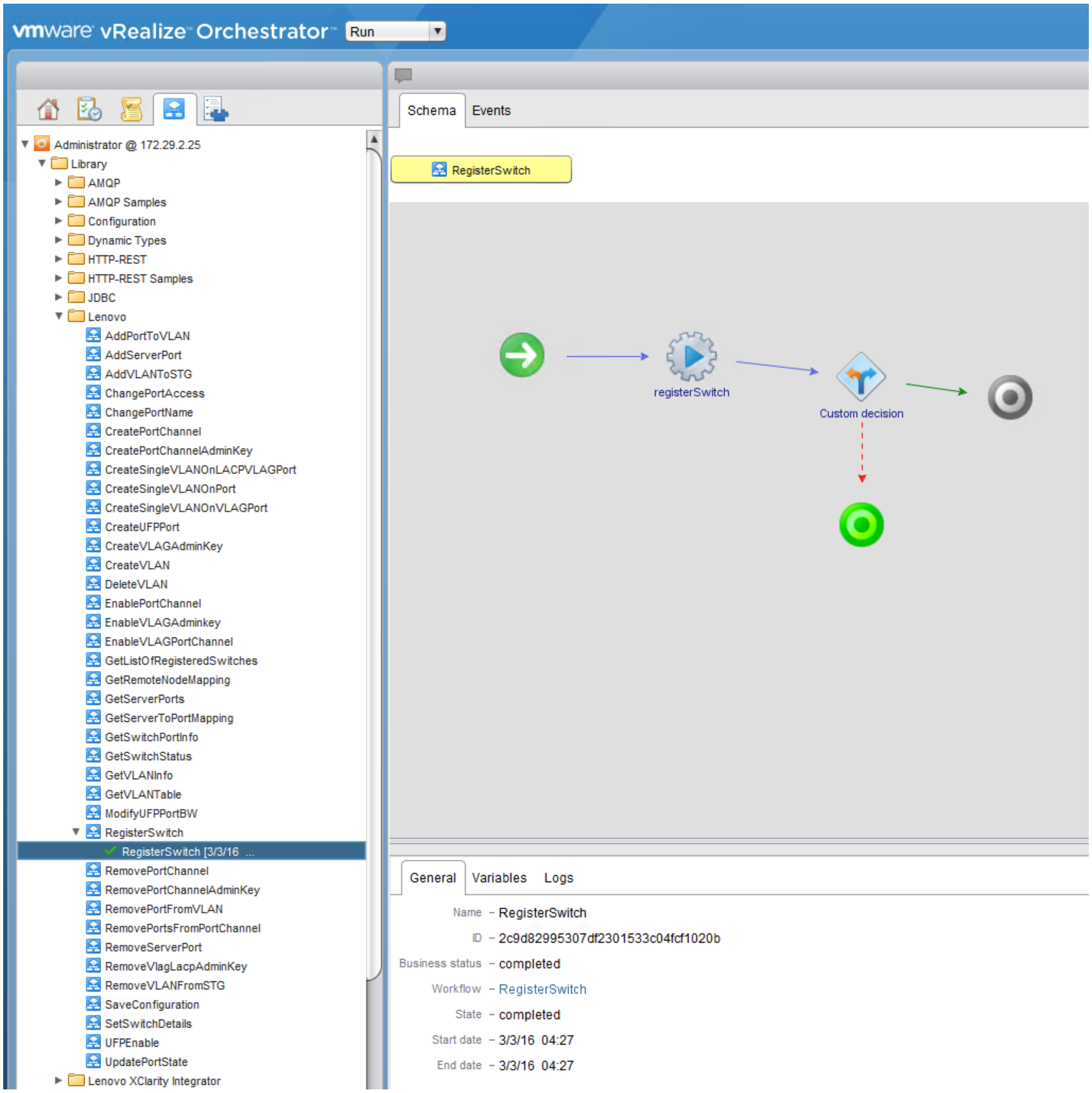
9.6.2 Lenovo network integration plug-ins

Lenovo also provides network integration plug-ins for VMware vRealize Orchestrator and vRealize Log Insight. For more information, see this website:

shop.lenovo.com/us/en/systems/software/systems-management/network-management

The Lenovo networking plug-In for VMware vRealize Orchestrator enables you to:

- Reduce new service delivery time on RackSwitch G8272, Flex System EN4093R, and Flex Systems Interconnect Fabric
- Leverage a comprehensive library of Lenovo Networking workflows and actions capturing network configuration best practices for rapid service orchestration
- Implement more Lenovo switch configuration activity through vRealize Orchestrator and vCenter with less reliance upon native switch interfaces.



Name – RegisterSwitch
ID – 2c9d82995307df2301533c04fcf1020b
Business status – completed
Workflow – RegisterSwitch
State – completed
Start date – 3/3/16 04:27
End date – 3/3/16 04:27

Figure 54: Lenovo networking plug-In for VMware vRealize Orchestrator interface

Lenovo Networking Content Pack for VMware vRealize Log Insight enables you to:

- Increase network reliability by allowing system or network administrators to monitor networks that feature Lenovo branded RackSwitch switches
- Gain access to extremely detailed switch log entries to facilitate deep insights into the network status

- Reduce initial provisioning time of Log Insight by using the 9 prebuilt dashboards and 12 predefined alarms featured in the pack

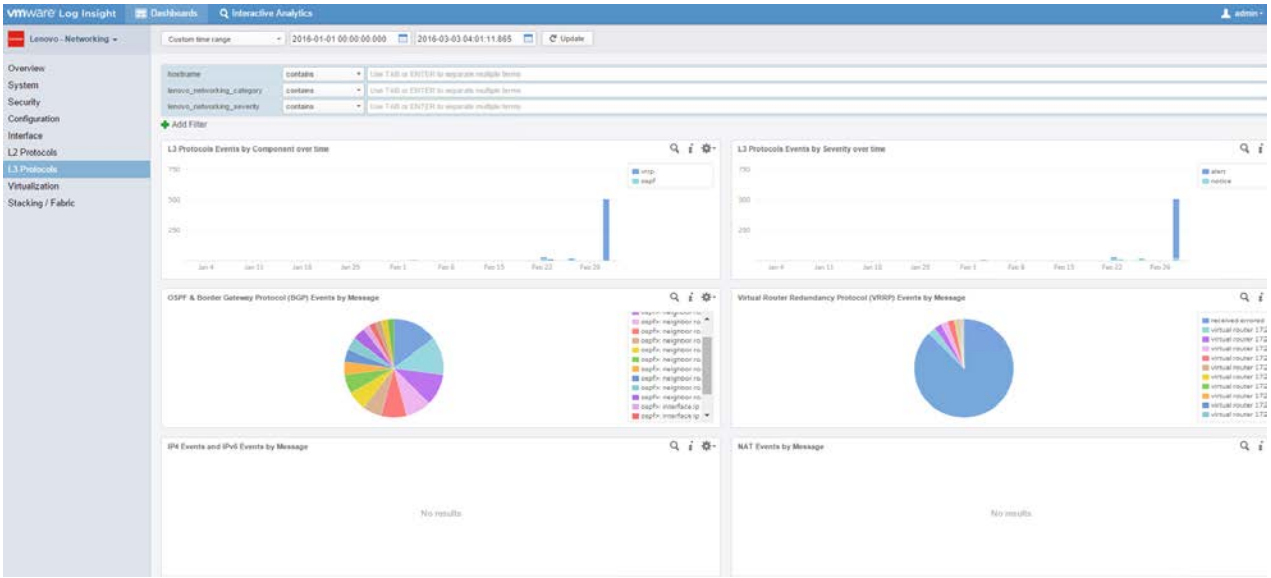


Figure 55: Lenovo Networking Content Pack for VMware vRealize Log Insight interface

9.6.3 Lenovo plug-ins compatibility

Table 54 below lists current versions of Lenovo integration plugins and the required or supported VMware vCenter and vRealize Suite products.

Table 48: Plug-in compatibility

Component Name	Version	Supported Product Versions
Lenovo XClarity Administrator(LXCA)	2.1.0	VMware vCenter 6.0U2/6.5/6.7, ESXi 6.0U2/6.5 U1/6.7
Lenovo XClarity Integrator(LXCI) for vCenter	5.4.0	Lenovo XClarity Administrator 1.4.x, 2.x VMware vCenter 5.x U1/U2/U3, 6.0 U1/U2/U3, 6.5 U1/U2,6.7
Lenovo XClarity Administrator content pack for VMWare vRealize Log Insight	2.0	Lenovo XClarity Administrator 1.1 or higher VMware vRealize Log Insight 3.0 or higher
Lenovo XClarity Integrator for VMware vRealize Automation	1.0.1	VMware vRealize Automation 7.x
Lenovo XClarity Integrator for VMware vRealize Orchestrator	1.1.0	VMware vRealize Automation 7.0 VMware vRealize Orchestrator 6.0/7.0
Lenovo Network Plugin for VMware vRealize Orchestrator	1.4.0	VMware vRealize Orchestrator 7.4.x
Lenovo Networking Content Pack for VMware vRealize Log Insight	1.2	CNOS 10.3, ENOS 8.1, VMware vRealize Log Insight 4.3.x, 4.5.x

9.7 Deployment example

There are 3 strategies to deploy vCloud Suite:

- Use VLANs for both user VMs and the Nutanix CVM
- Use NSX and VXLANs for user VMs and the Nutanix CVM in a VLAN
- Use NSX and VXLANs for both user VMs and the Nutanix CVM

This section describes an example deployment of vRealize Suite 7.2 using VLANs for all VMs. The term vRA is used as a short-hand to denote all of the vRealize Automation VMs.

The Lenovo ESXi 6.5 U1 image is used for this example deployment. To use ESXi 6.5 U1, download the Lenovo ThinkSystem custom image from the following website: my.vmware.com/web/vmware/info/slug/datacenter_cloud_infrastructure/vmware_vsphere/6_5#custom_iso.

9.7.1 Physical Model

Four physical servers are used for each of the shared edge and compute, management, and additional compute clusters. Lenovo ThinkSystem HX3320 appliances are used for the shared edge and compute cluster and management cluster. The servers for the additional compute cluster are workload- dependent and Lenovo ThinkSystem HX3520-G appliances are used in this case. Lenovo RackSwitch G8272 switches are used for the 10 GbE network. See also the “VMware vCloud Suite deployment model” on page 23.

Figure 59 shows a view of the physical 10 GbE network and connections to the external internet.

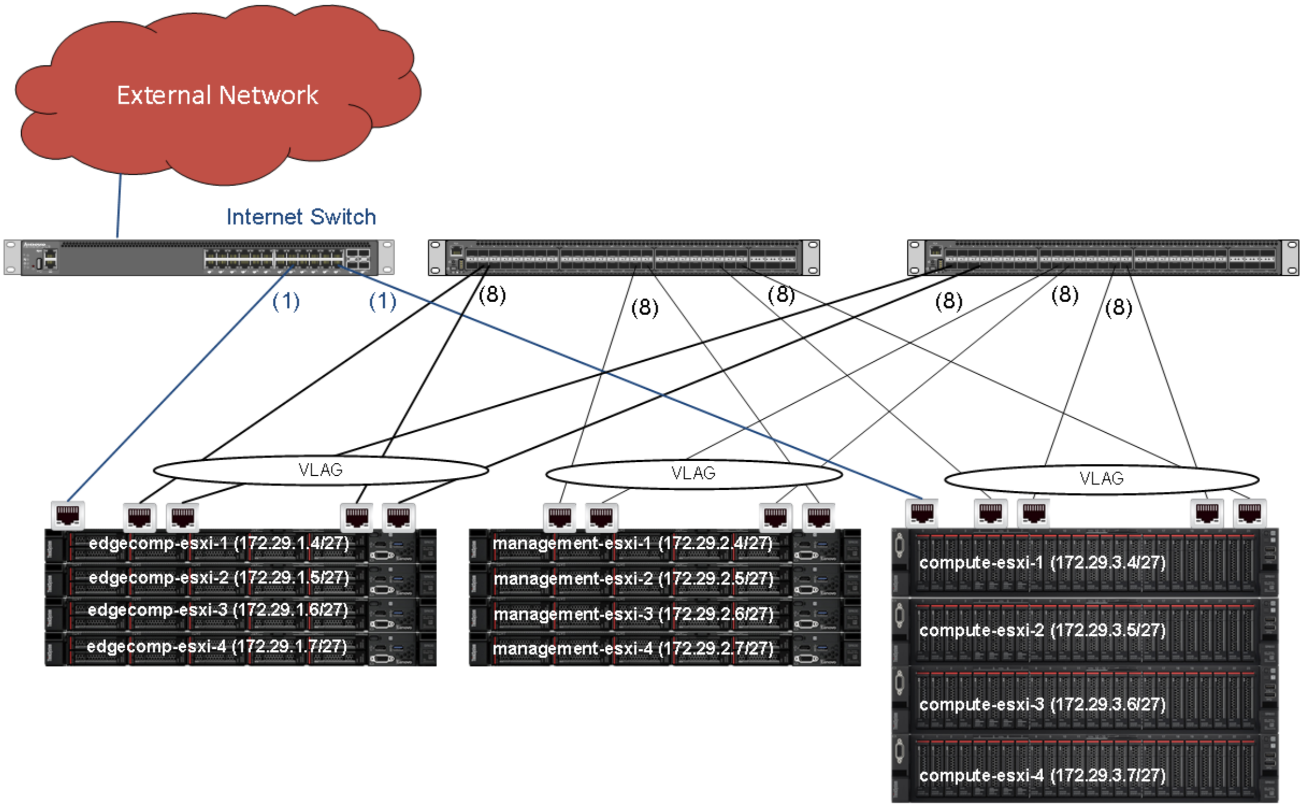

External Network
Internet Switch
(1)
(1)
(8)
(8)
(8)
(8)
(8)
(8)
VLAG
VLAG
VLAG
edgecomp-esxi-1 (172.29.1.4/27)
edgecomp-esxi-2 (172.29.1.5/27)
edgecomp-esxi-3 (172.29.1.6/27)
edgecomp-esxi-4 (172.29.1.7/27)
management-esxi-1 (172.29.2.4/27)
management-esxi-2 (172.29.2.5/27)
management-esxi-3 (172.29.2.6/27)
management-esxi-4 (172.29.2.7/27)
compute-esxi-1 (172.29.3.4/27)
compute-esxi-2 (172.29.3.5/27)
compute-esxi-3 (172.29.3.6/27)
compute-esxi-4 (172.29.3.7/27)

Figure 56: Networking Overview

For the shared edge and compute, management and additional compute clusters, the nodes use VLAG technology and as such are using a LAG configuration within the vSphere Distributed Switches. It is recommended to use VLAG for all the clusters connected to the same set of switches.

The servers in the shared edge and compute cluster and the additional compute cluster are connected to a 1G switch. This switch in turn is connected to the internet via a gateway and firewall (not shown).

9.7.2 IP/VLAN mapping

It is recommended to use different VLANs in each of the clusters (shared edge and compute, management, and compute). These different VLANs can be connected together by leveraging Layer 3 routing either in the G8272 physical switch or in upstream switches.

This example deployment uses the following five VLANs:

- Management (includes Nutanix CVM)
- vMotion
- FT
- vRA
- Comp (for customer specific computation data traffic)

It is a Nutanix best practice to include the CVM as part of the management VLAN. Alternatively the Nutanix CVM could be executed in its own Storage VLAN. Local storage traffic between VMs and the local CVM is done with a Nutanix specific vSwitch (vSwitchNutanix) using the svm-iscsi-pg port group. Do not modify vSwitchNutanix, as this provides critical CVM-to-hypervisor storage communication on CVM eth1.

Table 55 lists example IP address ranges for the VLANs in each cluster where RID means Rack ID.

Table 49: Network Segments

Cluster	**Shared Edge and Compute (RID 1)**		**Management (RID 2)**		**Compute (RID 3)**	
	Subnet	**VLAN**	**Subnet**	**VLAN**	**Subnet**	**VLAN**
Manage	172.29.1.0/27	101	172.29.2.0/27	201	172.29.3.0/27	301
vMotion	172.29.1.32/27	102	172.29.2.32/27	202	172.29.3.32/27	302
FT	172.29.1.64/27	103	172.29.2.64/27	203	172.29.3.64/27	303
vRA	N/A		172.29.2.192/27	207	N/A	
Comp	172.29.1.192/27	109	N/A		172.29.1.192/27	109

In this example, each cluster needs a minimum of five network segments within the 172.29.RID.x address range. Each segment does not require more than 30 IP addresses; therefore, a 255.255.255.224 (/27) netmask provides enough addresses. The same VLAN IDs can be used across racks with different IP segments as shown by VLAN 109.

9.7.3 Layer 3 Routing with VRRP

Virtual Routing Redundancy Protocol (VRRP) should be enabled in the switches for layer 3 routing. Each switch defaults to IP routing. VMs can use the respective routing IP to reach the switches. Routing occurs through either one of the switches and this causes intermittent routing failures when used with VLAG.

Layer 3 routing with VRRP removes this limitation by using a single routing IP address for both the switches. Each subnet requires 2 IPs reserved for normal layer 3 routers and one IP reserved for VRRP router.

Table 56 lists the example layer 3 routing IPs for each of the VLANs and the clusters.

Table 50: Layer 3 Example Routing

Cluster	**VLAN**	**Subnet**	**L3 Interface IP**		**VRRP Router IP**	
			G8272(1)	G8272(2)	G8272(1) Master	G8272(2) Backup
Shared Edge and Compute	Management	172.29.1.0/27	172.29.1.1	172.29.1.2	172.29.1.3	172.29.1.3
	vMotion	172.29.1.32/27	172.29.1.33	172.29.1.34	172.29.1.35	172.291.35
	FT	172.29.1.64/27	172.29.1.65	172.29.1.66	172.29.1.67	172.29.1.67
	Comp	172.29.1.192/27	172.29.1.193	172.29.1.194	172.29.1.195	172.29.1.195
Management	Management	172.29.2.0/27	172.29.2.1	172.29.2.2	172.29.2.3	172.29.2.3
	vMotion	172.29.2.32/27	172.29.2.33	172.29.2.34	172.29.2.35	172.29.2.35
	FT	172.29.2.64/27	172.29.2.65	172.29.2.66	172.29.2.67	172.29.2.67
	vRA	172.29.2.192/27	172.29.2.193	172.29.2.194	172.29.2.195	172.29.2.195
Additional Compute	Management	172.29.3.0/27	172.29.3.1	172.29.3.2	172.29.3.3	172.29.3.3
	vMotion	172.29.3.32/27	172.29.3.33	172.29.3.34	172.29.3.35	172.29.3.35
	FT	172.29.3.64/27	172.29.3.65	172.29.3.66	172.29.3.67	172.29.3.67
	Comp	172.29.1.192/27	172.29.1.193	172.29.1.194	172.29.1.195	172.29.1.195

9.7.4 Management cluster

Although a single distributed switch can be deployed across various clusters, it is recommended to deploy unique distributed switches per cluster. Figure 60 shows the distributed switch for the management cluster.

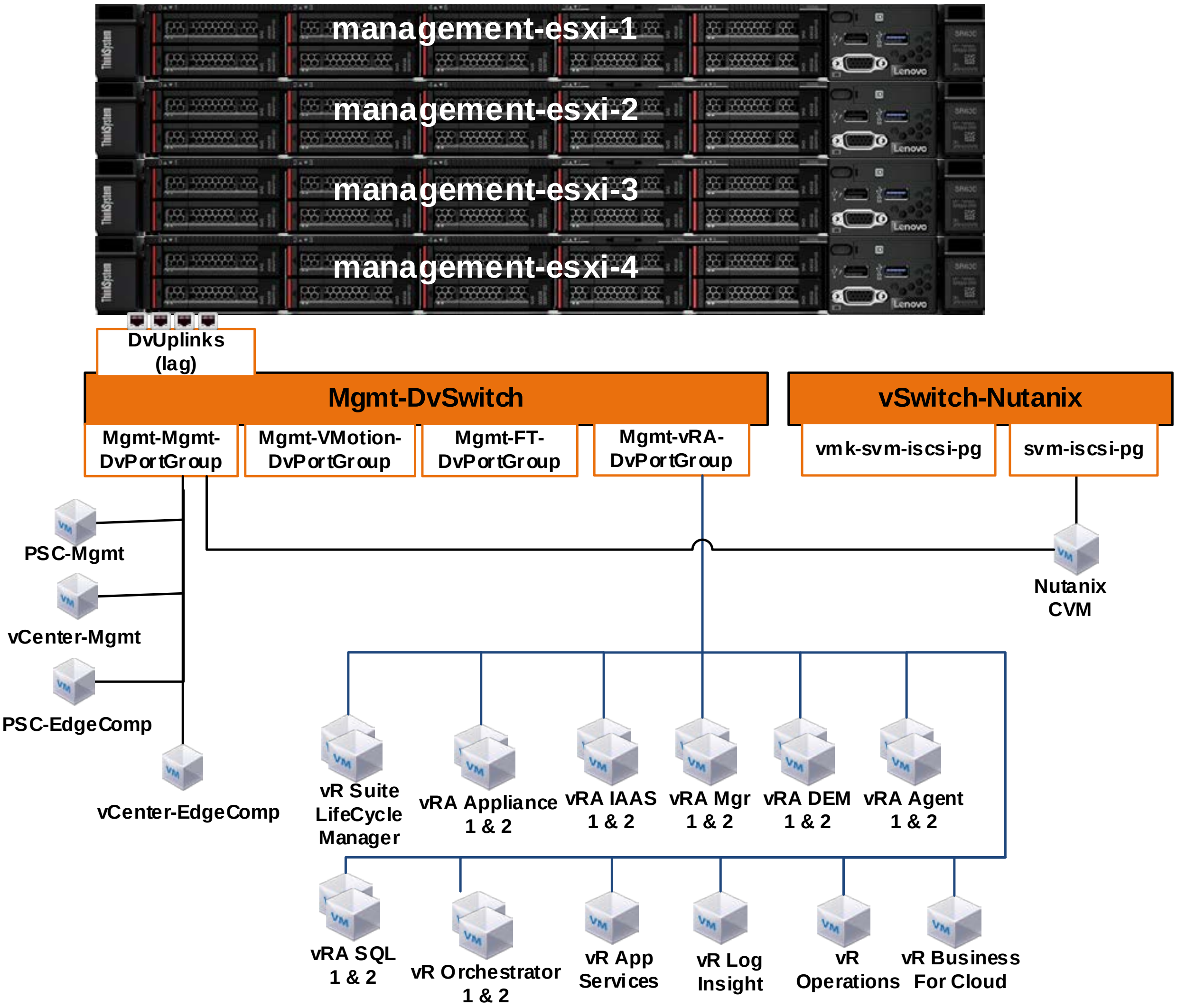


Figure 57: Management Cluster VDS

The infrastructure uses single SSO domain and all platform service controllers and vCenter servers are connected to this domain. All VMs in the management cluster can be configured on the same network segment. Separate vCenter and platform service controller instances are used for management cluster and shared edge and compute cluster. It is recommended that vCenter is deployed in a highly available configuration (depending on each customer's needs) which can result in multiple vCenters, multiple PSCs, and physical load balancers.

9.7.5 Shared Edge and Compute cluster

Figure 61 shows the distributed switches for the edge and compute cluster. One DVS is used for accessing the Internet via a 1GbE port on the server. The other DVS is used for all of the edge and compute network flows.

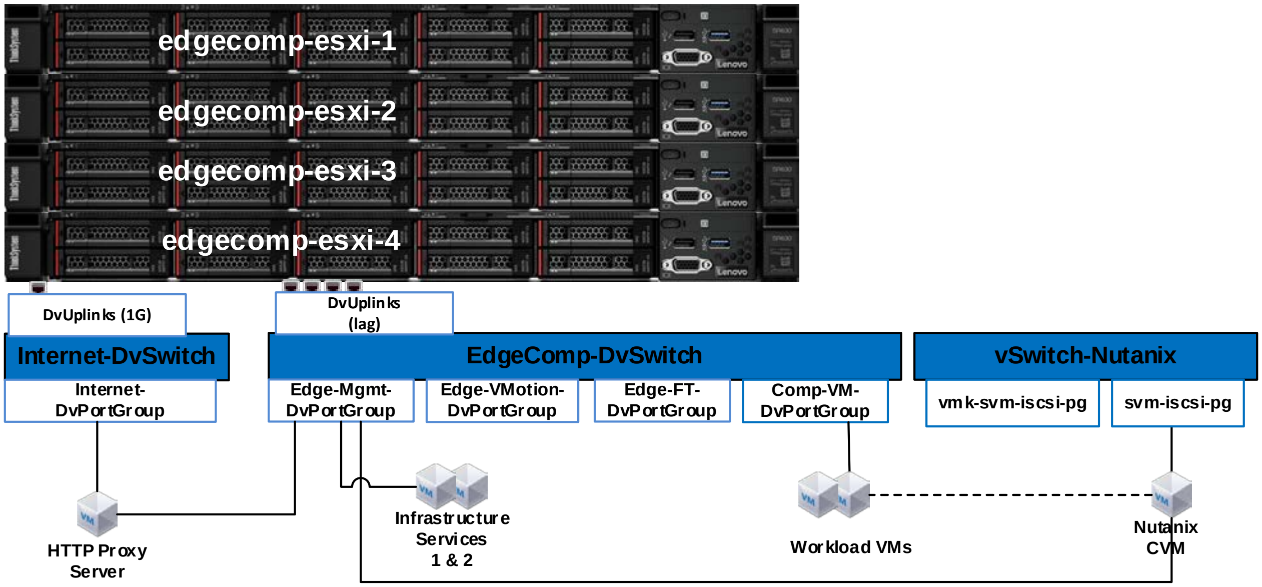


Figure 58: Shared Edge and Compute Cluster VDS

Infrastructure service VMs, such as Active Directory, DHCP, DNS, and NTP might exist in the customers' environment and these services are accessed through a clustered configuration, as shown in this example. However, if there is a requirement to virtualize these services within this environment, then they can be accessed through Edge-Mgmt-DvPortGroup, as shown in Figure 61.

The Comp-VM-DvPortGroup is used for the workload VMs.

9.7.6 Compute cluster

Figure 62 shows the distributed switch for the compute cluster. Workload-specific DvPortGroups and VMs are not shown because they are highly dependent on the specific workloads. More than one compute cluster can be used to add more workloads.

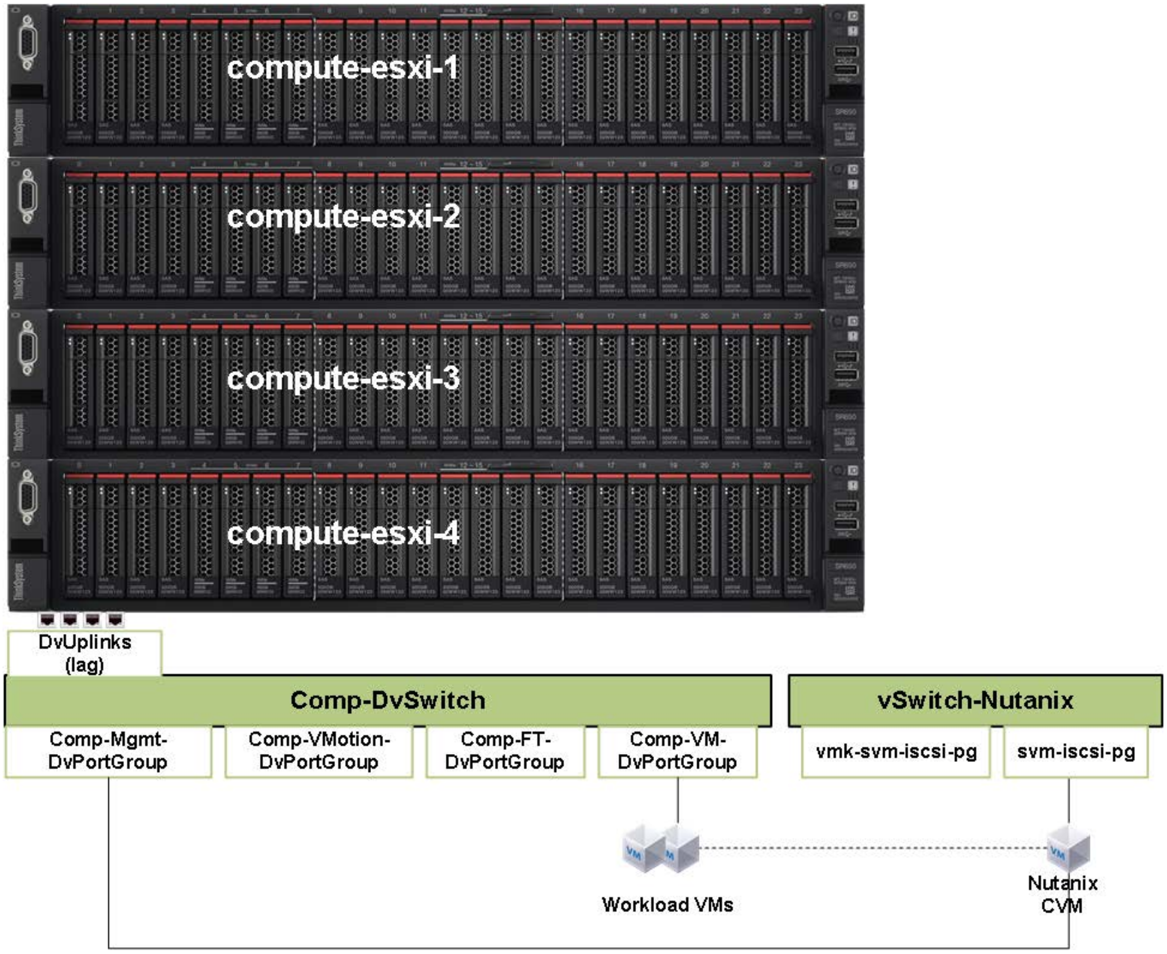


Figure 59: Compute Cluster VDS

Resources

- Nutanix Portal (requires registration)

  portal.nutanix.com

- Nutanix Bible

  nutanixbible.com/

- Nutanix Tech Note: VMware vSphere Networking on Nutanix

  go.nutanix.com/rs/nutanix/images/Nutanix_TechNote-VMware_vSphere_Networking_with_Nutanix.pdf

- VMware vSphere

  vmware.com/products/datacenter-virtualization/vsphere

- SAP Support Portal (requires registration)

  support.sap.com

Document History

Version 1.0	31 October 2017	• Initial version
Version 1.1	15 November 2017	• Added chapter for Microsoft SQL Server workload • Added rack deployment models with more detail on ThinkAgile SXN
Version 1.2	10 December 2017	• Added chapter for Microsoft Exchange 2013 workload • Added deployment model for Microsoft Exchange
Version 1.3	22 January 2018	• Added Nutanix AHV hypervisor and Citrix Workspace Appliance (CWA) for Citrix XenDesktop • Added Microsoft Exchange 2016 • Added ThinkSystem NE1072T switch • Added ThinkSystem NE2572 switch
Version 1.4	29 January 2018	• Added chapter for VMware vCloud Suite • Added deployment model for VMware vCloud Suite • Added performance results for VDI graphics acceleration • Add Microsoft Exchange 2016 results for all flash appliances
Version 1.5	18 August 2018	• Added certified nodes • Added pre-configured deployment ready solutions for Citrix XenDesktop and VMware Horizon • Removed ThinkAgile SXN integrated solution
Version 1.6	2 October 2018	• Added HX782x appliances and certified nodes • Added chapter for SAP Business Applications
Version 1.7	12 December 2018	• Added chapter for Red Hat OpenShift Container Platform
Version 1.8	23 April 2019	• Updated chapter on Microsoft SQL Server to use SQL Server 2019 and Intel Xeon Scalable Processor gen 2 CPUs.
Version 1.9	10 June 2019	• Added Citrix XenDesktop results for Xeon SP2 processors • Added VMware Horizon results for Xeon SP2 processors
Version 2.0	16 July 2019	• Updated performance data for VDI Power Worker profile from the Login VSI benchmark
Version 2.1	25 September 2019	• Updated chapter SAP HANA to include HX332x and HX752x models based on Intel Xeon SP 1.Gen processors • Added HX7820 SSD Only descriptions to SAP HANA section • Removed requirement for Mellanox ConnectX Adapters with SAP HANA

Version 2.2	15 November 2019	• Added Citrix Workspace appliance • Added SAP HANA models for Intel Xeon SP Gen 2. Processors
Version 2.3	10 December 2020	• Moved the SAP Solutions chapter to a separate Reference Architecture document - lenovopress.com/lp1413

Availability*: Offers, prices, specifications and availability may change without notice. Lenovo is not responsible for photographic or typographical errors.* ***Warranty:*** *For a copy of applicable warranties, write to: Lenovo Warranty Information, 1009 Think Place, Morrisville, NC, 27560, Lenovo makes no representation or warranty regarding third party products or services.* ***Trademarks:*** *Lenovo, the Lenovo logo, ThinkSystem, System x, ThinkServer are trademarks or registered trademarks of Lenovo. Microsoft and Windows are registered trademarks of Microsoft Corporation. Intel, the Intel logo, Xeon and Xeon Inside are registered trademarks of Intel Corporation in the U.S. and other countries. Other company, product, and service names may be trademarks or service marks of others.*

12-2020