EMC²

FUNCTIONAL DEPLOYMENT ARCHITECTURES FOR VMWARE-BASED HADOOP DEPLOYMENTS WITH ISILON ONEFS STORAGE

How to deploy and manage a VMware compute cluster with Isilon OneFS storage

ABSTRACT

This white paper outlines the benefits and a high-level approach to deploying virtualized Hadoop clusters with Isilon OneFS storage.

January 2018

Publication History

Version	Date	Description
1.00	3 January 2018	Initial version.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This document is a high-level design document that highlights the best practices for deploying and managing virtual Hadoop compute architectures with a shared Isilon OneFS storage cluster. The whitepaper will provide guidance and functional reference architecture to deploy a VMWare VSphere-based virtual compute platform to leverage the decoupled storage capabilities of using Isilon OneFS as the HDFS storage component.

Audience

This guide is intended for VMWare administrators, Hadoop systems administrators, storage administrators, IT architects, and IT managers who will be running OneFS with virtual Hadoop.

OVERVIEW

The ability to decouple the storage requirements from the compute and DataNode functionality opens up the ability to virtualize the Hadoop platform, leveraging only the Virtual Machines (VMs) to provide all the required compute resources. When a DataNode storage requirement existed, it made no sense to deploy DataNodes on VMs from a provisioning or a storage perspective. The ability to utilize OneFS as the HDFS storage layer changes the deployment model dramatically, increasing the options and flexibility of your Hadoop clusters significantly. It should be noted that virtual compute Hadoop-based clusters will likely have performance limitations when compared to hardware-based compute infrastructure. However, the flexibility and ease of management may provide many additional benefits outside of just performance that make a virtualized cluster extremely appealing under many situations.

Why Virtualize Compute

Many reasons exist as to why utilizing a virtualized computing platform would be appropriate when we can decouple the data and storage dependencies from the virtual machine.

- Ease of deployment and management of virtual machines and the existence of current virtualization environment
- The use case is appropriate to the performance available from virtual machines; Sandbox, Dev, QA, and so on
- Flexibility of virtual machines availability; grow and shrink footprint as needed
- Physical limitations prevent deployment of Hadoop; servers, storage, racks, power, and network availability is limited

Why Isilon

Isilon is implemented as a node-based clustered storage appliance running a combined operating and file system known as OneFS. It was originally developed to solve storage challenges at scale, and customers have relied on its enterprise features for over a decade. The unique capabilities and integration of support for the HDFS protocol have made it an ideal platform for scale-out Hadoop data storage. The ability to provide all the HDFS storage requirements to a compute-only based clusters clearly removes the dependency of Direct Attached Storage (DAS) to servers and allows the compute resources to become virtualized.

As Isilon OneFS assumes the role of both the NameNode and all DataNodes in the Hadoop HDFS cluster, this provides a number of benefits as follows:

- Since all nodes in an Isilon OneFS cluster function as a NameNode, no secondary NameNode or High Availability (HA) is required; any configured OneFS nodes will function as a NameNode and will provide continuous data access without complex HA configuration. This substantially increases operational availability and reduces administrative overhead.
- Host nodes in the compute cluster no longer function as DataNodes. The roles that remain on these nodes are now Node Manager Roles and other dedicated master roles for services, for example, Hive, HBase, Spark, and Impala, as quantity and distribution are dictated. This removal of data and data management allows significantly more freedom around compute server selection and the required quantity. This alone can provide substantial consolidation of the physical space and power draw requirements of a given configuration.

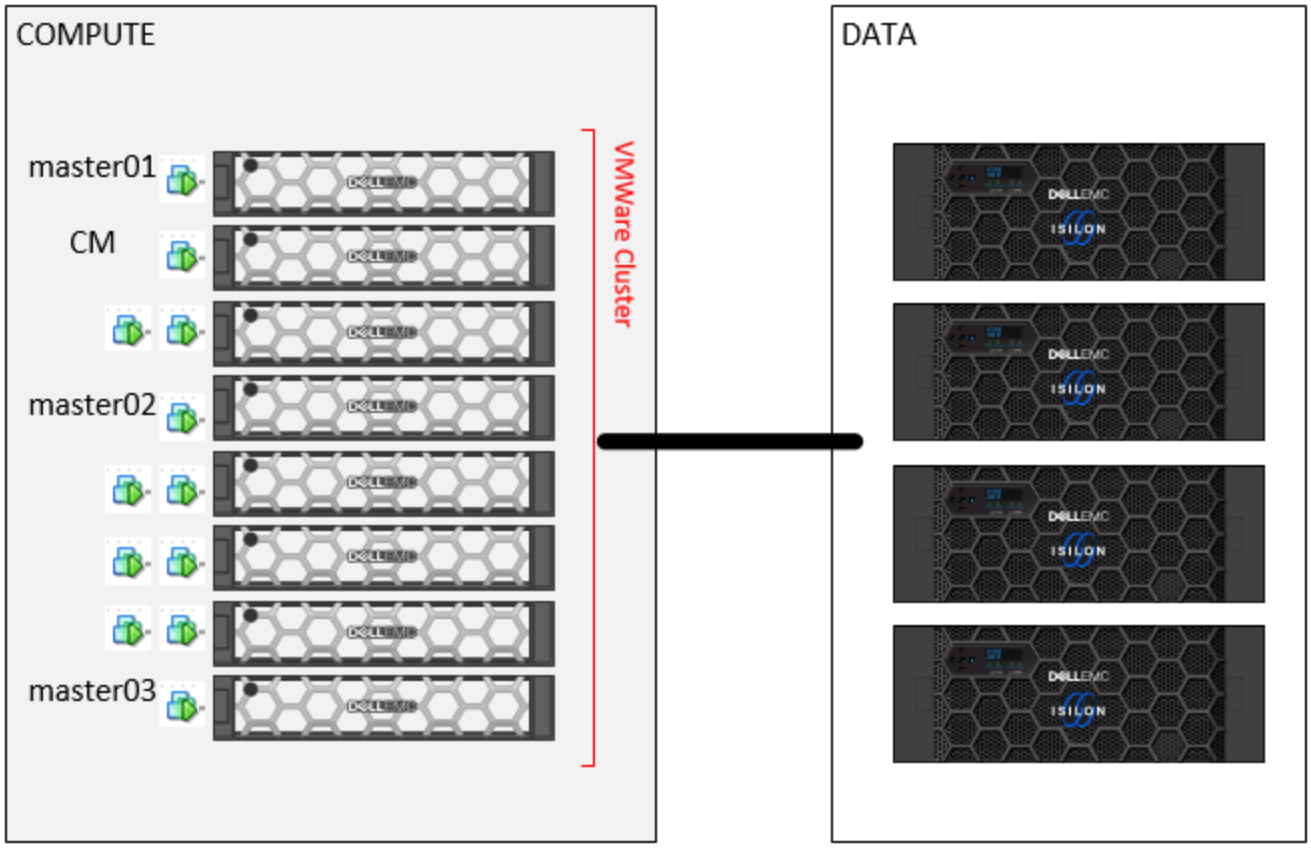

Figure 1. Logical Virtualized Hadoop Cluster and Isilon Data Storage

VMWare vSphere Considerations

Even with the removal of the HDFS DataNode storage requirement from the VMware environment, careful planning and design decisions must be made with regard to the architecture and functional considerations of the vSphere clusters storage and networking design.

PHYSICAL ESXI AND ISILON CLUSTER TOPOLOGIES

The ESXi and vSphere cluster should be implemented to provide standard High Availability capabilities to the VMs it hosts. The physical topology shown in Figure 2. provides a high-level overview of the physical topologies that should be looked at when deploying the ESXi hosts and the Isilon cluster. In general ESXi hosts and Isilon nodes should be deployed to provide as much bandwidth and low latency connectivity between the hosts.

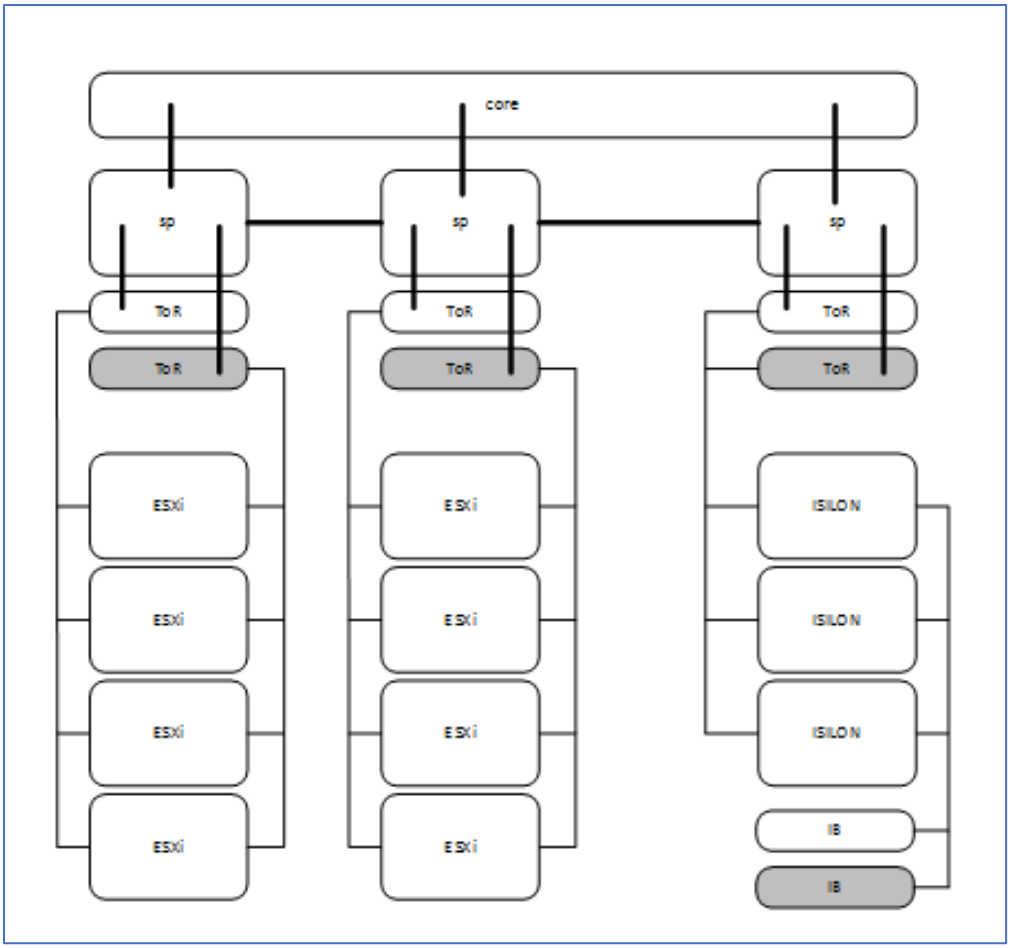

Figure 2. Physical Topology

With the ability to use compute only VMs, the ESXi vSphere clusters can be an existing multitenant VMWare clusters or dedicated host cluster only hosting and managing the Hadoop VMs. The type and configuration of the ESXi Servers, Storage, and Network will clearly impact performance obtainable from the Hadoop compute cluster and should be managed appropriately. Some key components to focus on should be as follows:

ESXi Servers

Multiple CPU servers with large amounts of RAM should be used for hosting VMware ESXi. The number of servers in the vSphere cluster will depend on the size and number of VMs to be deployed in the environment.

Storage

ESXi itself requires very little disk space and can be installed on a small local disk, even an onboard SSD of some type. Since the architecture of the compute cluster does not require DataNodes, the majority of the compute Hadoop VMs require relatively small disks; enough for an OS installation and logging. The master nodes in the Hadoop cluster will require more, and we will discuss this later in the document. Also, the compute-only nodes will require some additional disk to act as temporary disk space for 'Shuffle'. This local Shuffle space that is allocated to compute VMs can reside in a few places:

- Local disk within the server the VM resides on; this will require pinning of the VM to this server, use dedicated SSD
- An allocated VMDK on SAN, VSAN or ScaleIO attached to the compute VM
- An NFS mount point on Isilon

The type of storage that is used will affect the performance of the Hadoop cluster, but in many of these scenarios, the primary purpose of a VM-based Hadoop cluster should not be considered a High-Performance Compute cluster.

Networking

At a minimum, the ESXi and Isilon networking should be 10 Gbps from end to end, the port density of the switches should be enough to accommodate all the NICs the servers and Isilon have available to maximize throughput. We also recommend that you have enough uplinks between switches to provide realistic Inter-Switch Links (ISLs) to support the infrastructure deployed. It is also recommended to isolate Hadoop data traffic to its own network to maximize throughput and minimize latency between compute and storage. Additional vSphere best practices for networking should be implemented per VMware Best Practice; some will be addressed later in the document but many are beyond the scope of it.

LOGICAL VIRTUAL CLUSTER TOPOLOGY

VMs Allocated

The primary advantage of using virtual machines for the Hadoop cluster is the ability to easily allocate—then grow or shrink the cluster—without the traditional restrictions on servers, rack space, power, and networking. Since no HDFS data protection is required, the compute nodes can be added and removed without having to reallocate or manage storage or the data protection overhead. Based on the Hadoop services to be deployed, a number of Master Role nodes will be required to act as the primary masters for those roles, while the number of compute nodes can be deployed and scaled outside of these Master nodes.

We recommend that you have the following nodes deployed:

- Minimum of three master nodes VMs
- Hadoop Cluster Management Server – Ambari or Cloudera Manager VM
- Number of compute nodes will depend on the size and performance requirements of the cluster

VM Placement

In a VMotion-enabled vSphere clusters, create strong negative affinity rules within the Distributed Resource Scheduler (DRS) to keep Master nodes on separate ESXi hosts to limit node unavailability from host issues. If no VMotion is in use, manually place Master node VMs on different physical hosts.

The placement of compute nodes, VMotion automation and thresholds should be set to maximize resource availability within the vSphere cluster. Additional DRS rules should be added to minimize the grouping of larger amounts of compute node on the same host depending on the size of the cluster.

Management VM

Create a dedicated management VM to host the cluster management and administration tools:

Cloudera Manager for Cloudera CDH Deployments

Role	Cloudera Manager	Management Services	Cloudera Agent	Navigator
Admin Role	**X**	**ALL**	**X**	**X**

Table 1. Cloudera Manager Virtual Machine deployed roles

Ambari Server for HDP Deployments

Role	Ambari Server	Ambari Agent		
Admin Role	**X**	**X**		

Table 2. Ambari Server Virtual Machine deployed roles

Role Allocation and VMs

Since we are deploying the cluster without any HDFS DataNodes, the Master nodes do not require any NameNode or Secondary NameNode services.

The following table outlines the role assignment to each node in the cluster. This represents a minimum configuration if a large number of compute node VMs is to be deployed, or additional requirements are to be met, then additional Master nodes should be deployed and leveraged.

Master Node VM's

Role	YARN	ZOOKEEPER	OOZIE	HIVE	IMPALA	HBASE	AGENT	HUE
Master1	Resource Manager	X				Master	X	
Master2		X	X			Master	X	X
Master3	History Server	X		MetaStore HiveServer2 WebHCat	StateStore Catalog	Master	X	

Table 3. Master Roles Virtual Machine deployments

No NameNode or Secondary NameNode services are required as Isilon will provide NameNode and DataNode services. Additional Master services may not be listed here depending on services deployed. For additional information on deploying Hadoop cluster services with Isilon, review the Isilon and Hadoop Deployment Guides and the specific installation guides for each Hadoop distribution.

Compute Node VM's

Role	AGENT	NODE MANAGER	HBase Region Sever	Impala Daemon
Compute Node 1 – (X)	Cloudera or Ambari Agent	X	X	X

Table 4. Compute Node Roles Virtual Machine deployments

No DataNode service are required, All HDFS data storage is Isilon-based.

The following illustrates an example of service deployment and distribution using a Cloudera Manager based deployment of Cloudera CDH on virtual machines.

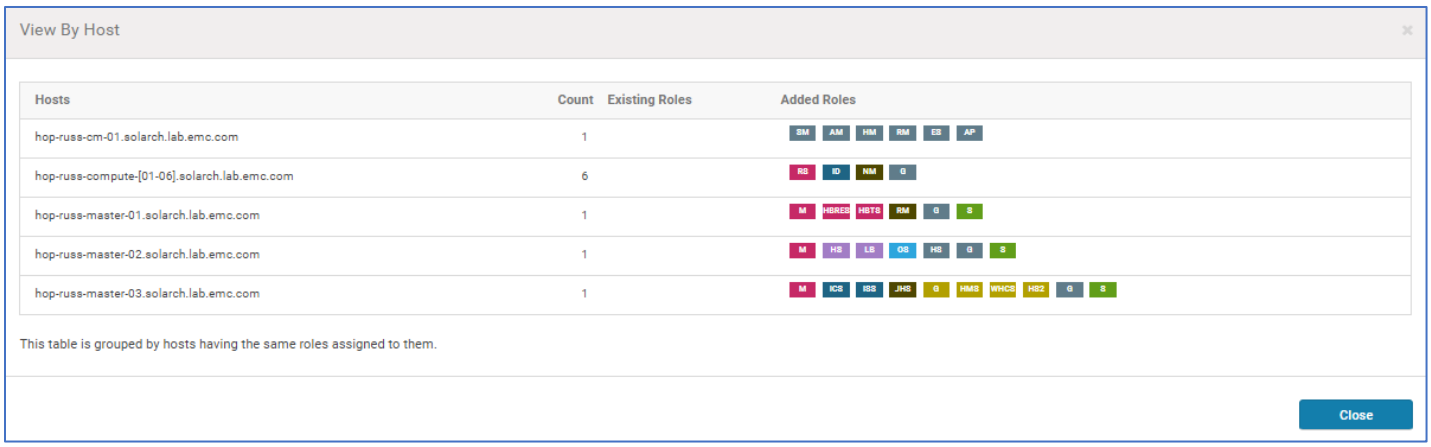

Figure 3. Cloudera Manager role distribution and allocation

Supportability/Compatibility Matrix

Virtualized deployments have no specific restriction compared to physical deployments. You should consult the Products Supported by OneFS compatibility matrix prior to deployment.

Environment Sizing and Platform Tuning Considerations

Outside of the already discussed base deployment and Master node roles, the number of deployed compute nodes will depend on a few factors:

- Available ESXi Host resource availability
- Isilon cluster configuration; Node type and quantity
- Size of the compute-only nodes to be deployed
- Expected performance profile: Sandbox, QA, Development, and test or lower tier compute profile

A general baseline of deployed compute VMs is a 2:1 ratio of compute node VMs to Isilon Node (can also support up to 3:1), for increased high IO a ratio of 1:1.5 for compute VMs to Isilon nodes may be more appropriate. In general lab testing, 2 VMs per Compute Node for Optimal Performance is seen.

VMWARE VSPHERE DEPLOYMENT CONFIGURATIONS

Virtual Machines

The configuration of individual virtual machines within a virtualized cluster will depend heavily on the resources available from the underlying ESXi hosts or cluster. All virtual machines need to meet the minimum requirements in order to deploy the services they are hosting. Some general guidelines are as follows.

Master Node VM:

- 32 GB Memory
- 4 vCPU Cores
- 128 GB Disk Space – OS VMDK Disk on shared storage VSAN, SIO or SAN

Compute Node VM:

- 16 GB Memory - minimum (allocate as much memory as possible to compute VM's, this will impact performance of the cluster)
- 2 vCPU Cores
- 64 GB Disk Space – OS Disk
- Shuffle Disk Space – Locally provisioned VMDK or Raw Device Mapping disk on local disk, VSAN, SIO, SAN or NFS based Isilon mount point

Management VM: Cloudera Manager or Ambari Server

- 2vCPU
- 4GB RAM
- 500GB Disk Space – OS VMDK Disk on shared storage VSAN, SIO or SAN

NETWORKING

VMware VSwitches should be deployed per VMWare's best practices to optimize network traffic from within the ESXi cluster and to the Isilon. The specific configuration of VSwitches is beyond the scope of this document but some general Best Practices should be followed if possible:

- Isolate Hadoop to Isilon traffic to maximize latency
- MTU size should be consistent end-to-end explicitly, use MTU 9000 if possible

STORAGE

Storage requirements fall into two main categories

1. **Host VM VMDK** – Host OS disks should be provisioned per VMware best practice and balanced across VMware datastores.

- VMDK Virtual disks created in "independent persistent" mode for optimal performance
- Eager Zeroed Thick virtual disks provide the best performance
- Partition alignment at the VMFS layer depends on the storage vendor. Misaligned storage can impact performance
- Disable SIOC
- Disable storage DRS

2. Temporary Shuffle Space

The amount of temporary shuffle space required will depend on the amount of HDFS storage to be utilized. As a rough guide, approximately 20% of the total used HDFS equivalent storage will be required. If 100TB of storage will be used on Isilon, then 20TB of shuffle space will be required and spread evenly across the compute nodes. Two primary options exist for provisioning this storage to each compute node VM.

- Local disk/SAN/ScaleIO disk presented as a per virtual machine VMDK or Raw Device Mapping

  Or

- Isilon based NFS mounts
  Using an NFS-based mount for each VM's shuffle space provide flexibility and again decouples the compute node VM from any local or dedicated shared storage. This leverages the power of the Isilon scale-out storage and likely provides adequate quantities of shuffle which may be tricky to provision with VMDKs presented to each virtual machine. When using NFS based storage:
    - Use a dedicated mount point per VM compute node
    - Review the Isilon NFS best practices for client-side settings: https://support.emc.com/kb/457328

The next section contains a short summary of the main NFS mount option considerations for NFS mounted Shuffle space. For a definitive discussion of the options available on your specific NFS client platform, please consult your version-specific documentation such as the Unix/Linux nfs(5) manual pages.

Mount Options for NFS-based shuffle space

Since per compute node VM shuffle space is specific to each VM compute node and not shared between hosts, we recommend that you assess the following mount point options to maximize NFS throughput to this shuffle space:

- nolock - Enables client-local lock management for greatly-improved performance
- noatime
- nodiratime
- actimeo=86400 – That is one day in seconds, but any reasonably-large value should suffice
- rdirplus - Improves 'ls -l' performance. This is normally the default, but re-stating it should not be a problem.
- nocto - This option may provide some performance benefit, but may also result in cascaded performance issues with memory management. Its utility must be evaluated on a case-by-case basis by empirical testing.
- nfsvers=3
- tcp

For example:

mount -t nfs -o nfsvers=3, tcp, nolock,noatime,nodiratime,rdirplus, nocto, actimeo=86400 <Isilon>:<export> <mount_point>

Mounts should be made persistent by adding into the clients /etc/fstab so they are available at all times.

NFS Exports

Since most NFS performance factors are associated with the client and client mount options, one can use a single NFS export to facilitate all of the mounts associated with a given Hadoop VM compute node landscape. No specific export tuning is required; the configuration of the exports should meet the requirements of your compute node VM access requirements.

The required permissions and access controls should be applied to each NFS export to meet the security requirements of the deployed configuration.

Isilon Connectivity

Isilon operates a DNS responder-based load balancer to distribute traffic across the Isilon cluster nodes. This feature is known as *SmartConnect Advanced*. We recommend that you use SmartConnect Advanced and Dynamic IP Pools for NFS. Dynamic SmartConnect Pools are required for NFS v3 IP failover functionality. Support for NFSv4 graceful failover is in OneFS 8.0.x, but it is not recommended for these shuffle space mount points. For additional information on implementing a Dynamic SmartConnect Zone, see the External Network Connectivity Guide.

In order to provide per node dedicated shuffle space, we recommend that you use a dedicated data path for each compute node VM similar to the following:

For example: where SCZ-FQDN is the Isilon SmartConnect DNS zone name

VM compute node1 - /mnt/Isilon/shuffle -- > SCZ-FQDN:/ifs/zone/hadoop-root/shuffle/node1

VM compute node2 - /mnt/Isilon/shuffle -- > SCZ-FQDN:/ifs/zone/hadoop-root/shuffle/node2

VM compute node3 - /mnt/Isilon/shuffle -- > SCZ-FQDN:/ifs/zone/hadoop-root/shuffle/node3

VM compute node(X) - /mnt/Isilon/shuffle area1 -- > SCZ-FQDN:/ifs/zone/hadoop-root/shuffle/node(X)

Recommendations:

- Assign the appropriate numbers of IP addresses to the SmartConnect Network Pool
- Use a Dynamic allocation methodology for the SmartConnect Network Pool

Validate and review all mounts

Confirm NFS mount options under Linux by using the 'mountstats' command for each mount point. The mountstats command reveals the values that were actually chosen by each mount negotiation process; for various reasons, those values may not be the same as what is specified in /etc/fstab and later shown by the 'mount' command. Review and assess the actual mount configuration for deviation and address accordingly.

Hadoop Clusters

Assign the mounted temporary shuffle storage to the allocated mount points in the VM while deploying the Hadoop cluster.

OS CONSIDERATIONS

All standard best practices for OS deployments should be followed when deploying the VMs.

- OS tuning guidelines for individual virtualized host should be applied on a per Operating System basis
- VMWare Tools should be installed on all VMs deployed
- All recommended OS configuration tunings and best practice for Hadoop cluster deployments should be followed

BEST PRACTICES

Since Isilon OneFS is providing all NameNode and DataNode services for the virtualized Hadoop compute cluster in the same manner that it does for a physical based compute clusters, Isilon OneFS should be deployed using the same deployment guides by following the same best practices. For supported versions, see the Hadoop Distributions and Products Supported by OneFS compatibility matrix. No specific tuning or best practices exist for virtualized Hadoop nodes versus physical Hadoop nodes, as Isilon is agnostic to this deployment model.

CONCLUSION

This whitepaper outlines the capabilities and options of deploying a virtualized Hadoop compute cluster with Isilon OneFS which provides many benefits to leveraging your existing datacenter infrastructure in an optimized manner. The ability to decouple storage from the Hadoop nodes is the key to successfully deploying a virtualized compute cluster, and Isilon OneFS provides the ideal platform for enterprise-grade HDFS storage services to your compute cluster, allowing the virtual machines to be optimized for the compute services only in the Hadoop cluster.