ONTAP and enterprise applications

Enterprise applications

NetApp
April 29, 2024

This PDF was generated from https://docs.netapp.com/us-en/ontap-apps-dbs/index.html on April 29, 2024. Always check docs.netapp.com for the latest.

Table of Contents

ONTAP and enterprise applications

Hyper-V

Deployment guidelines and storage best practices

Overview

Microsoft Windows Server is an enterprise-class operating system (OS) that covers networking, security, virtualization, private cloud, hybrid cloud, virtual desktop infrastructure, access protection, information protection, web services, application platform infrastructure, and much more.

> **This documentation replaces previously published technical reports *TR-4568: NetApp Deployment Guidelines and Storage Best Practices for Windows Server***

NetApp ONTAP® management software runs on NetApp storage controllers. It is available in multiple formats.

- A unified architecture supporting file, object, and block protocols. This enables the storage controllers to act as both NAS and SAN devices as well as object stores
- An All SAN Array (ASA) that focuses only on block protocols and optimizes I/O resume times (IORT) by adding symmetric active-active multipathing for connect hosts
- A software defined unified architecture
  - ONTAP Select running on VMware vSphere or KVM
  - Cloud Volumes ONTAP running as a cloud native instance
- First party offerings from hyperscale cloud providers
  - Amazon FSx for NetApp ONTAP
  - Azure NetApp Files
  - Google Cloud NetApp Volumes

ONTAP provides NetApp storage efficiency features such as NetApp Snapshot® technology, cloning, deduplication, thin provisioning, thin replication, compression, virtual storage tiering, and much more with enhanced performance and efficiency.

Together, Windows Server and ONTAP can operate in large environments and bring immense value to data center consolidation and private or hybrid cloud deployments. This combination also provides nondisruptive workloads efficiently and supports seamless scalability.

Intended audience

This document is intended for system and storage architects who design NetApp storage solutions for the Windows Server.

We make the following assumptions in this document:

- The reader has general knowledge of NetApp hardware and software solutions. See the System Administration Guide for Cluster Administrators for details.
- The reader has general knowledge of block-access protocols, such as iSCSI, FC and the file-access protocol SMB/CIFS. See the Clustered Data ONTAP SAN management for SAN-related information. See

the NAS management for CIFS/SMB-related information.

- The reader has general knowledge of the Windows Server OS and Hyper-V.

For a complete, regularly updated matrix of tested and supported SAN and NAS configurations, see the Interoperability Matrix Tool (IMT) on the NetApp Support site. With the IMT, you can determine the exact product and feature versions that are supported for your specific environment. The NetApp IMT defines the product components and versions that are compatible with NetApp supported configurations. Specific results depend on each customer's installation in accordance with published specifications.

NetApp storage and Windows Server environment

As mentioned in the Overview, NetApp storage controllers provide a truly unified architecture that supports file, block, and object protocols. This includes SMB/CIFS, NFS, NVMe/TCP, NVMe/FC, iSCSI, FC(FCP) and S3, and they create unified client and host access. The same storage controller can concurrently deliver block storage service in the form of SAN LUNs and file service as NFS and SMB/CIFS. ONTAP is also available as an All SAN Array (ASA) that optimizes host access though symmetric active-active multipathing with iSCSI and FCP, whereas the unified ONTAP systems use asymmetric active-active multipathing. In both modes, ONTAP uses ANA for NVMe over Fabrics (NVMe-oF) multipath management.

A NetApp storage controller running ONTAP software can support the following workloads in a Windows Server environment:

- VMs hosted on continuously available SMB 3.0 shares
- VMs hosted on Cluster Shared Volume (CSV) LUNs running on iSCSI or FC
- SQL Server databases on SMB 3.0 shares
- SQL Server databases on NVMe-oF, iSCSI or FC
- Other application workloads

In addition, NetApp storage efficiency features such as deduplication, NetApp FlexClone® copies, NetApp Snapshot technology, thin provisioning, compression, and storage tiering provide significant value for workloads running on Windows Server.

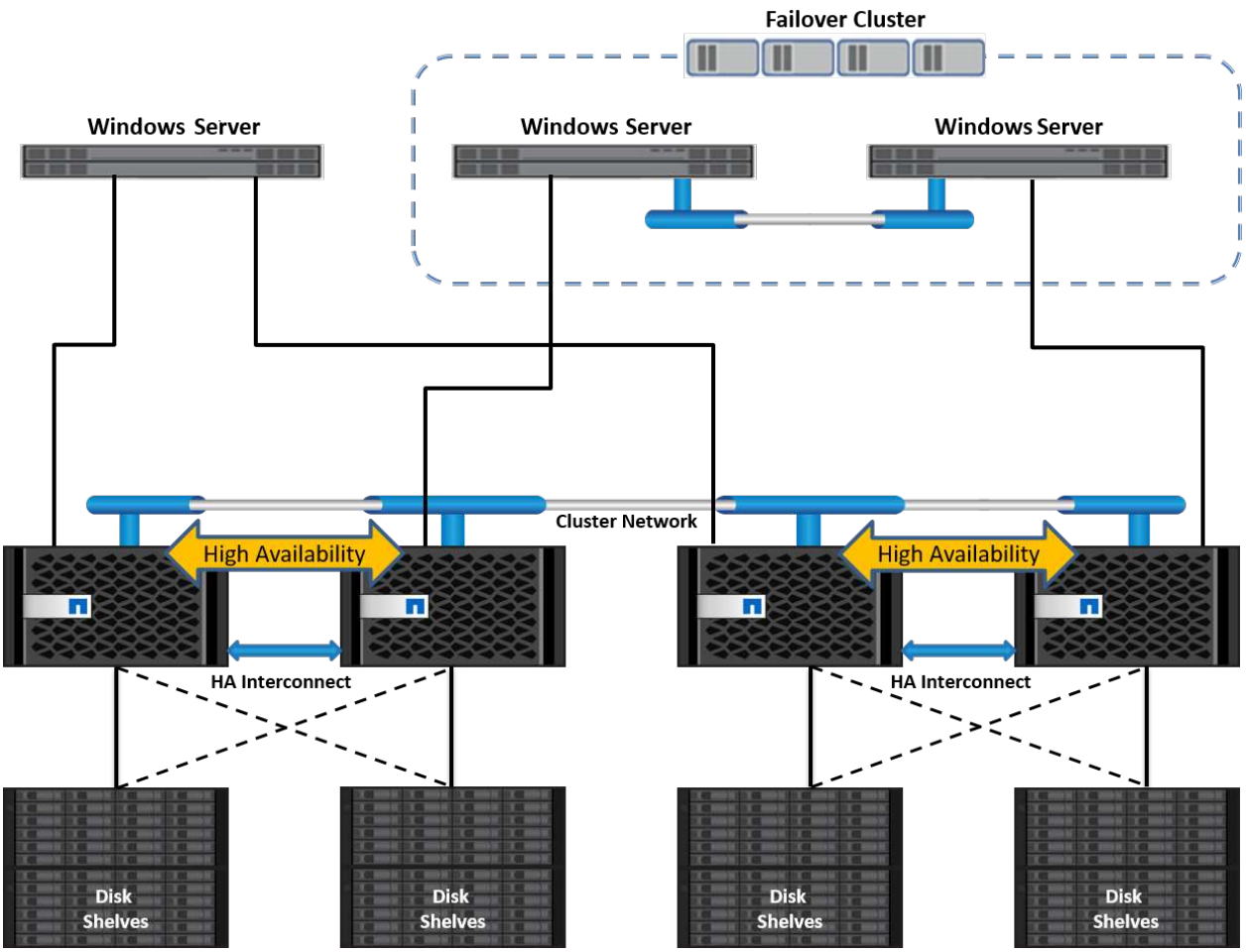
ONTAP data management

ONTAP is management software that runs on a NetApp storage controller. Referred to as a node, a NetApp storage controller is a hardware device with a processor, RAM, and NVRAM. The node can be connected to SATA, SAS, or SSD disk drives or a combination of those drives.

Multiple nodes are aggregated into a clustered system. The nodes in the cluster communicate with each other continuously to coordinate cluster activities. The nodes can also move data transparently from node to node by using redundant paths to a dedicated cluster network consisting of two 10Gb Ethernet switches. The nodes in the cluster can take over one another to provide high availability during any failover scenarios. Clusters are administered on a whole-cluster rather than a per-node basis, and data is served from one or more storage virtual machines (SVMs). A cluster must have at least one SVM to serve data.

The basic unit of a cluster is the node, and nodes are added to the cluster as part of a high-availability (HA) pair. HA pairs enable high availability by communicating with each other over an HA interconnect (separate from the dedicated cluster network) and by maintaining redundant connections to the HA pair's disks. Disks are

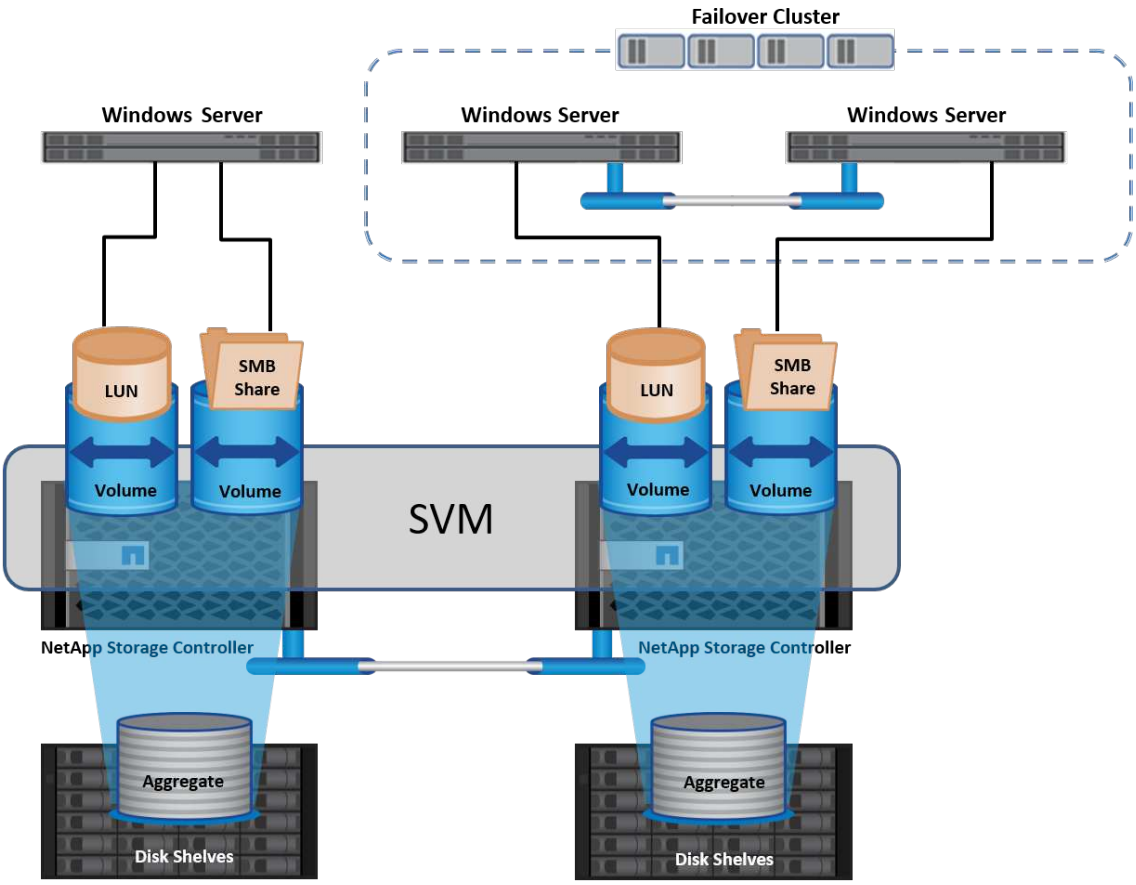
not shared between HA pairs, although shelves might contain disks that belong to either member of an HA pair. The following figure depicts a NetApp storage deployment in a Windows Server environment.



Storage Virtual Machines

An ONTAP SVM is a logical storage server that provides data access to LUNs and/or a NAS namespace from one or more logical interfaces (LIFs). The SVM is thus the basic unit of storage segmentation that enables secure multitenancy in ONTAP. Each SVM is configured to own storage volumes provisioned from a physical aggregate and logical interfaces (LIFs) assigned either to a physical Ethernet network or to FC target ports.

Logical disks (LUNs) or CIFS shares are created inside an SVM's volumes and are mapped to Windows hosts and clusters to provide them with storage space, as shown in the following figure. SVMs are node independent and cluster based; they can use physical resources such as volumes or network ports anywhere in the cluster.



Provisioning NetApp storage for Windows Server

Storage can be provisioned to Windows Server in both SAN and NAS environments. In a SAN environment, the storage is provided as disks from LUNs on NetApp volume as block storage. In a NAS environment, the storage is provided as CIFS/SMB shares on NetApp volumes as file storage. These disks and shares can be applied in Windows Server as follows:

- Storage for Windows Server hosts for application workloads
- Storage for Nano Server and containers
- Storage for individual Hyper-V hosts to store VMs
- Shared storage for Hyper-V clusters in the form of CSVs to store VMs
- Storage for SQL Server databases

Managing NetApp storage

To connect, configure, and manage NetApp storage from Windows Server 2016, use one of the following methods:

- **Secure Shell (SSH).** Use any SSH client on Windows Server to run NetApp CLI commands.
- **System Manager.** This is NetApp's GUI-based manageability product.
- **NetApp PowerShell Toolkit.** This is the NetApp PowerShell Toolkit for automating and implementing custom scripts and workflows.

NetApp PowerShell Toolkit

NetApp PowerShell Toolkit (PSTK) is a PowerShell module that provides end-to-end automation and enables storage administration of NetApp ONTAP. The ONTAP module contains over 2,000 cmdlets and helps with the administration of FAS, NetApp All Flash FAS (AFF), commodity hardware, and cloud resources.

Things to remember

- NetApp does not support Windows Server storage spaces. Storage spaces are used only for JBOD (just a bunch of disks) and does not work with any type of RAID (direct-attached storage [DAS] or SAN).
- Clustered storage pools in Windows Server are not supported by ONTAP.
- NetApp supports the shared virtual hard disk format (VHDX) for guest clustering in Windows SAN environments.
- Windows Server does not support creating storage pools using iSCSI or FC LUNs.

Further reading

- For more information about the NetApp PowerShell Toolkit, visit the NetApp Support Site.
- For information about NetApp PowerShell Toolkit best practices, see TR-4475: NetApp PowerShell Toolkit Best Practices Guide.

Networking best practices

Ethernet networks can be broadly segregated into the following groups:

- A client network for the VMs
- One more more storage networks (iSCSI or SMB connecting to the storage systems)
- A cluster communication network (heartbeat and other communication between the nodes of the cluster)
- A management network (to monitor and troubleshoot the system)
- A migration network (for host live migration)
- VM replication (a Hyper-V Replica)

Best practices

- NetApp recommends having dedicated physical ports for each of the preceding functionalities for network isolation and performance.
- For each of the preceding network requirements (except for the storage requirements), multiple physical network ports can be aggregated to distribute load or provide fault tolerance.
- NetApp recommends having a dedicated virtual switch created on the Hyper-V host for guest storage connection within the VM.
- Make sure that the Hyper-V host and guest iSCSI data paths use different physical ports and virtual switches for secure isolation between the guest and the host.
- NetApp recommends avoiding NIC teaming for iSCSI NICs.
- NetApp recommends using ONTAP multipath input/output (MPIO) configured on the host for storage purposes..
- NetApp recommends using MPIO within a guest VM if using guest iSCSI initiators. MPIO usage must be avoided within the guest if you use pass-through disks. In this case, installing MPIO on the host should suffice.

- NetApp recommends not applying QoS policies to the virtual switch assigned for the storage network.
- NetApp recommends not using automatic private IP addressing (APIPA) on physical NICs because APIPA is nonroutable and not registered in the DNS.
- NetApp recommends turning on jumbo frames for CSV, iSCSI, and live migration networks to increase the throughput and reduce CPU cycles.
- NetApp recommends unchecking the option Allow Management Operating System to Share This Network Adapter for the Hyper-V virtual switch to create a dedicated network for the VMs.
- NetApp recommends creating redundant network paths (multiple switches) for live migration and the iSCSI network to provide resiliency and QoS.

Provisioning in SAN environments

ONTAP SVMs support the block protocols iSCSI and FC. When an SVM is created with block protocol iSCSI or FC, the SVM gets either an iSCSI Qualified Name (IQN) or an FC worldwide name (WWN), respectively. This identifier presents a SCSI target to hosts that access NetApp block storage.

Provisioning NetApp LUN on Windows Server

Prerequisites

Using NetApp storage in SAN environments in Windows Server has the following requirements:

- A NetApp cluster is configured with one or more NetApp storage controllers.
- The NetApp cluster or storage controllers have a valid iSCSI license.
- iSCSI and/or FC configured ports are available.
- FC zoning is performed on an FC switch for FC.
- At least one aggregate is created.
- An SVM should have one LIF per Ethernet network or Fibre Channel fabric on every storage controller that is going to serve data using iSCSI or Fibre Channel.

Deployment

1. Create a new SVM with block protocol iSCSI and/or FC enabled. A new SVM can be created with any of the following methods:
    - CLI commands on NetApp storage
    - ONTAP System Manager
    - NetApp PowerShell Toolkit
1. Configure the iSCSI and/or FC protocol.
2. Assign the SVM with LIFs on each cluster node.
3. Start the iSCSI and/or FC service on the SVM.

    .

4. Create iSCSI and/or FC port sets using the SVM LIFs.

5. Create an iSCSI and/or FC initiator group for Windows using the port set created.
6. Add an initiator to the initiator group. The initiator is the IQN for iSCSI and WWPN for FC. They can be queried from Windows Server by running the PowerShell cmdlet Get-InitiatorPort.

```
# Get the IQN for iSCSI
Get-InitiatorPort | Where \{$_.ConnectionType -eq 'iSCSI'} | Select-
Object -Property NodeAddress
```

```
# Get the WWPN for FC
Get-InitiatorPort | Where \{$_.ConnectionType -eq 'Fibre Channel'} |
Select-Object -Property PortAddress
```

```
# While adding initiator to the initiator group in case of FC, make sure
to provide the initiator(PortAddress) in the standard WWPN format
```

The IQN for iSCSI on Windows Server can also be checked in the configuration of the iSCSI initiator properties.

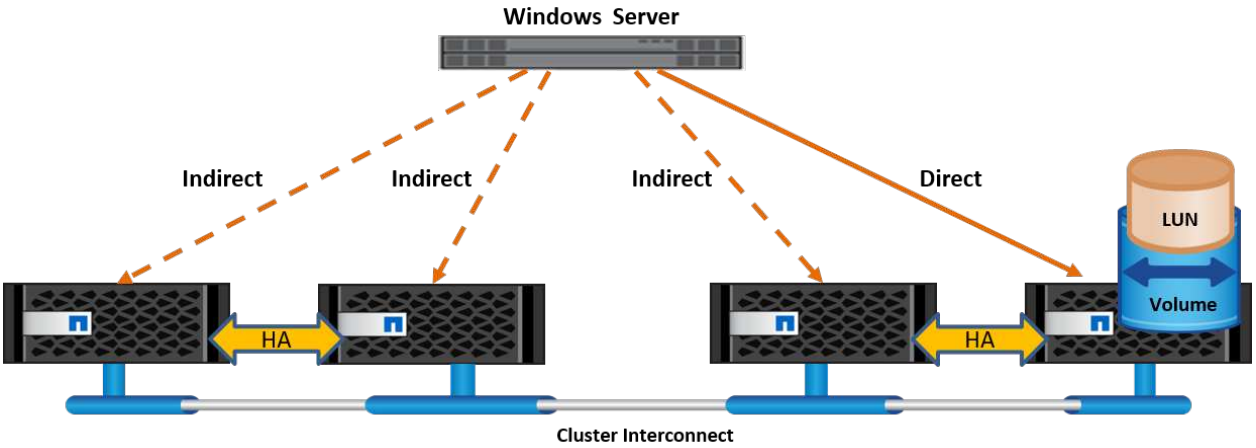
- Create a LUN using Create LUN wizard and associate it with the initiator group created.

Host integration

Windows Server uses Asymmetrical Logical Unit Access (ALUA) extension MPIO to determine direct and indirect paths to LUNs. Even though every LIF owned by an SVM accepts read/write requests for its LUNs, only one of the cluster nodes actually owns the disks backing that LUN at any given moment. This divides available paths to a LUN into two types, direct or indirect, as shown in the following figure.

A direct path for a LUN is a path on which an SVM's LIFs and the LUN being accessed reside on the same node. To go from a physical target port to disk, it is not necessary to traverse the cluster network.

Indirect paths are data paths on which an SVM's LIFs and the LUN being accessed reside on different nodes. Data must traverse the cluster network to go from a physical target port to disk.



MPIO

NetApp ONTAP provide highly available storage in which multiple paths from the storage controller to the Windows Server can exist. Multipathing is the ability to have multiple data paths from a server to a storage array. Multipathing protects against hardware failures (cable cuts, switch and host bus adapter [HBA] failure, and so on), and it can provide higher performance limits by using the aggregate performance of multiple connections. When one path or connection becomes unavailable, the multipathing software automatically shifts the load to one of the other available paths. The MPIO feature combines the multiple physical paths to the storage as a single logical path that is used for data access to provide storage resiliency and load balancing. To use this feature, the MPIO feature must be enabled on Windows Server.

Enable MPIO

To enable MPIO on Windows Server, complete the following steps:

1. Log in to Windows Server as a member of the administrator group.

7. Start Server Manager.
8. In the Manage section, click Add Roles and Features.
9. In the Select Features page, select Multipath I/O.

Configure MPIO

When using the iSCSI protocol, you must tell Windows Server to apply multipath support to iSCSI devices in the MPIO properties.

To configure MPIO on Windows Server, complete the following steps:

1. Log on to Windows Server as a member of the administrator group.

10. Start Server Manager.
11. In the Tools section, click MPIO.
12. In MPIO Properties on Discover Multi-Paths, select Add Support for iSCSI Devices and click Add. A prompt then asks you to restart the computer.
13. Reboot Windows Server to see the MPIO device listed in the MPIO Devices section of MPIO Properties.

Configure iSCSI

To detect iSCSI block storage on Windows Server, complete the following steps:

1. Log on to Windows Server as a member of the administrator group.

14. Start Server Manager.
15. In the Tools section, click iSCSI Initiator.
16. Under the Discovery tab, click Discover Portal.
17. Provide the IP address of the LIFs associated with the SVM created for the NetApp storage for SAN protocol. Click Advanced, configure the information in the General tab, and click OK.
18. The iSCSI initiator automatically detects the iSCSI target and lists it in the Targets tab.
19. Select the iSCSI target in Discovered Targets. Click Connect to open the Connect To Target window.
20. You must create multiple sessions from the Windows Server host to the target iSCSI LIFs on the NetApp

storage cluster. To do so, complete the following steps:

a. In the Connect to Target window, select Enable MPIO and click Advanced.
b. In Advanced Settings under the General tab, select the local adapter as the Microsoft iSCSI initiator and select the Initiator IP and Target Portal IP.
c. You must also connect using the second path. Therefore, repeat step 5 through step 8, but this time select the Initiator IP and Target Portal IP for the second path.
d. Select the iSCSI target in Discovered Targets on the iSCSI Properties main window and click Properties.
e. The Properties window shows that multiple sessions have been detected. Select the session, click Devices, and then click the MPIO to configure the load balancing policy. All the paths configured for the device are displayed and all load balancing policies are supported. NetApp generally recommends round robin with subset, and this setting is the default for arrays with ALUA enabled. Round robin is the default for active-active arrays that do not support ALUA.

Detect block storage

To detect iSCSI or FC block storage on Windows Server, complete the following steps:

1. Click Computer Management in the Tools section of the Server Manager.
2. In Computer Management, click the Disk Management in Storage section and then click More Actions and Rescan Disks. Doing so displays the raw iSCSI LUNs.
3. Click the discovered LUN and make it online. Then select Initialize Disk using the MBR or GPT partition. Create a new simple volume by providing the volume size and drive letter and format it using FAT, FAT32, NTFS, or the Resilient File System (ReFS).

Best practices

- NetApp recommends enabling thin provisioning on the volumes hosting the LUNs.
- To avoid multipathing problems, NetApp recommends using either all 10Gb sessions or all 1Gb sessions to a given LUN.
- NetApp recommends that you confirm that ALUA is enabled on the storage system. ALUA is enabled by default on ONTAP.
- On the Windows Server host to where the NetApp LUN is mapped, enable iSCSI Service (TCP-In) for Inbound and iSCSI Service (TCP-Out) for Outbound in the firewall settings. These settings allow iSCSI traffic to pass to and from the Hyper-V host and NetApp controller.

Provisioning NetApp LUNs on Nano Server

Prerequisites

In addition to the prerequisites mentioned in the previous section, the storage role must be enabled from the Nano Server side. For example, Nano Server must be deployed using the -Storage option. To deploy Nano Server, see the section "Deploy Nano Server."

Deployment

To provision NetApp LUNs on a Nano Server, complete the following steps:

1. Connect to the Nano Server remotely using instructions in the section "Connect to Nano Server."
2. To configure iSCSI, run the following PowerShell cmdlets on the Nano Server:

```
# Start iSCSI service, if it is not already running
Start-Service msiscsi
```

```
# Create a new iSCSI target portal
New-IscsiTargetPortal â€"TargetPortalAddress <SVM LIF>
```

```
# View the available iSCSI targets and their node address
Get-IscsiTarget
```

```
# Connect to iSCSI target
Connect-IscsiTarget -NodeAddress <NodeAddress>
```

```
# NodeAddress is retrived in above cmdlet Get-IscsiTarget
# OR
Get-IscsiTarget | Connect-IscsiTarget
```

```
# View the established iSCSI session
Get-IscsiSession
```

```
# Note the InitiatorNodeAddress retrieved in the above cmdlet Get-
IscsiSession. This is the IQN for Nano server and this needs to be added
in the Initiator group on NetApp Storage
```

```
# Rescan the disks
Update-HostStorageCache
```

3. Add an initiator to the initiator group.

```
Add the InitiatorNodeAddress retrieved from the cmdlet Get-IscsiSession
to the Initiator Group on NetApp Controller
```

4. Configure MPIO.

```
# Enable MPIO Feature
Enable-WindowsOptionalFeature -Online -FeatureName MultipathIo
```

```
# Get the Network adapters and their IPs
Get-NetIPAddress â€"AddressFamily IPv4 â€"PrefixOrigin <Dhcp or Manual>
```

```
# Create one MPIO-enabled iSCSI connection per network adapter
Connect-IscsiTarget -NodeAddress <NodeAddress> -IsPersistent $True â
€"IsMultipathEnabled $True â€"InitiatorPortalAddress <IP Address of
ethernet adapter>
```

```
# NodeAddress is retrieved from the cmdlet Get-IscsiTarget
# IPs are retrieved in above cmdlet Get-NetIPAddress
```

```
# View the connections
Get-IscsiConnection
```

5. Detect block storage.

```
# Rescan disks
Update-HostStorageCache
```

```
# Get details of disks
Get-Disk
```

```
# Initialize disk
Initialize-Disk -Number <DiskNumber> -PartitionStyle <GPT or MBR>
```

```
# DiskNumber is retrived in the above cmdlet Get-Disk
# Bring the disk online
Set-Disk -Number <DiskNumber> -IsOffline $false
```

```
# Create a volume with maximum size and default drive letter
New-Partition -DiskNumber <DiskNumber> -UseMaximumSize
-AssignDriveLetter
```

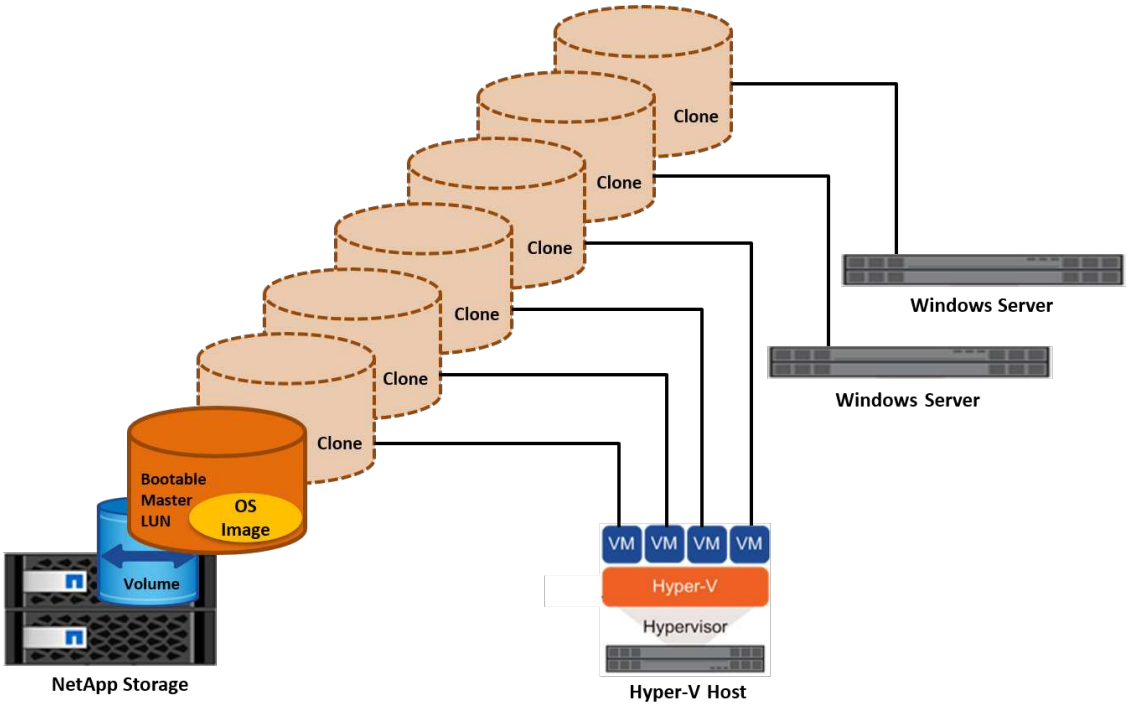
```
# To choose the size and drive letter use -Size and -DriveLetter
parameters
# Format the volume
Format-Volume -DriveLetter <DriveLetter> -FileSystem <FAT32 or NTFS or
REFS>
```

Boot from SAN

A physical host (server) or a Hyper-V VM can boot the Windows Server OS directly from a NetApp LUN instead of its internal hard disk. In the boot-from-SAN approach, the OS image to boot from resides on a NetApp LUN that is attached to a physical host or VM. For a physical host, the HBA of the physical host is configured to use the NetApp LUN for booting. For a VM, the NetApp LUN is attached as a pass-through disk for booting.

NetApp FlexClone approach

Using NetApp FlexClone technology, boot LUNs with an OS image can be cloned instantly and attached to the servers and VMs to rapidly provide clean OS images, as show in the following figure.



Boot from SAN for physical host

Prerequisites

- The physical host (server) has a proper iSCSI or FC HBA.
- You have downloaded a suitable HBA device driver for the server supporting Windows Server.
- The server has a suitable CD/DVD drive or virtual media to insert the Windows Server ISO image and the HBA device driver has been downloaded.
- A NetApp iSCSI or FC LUN is provisioned on the NetApp storage controller.

Deployment

To configure booting from SAN for a physical host, complete the following steps:

1. Enable BootBIOS on the server HBA.
2. For iSCSI HBAs, configure the Initiator IP, iSCSI node name, and adapter boot mode in the boot BIOS settings.
3. When creating an initiator group for iSCSI and/or FC on a NetApp storage controller, add the server HBA initiator to the group. The HBA initiator of the server is the WWPN for the FC HBA or iSCSI node name for iSCSI HBA.
4. Create a LUN on the NetApp storage controller with a LUN ID of 0 and associate it with the initiator group created in the previous step. This LUN serves as a boot LUN.
5. Restrict the HBA to a single path to the boot LUN. Additional paths can be added after Windows Server is installed on the boot LUN to exploit the multipathing feature.
6. Use the HBA's BootBIOS utility to configure the LUN as a boot device.
7. Reboot the host and enter the host BIOS utility.
8. Configure the host BIOS to make the boot LUN the first device in the boot order.
9. From the Windows Server ISO, launch the installation setup.
10. When the installation asks, "Where Do You Want to Install Windows?," click Load Driver at the bottom of the installation screen to launch the Select Driver to Install page. Provide the path of the HBA device driver downloaded earlier and finish the installation of the driver.
11. Now the boot LUN created previously must be visible on the Windows installation page. Select the boot LUN for installation of Windows Server on the boot LUN and finish the installation.

Boot from SAN for virtual machine

To configure booting from SAN for a VM, complete the following steps:

Deployment

1. When creating an initiator group for iSCSI or FC on a NetApp storage controller, add the IQN for iSCSI or the WWN for FC of the Hyper-V server to the controller.
2. Create LUNs or LUN clones on the NetApp storage controller and associate them with the initiator group created in the previous step. These LUNs serve as boot LUNs for the VMs.
3. Detect the LUNs on the Hyper-V server, bring them online, and initialize them.
4. Bring the LUNs offline.
5. Create VMs with the option Attach a Virtual Hard Disk Later on the Connect Virtual Hard Disk page.
6. Add a LUN as a pass-through disk to a VM.

a. Open the VM settings.
b. Click IDE Controller 0, select Hard Drive, and click Add. Selecting IDE Controller 0 makes this disk the first boot device for the VM.
c. Select Physical Hard Disk in the Hard Disk options and select a disk from the list as a pass-through disk. The disks are the LUNs configured in the previous steps.

7. Install Windows Server on the pass-through disk.

Best practices

- Make sure that the LUNs are offline. Otherwise, the disk cannot be added as a pass-through disk to a VM.
- When multiple LUNs exist, be sure to note the disk number of the LUN in disk management. Doing so is necessary because disks listed for the VM are listed with the disk number. Also, the selection of the disk as a pass-through disk for the VM is based on this disk number.
- NetApp recommends avoiding NIC teaming for iSCSI NICs.
- NetApp recommends using ONTAP MPIO configured on the host for storage purposes.

Provisioning in SMB environments

ONTAP provides resilient and high performance NAS storage for Hyper-V virtual machines using the SMB3 protocol.

When an SVM is created with the CIFS protocol, a CIFS server runs on top of the SVM that is part of the Windows Active Directory Domain. SMB shares can be used for a home directory and to host Hyper-V and SQL Server workloads. The following SMB 3.0 features are supported in ONTAP:

- Persistent handles (continuously available file shares)
- Witness protocol
- Clustered client failover
- Scale-out awareness
- ODX
- Remote VSS

Provisioning SMB shares on Windows Server

Prerequisites

Using NetApp storage in NAS environments in Windows Server has the following requirements:

- ONTAP cluster have a valid CIFS license.
- At least one aggregate is created.
- One data logical interface (LIF) is created and the data LIF must be configured for CIFS.
- A DNS-configured Windows Active Directory domain server and domain administrator credentials are present.
- Each node in the NetApp cluster is time synchronized with the Windows domain controller.

Active Directory Domain Controller

A NetApp storage controller can be joined to and operate within an Active Directory similar to a Windows Server. During the creation of the SVM, you can configure the DNS by providing the domain name and name server details. The SVM attempts to search for an Active Directory domain controller by querying the DNS for an Active Directory/Lightweight Directory Access Protocol (LDAP) server in a manner similar to Windows Server.

For the CIFS setup to work properly, the NetApp storage controllers must be time synchronized with the Windows domain controller. NetApp recommends having a time skew between the Windows domain controller and the NetApp storage controller of not more than five minutes. It is a best practice to configure the Network Time Protocol (NTP) server for the ONTAP cluster to synchronize with an external time source. To configure the Windows domain controller as the NTP server, run the following command on your ONTAP cluster:

```
$domainControllerIP = "<input IP Address of windows domain controller>"
cluster::> system services ntp server create â€“server $domainControllerIP
```

Deployment

1. Create a new SVM with the NAS protocol CIFS enabled. A new SVM can be created with any of the following methods:
   - CLI commands on NetApp ONTAP
   - System Manager
   - The NetApp PowerShell Toolkit
2. Configure the CIFS protocol
   a. Provide the CIFS server name.
   b. Provide the Active Directory to which the CIFS server must be joined. You must have the domain administrator credentials to join the CIFS server to the Active Directory.
3. Assign the SVM with LIFs on each cluster node.
4. Start the CIFS service on the SVM.
5. Create a volume with the NTFS security style from the aggregate.
6. Create a qtree on the volume (optional).
7. Create shares that correspond to the volume or qtree directory so that they can be accessed from Windows Server. Select Enable Continuous Availability for Hyper-V during the creation of the share if the share is used for Hyper-V storage. Doing so enables high availability for file shares.
8. Edit the share created and modify the permissions as required for accessing the share. The permissions for the SMB share must be configured to grant access for the computer accounts of all the servers accessing this share.

Host integration

The NAS protocol CIFS is natively integrated into ONTAP. Therefore, Windows Server does not require any additional client software to access data on NetApp ONTAP. A NetApp storage controller appears on the network as a native file server and supports Microsoft Active Directory authentication.

To detect the CIFS share created previously with Windows Server, complete the following steps:

1. Log in to Windows Server as a member of the administrator group.
2. Go to run.exe and type the complete path of the CIFS share created to access the share.
3. To permanently map the share onto the Windows Server, right-click This PC, click Map Network Drive, and provide the path of the CIFS share.
4. Certain CIFS management tasks can be performed using Microsoft Management Console (MMC). Before performing these tasks, you must connect the MMC to the NetApp ONTAP storage using the MMC menu commands.
    a. To open the MMC in Windows Server, click Computer Management in the Tools section of Server Manager.
    b. Click More Actions and Connect to Another Computer, which opens the Select Computer dialog.
    c. Enter the name of the CIFS server or the IP address of the SVM LIF to connect to the CIFS server.
    d. Expand System Tools and Shared Folders to view and manage open files, sessions, and shares.

Best practices

- To confirm that there is no downtime when a volume is moved from one node to another or in the case of a node failure, NetApp recommends that you enable the continuous availability option on the file share.
- When provisioning VMs for a Hyper-V-over-SMB environment, NetApp recommends that you enable copy offload on the storage system. Doing so reduces the VMs' provisioning time.
- If the storage cluster hosts multiple SMB workloads such as SQL Server, Hyper-V, and CIFS servers, NetApp recommends hosting different SMB workloads on separate SVMs on separate aggregates. This configuration is beneficial because each of these workloads warrants unique storage networking and volume layouts.
- NetApp recommends connecting Hyper-V hosts and the NetApp ONTAP storage with a 10GB network if one is available. In the case of 1GB network connectivity, NetApp recommends creating an interface group consisting of multiple 1GB ports.
- When migrating VMs from one SMB 3.0 share to another, NetApp recommends enabling the CIFS copy offload functionality on the storage system so that migration is faster.

Things to remember

- When you provision volumes for SMB environments, the volumes must be created with the NTFS security style.
- Time settings on nodes in the cluster should be set up accordingly. Use the NTP if the NetApp CIFS server must participate in the Windows Active Directory domain.
- Persistent handles work only between nodes in an HA pair.
- The witness protocol works only between nodes in an HA pair.
- Continuously available file shares are supported only for Hyper-V and SQL Server workloads.
- The SMB multichannel is supported from ONTAP 9.4 onwards.
- RDMA is not supported.
- ReFS is not supported.

Provisioning SMB shares on Nano Server

Nano Server does not require additional client software to access data on the CIFS share on a NetApp storage controller.

To copy files from Nano Server to a CIFS share, run the following cmdlets on the remote server:

```
$ip = "<input IP Address of the Nano Server>"
```

```
# Create a New PS Session to the Nano Server
$session = New-PSSession -ComputerName $ip -Credential ~\Administrator
```

```
Copy-Item -FromSession $s -Path C:\Windows\Logs\DISM\dism.log -Destination
\\cifsshare
```

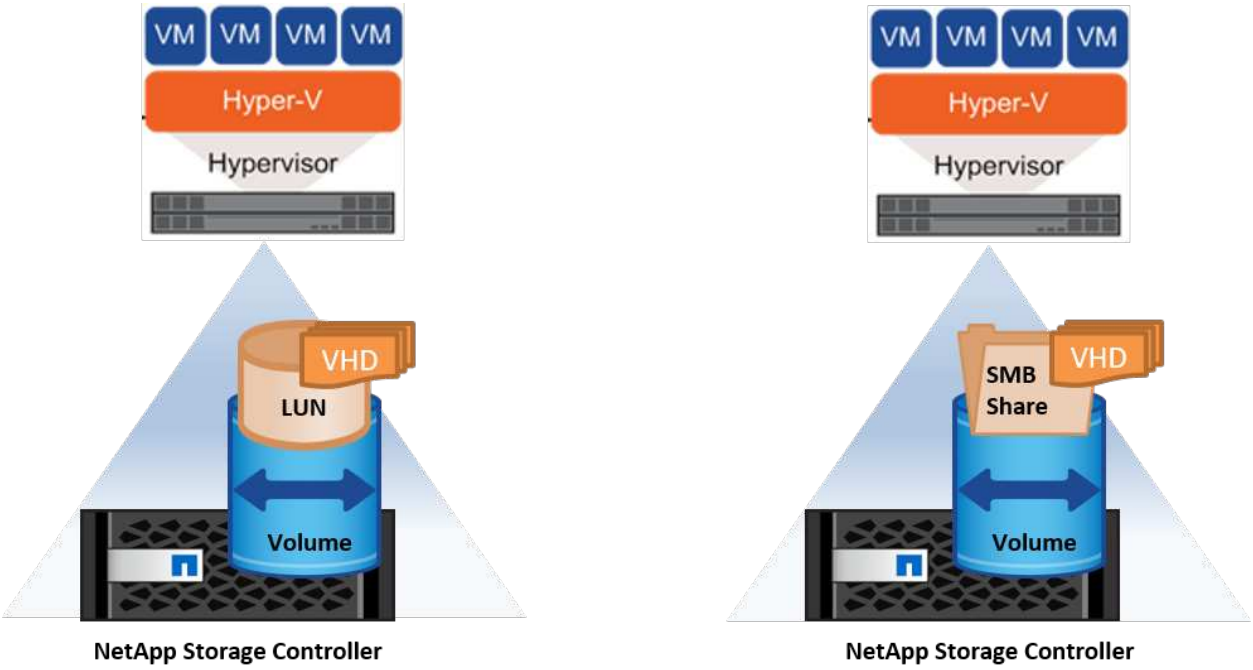
- `cifsshare` is the CIFS share on the NetApp storage controller.
- To copy files to Nano Server, run the following cmdlet:

  ```
  Copy-Item -ToSession $s -Path \\cifsshare\<file> -Destination C:\
  ```

To copy the entire contents of a folder, specify the folder name and use the -Recurse parameter at the end of the cmdlet.

Hyper-V storage infrastructure on NetApp

A Hyper-V storage infrastructure can be hosted on ONTAP storage systems. Storage for Hyper-V to store the VM files and its disks can be provided using NetApp LUNs or NetApp CIFS shares, as shown in the following figure.



Hyper-V Storage on NetApp LUNs

- Provision a NetApp LUN on the Hyper-V server machine. For more information, see the section "Provisioning in SAN Environments."
- Open Hyper-V Manager from the Tools section of Server Manager.
- Select the Hyper-V server and click Hyper-V Settings.
- Specify the default folder to store the VM and its disk as the LUN. Doing so sets the default path as the LUN for the Hyper-V storage. If you want to specify the path explicitly for a VM, then you can do so during VM creation.

Hyper-V Storage on NetApp CIFS

Before beginning the steps listed in this section, review the section "Provisioning in SMB Environments." To configure Hyper-V storage on the NetApp CIFS share, complete the following steps:

1. Open Hyper-V Manager from the Tools section of Server Manager.
2. Select the Hyper-V server and click Hyper-V Settings.
3. Specify the default folder to store the VM and its disk as the CIFS share. Doing so sets the default path as the CIFS share for the Hyper-V storage. If you want to specify the path explicitly for a VM, then you can do so during VM creation.

Each VM in Hyper-V can in turn be provided with the NetApp LUNs and CIFS shares that were provided to the physical host. This procedure is the same as for any physical host. The following methods can be used to provision storage to a VM:

- Adding a storage LUN by using the FC initiator within the VM
- Adding a storage LUN by using the iSCSI initiator within the VM
- Adding a pass-through physical disk to a VM
- Adding VHD/VHDX to a VM from the host

Best practices

- When a VM and its data are stored on NetApp storage, NetApp recommends running NetApp deduplication at the volume level at regular intervals. This practice results in significant space savings when identical VMs are hosted on a CSV or SMB share. Deduplication runs on the storage controller and it does not affect the host system and VM performance.
- When using iSCSI LUNs for Hyper-V, make sure to enable `iSCSI Service (TCP-In) for Inbound` and `iSCSI Service (TCP-Out) for Outbound` in the firewall settings on the Hyper-V host. Doing so allows iSCSI traffic to pass to and from the Hyper-V host and the NetApp controller.
- NetApp recommends unchecking the option Allow Management Operating System to Share This Network Adapter for the Hyper-V virtual switch. Doing so creates a dedicated network for the VMs.

Things to remember

- Provisioning a VM by using virtual Fibre Channel requires an N_Port ID Virtualizationâ€“enabled FC HBA. A maximum of four FC ports is supported.
- If the host system is configured with multiple FC ports and presented to the VM, then MPIO must be installed in the VM to enable multipathing.
- Pass-through disks cannot be provisioned to the host if MPIO is being used on that host, because pass-

through disks do not support MPIO.

- Disk used for VHD/VHDx files should use 64K formatting for allocation.

Further reading

- For information about FC HBAs, see the NetApp Interoperability Matrix.
- For more information about virtual Fibre Channel, see the Microsoft Hyper-V Virtual Fibre Channel Overview page.

Offloaded data transfer

Microsoft ODX, also known as copy offload, enables direct data transfers within a storage device or between compatible storage devices without transferring the data through the host computer. NetApp ONTAP supports the ODX feature for both CIFS and SAN protocols. ODX can potentially improve performance if copies are within same volume, reduce utilization of CPU and memory on the client, and reduce network I/O bandwidth utilization.

With ODX, it is faster and efficient to copy files within the SMB shares, within the LUNs, and between the SMB shares and LUNs if it's in same volume. This approach is more helpful in a scenario for which multiple copies of the golden image of an OS (VHD/VHDX) are required within same volume. Several copies of the same golden image can be made in significantly less time if copies are within same volume. ODX is also applied in Hyper-V storage live migration for moving VM storage.

If copy is across volumes, there may not be significant performance gains compared to host-based copies.

To enable the ODX feature on CIFS, run the following CLI commands on the NetApp storage controller:

1. Enable ODX for CIFS.
   #set the privilege level to diagnostic
   cluster::> set -privilege diagnostic

   ```
   #enable the odx feature
   cluster::> vserver cifs options modify -vserver <vserver_name> -copy
   -offload-enabled true
   ```

   ```
   #return to admin privilege level
   cluster::> set privilege admin
   ```

2. To enable the ODX feature on SAN, run the following CLI commands on the NetApp storage controller:
   #set the privilege level to diagnostic
   cluster::> set -privilege diagnostic

   ```
   #enable the odx feature
   cluster::> copy-offload modify -vserver <vserver_name> -scsi enabled
   ```

```
#return to admin privilege level
cluster::> set privilege admin
```

Things to remember

- For CIFS, ODX is available only when both the client and the storage server support SMB 3.0 and the ODX feature.
- For SAN environments, ODX is available only when both the client and the storage server support the ODX feature.

Further reading

For information about ODX, see Improving Microsoft Remote Copy Performance and Microsoft Offloaded Data Transfers .

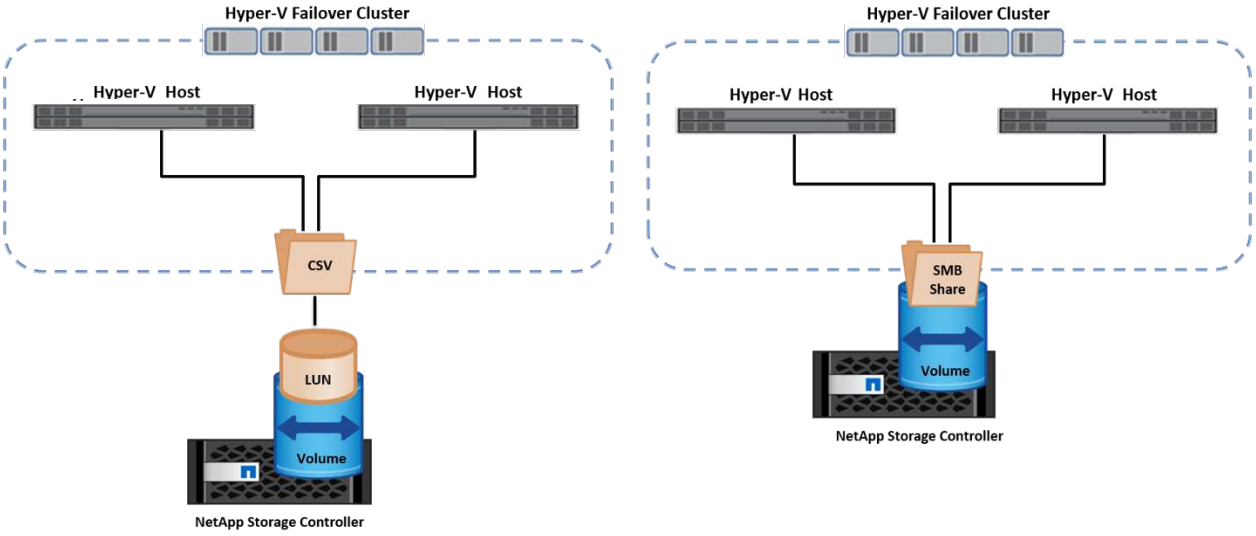
Hyper-V clustering: High availability and scalability for virtual machines

Failover clusters provide high availability and scalability to Hyper-V servers. A failover cluster is a group of independent Hyper-V servers that work together to increase availability and scalability for the VMs.

Hyper-V clustered servers (called nodes) are connected by the physical network and by cluster software. These nodes use shared storage to store the VM files, which include configuration, virtual hard disk (VHD) files, and Snapshot copies. The shared storage can be a NetApp SMB/CIFS share or a CSV on top of a NetApp LUN, as shown in Figure 6. This shared storage provides a consistent and distributed namespace that can be accessed simultaneously by all the nodes in the cluster. Therefore, if one node fails in the cluster, the other node provides service by a process called failover. Failover clusters can be managed by using the Failover Cluster Manager snap-in and the failover clustering Windows PowerShell cmdlets.

Cluster Shared Volumes

CSVs enable multiple nodes in a failover cluster to simultaneously have read/write access to the same NetApp LUN that is provisioned as an NTFS or ReFS volume. With CSVs, clustered roles can fail over quickly from one node to another without requiring a change in drive ownership or dismounting and remounting a volume. CSVs also simplify the management of a potentially large number of LUNs in a failover cluster. CSVs provide a general-purpose clustered file system that is layered above NTFS or ReFS.



Best practices

- NetApp recommends turning off cluster communication on the iSCSI network to prevent internal cluster communication and CSV traffic from flowing over the same network.
- NetApp recommends having redundant network paths (multiple switches) to provide resiliency and QoS.

Things to remember

- Disks used for CSV must be partitioned with NTFS or ReFS. Disks formatted with FAT or FAT32 cannot be used for a CSV.
- Disks used for CSVs should use 64K formatting for allocation.

Further reading

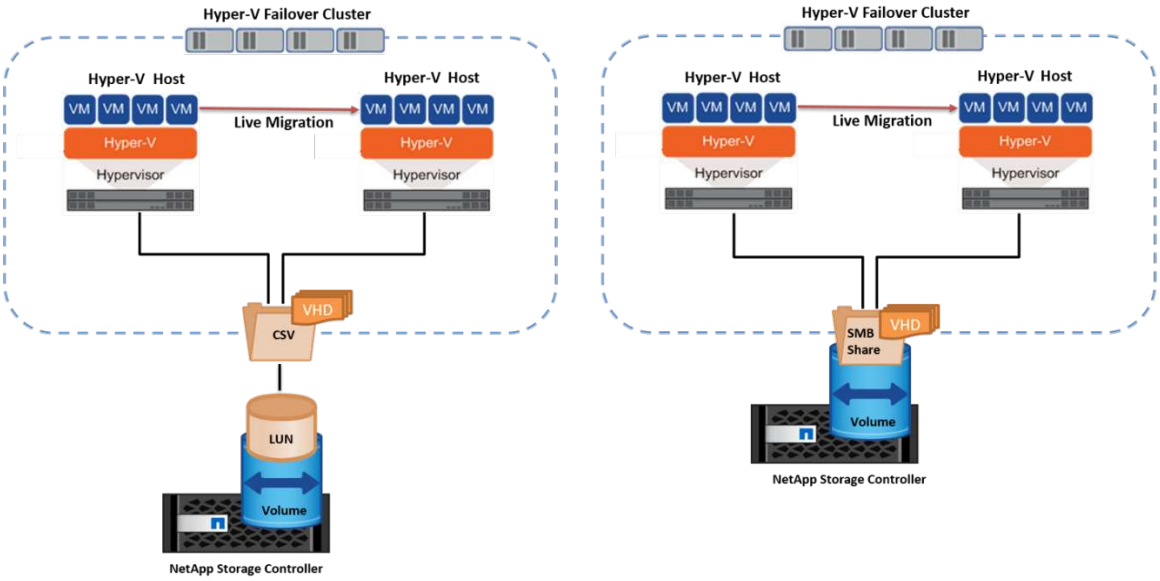
For information about deploying a Hyper-V cluster, see Appendix B: Deploy Hyper-V Cluster.

Hyper-V Live Migration: Migration of VMs

It is sometimes necessary during the lifetime of VMs to move them to a different host on the Windows cluster. Doing so might be required if the host is running out of system resources or if the host is required to reboot for maintenance reasons. Similarly, it might be necessary to move a VM to a different LUN or SMB share. This might be required if the present LUN or share is running out of space or yielding lower than expected performance. Hyper-V live migration moves running VMs from one physical Hyper-V server to another with no effect on VM availability to users. You can live migrate VMs between Hyper-V servers that are part of a failover cluster or between independent Hyper-V servers that are not part of any cluster.

Live Migration in a clustered environment

VMs can be moved seamlessly between the nodes of a cluster. VM migration is instantaneous because all the nodes in the cluster share the same storage and have access to the VM and its disk. The following figure depicts live migration in a clustered environment.



Best practice

- Have a dedicated port for live migration traffic.

- Have a dedicated host live migration network to avoid network-related issues during migration.

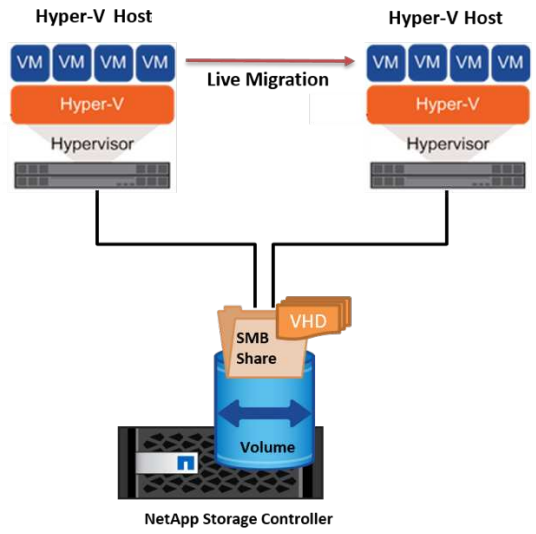
Further reading

For information about deploying live migration in a clustered environment, see Appendix C: Deploy Hyper-V Live Migration in a Clustered Environment.

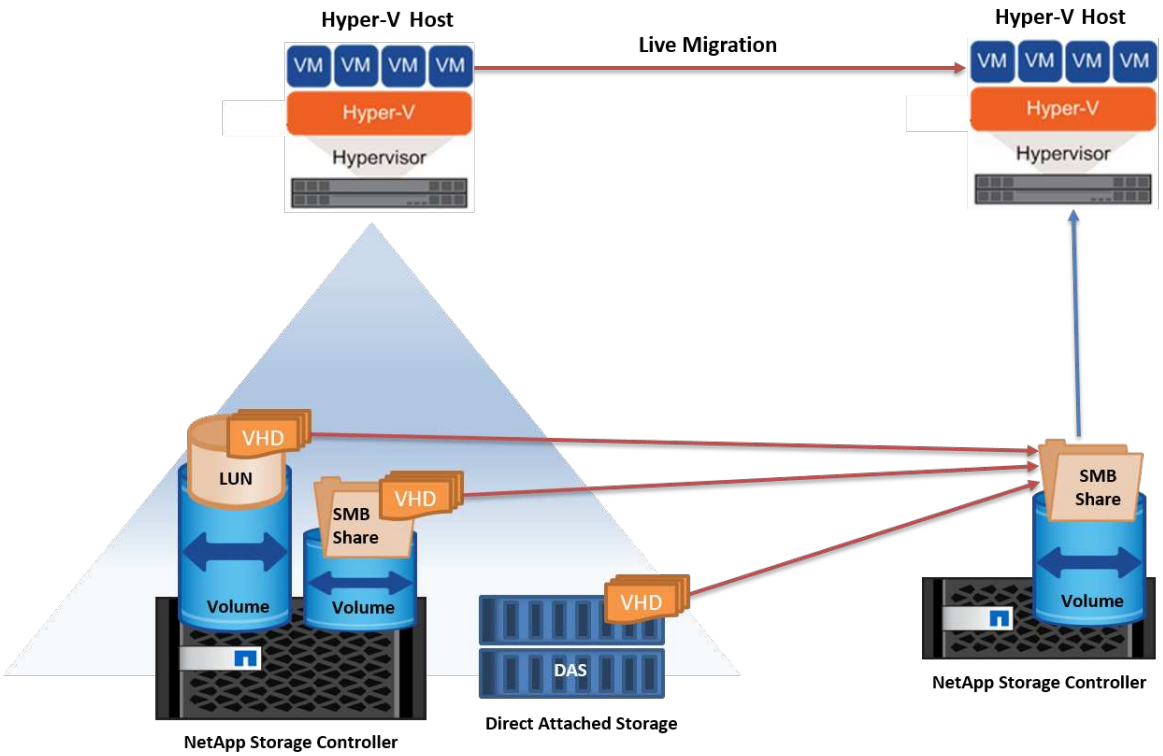
Live Migration outside a clustered environment

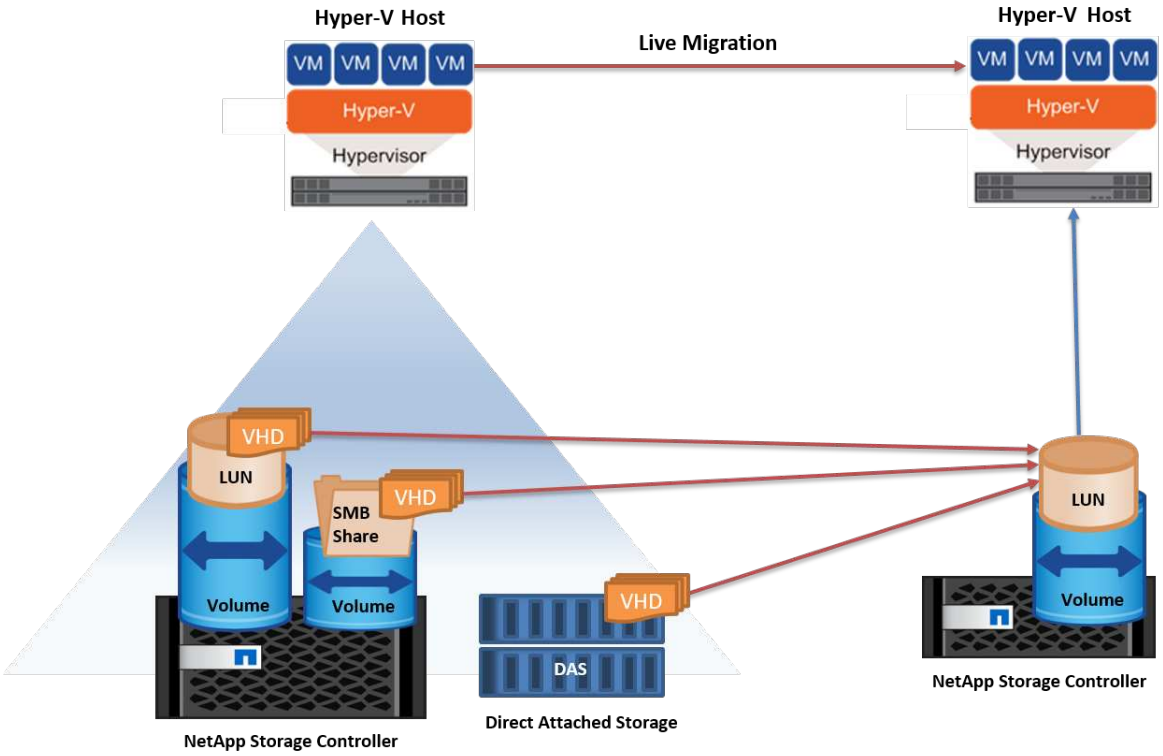
You can live migrate a VM between two nonclustered, independent Hyper-V servers. This process can use either shared or shared nothing live migration.

- In shared live migration, the VM is stored on an SMB share. Therefore, when you live migrate a VM, the VM's storage remains on the central SMB share for instant access by the other node, as shown in the following Figure.



- In shared nothing live migration, each Hyper-V server has its own local storage (it can be an SMB share, a LUN, or DAS), and the VM's storage is local to its Hyper-V server. When a VM is live migrated, the VM's storage is mirrored to the destination server over the client network and then the VM is migrated. The VM stored on DAS, a LUN, or an SMB/CIFS share can be moved to an SMB/CIFS share on the other Hyper-V server, as shown in the following figure. It can also be moved to a LUN, as shown in the second figure.





Further reading

For information about deploying live migration outside a clustered environment, see Appendix D: Deploy Hyper-V Live Migration Outside of a Clustered Environment.

Hyper-V Storage Live Migration

During the lifetime of a VM, you might need to move the VM storage (VHD/VHDX) to a different LUN or SMB share. This might be required if the present LUN or share is running out of space or yielding lower than expected performance.

The LUN or the share that currently hosts the VM can run out of space, be repurposed, or provide reduced performance. Under these circumstances, the VM can be moved without downtime to another LUN or share on a different volume, aggregate, or cluster. This process is faster if the storage system has copy-offload capabilities. NetApp storage systems are copy-offload enabled by default for CIFS and SAN environments.

The ODX feature performs full-file or sub-file copies between two directories residing on remote servers. A copy is created by copying data between the servers (or the same server if both the source and the destination files are on the same server). The copy is created without the client reading the data from the source or writing to the destination. This process reduces processor and memory use for the client or server and minimizes network I/O bandwidth. The copy is faster if its within same volume. If copy is across volumes, there may not be significant performance gains compared to host-based copies. Before proceeding with a copy operation on the host, confirm that the copy offload settings are configured on the storage system.

When VM storage live migration is initiated from a host, the source and the destination are identified, and the copy activity is offloaded to the storage system. Because the activity is performed by the storage system, there is negligible use of the host CPU, memory, or network.

NetApp storage controllers support the following different ODX scenarios:

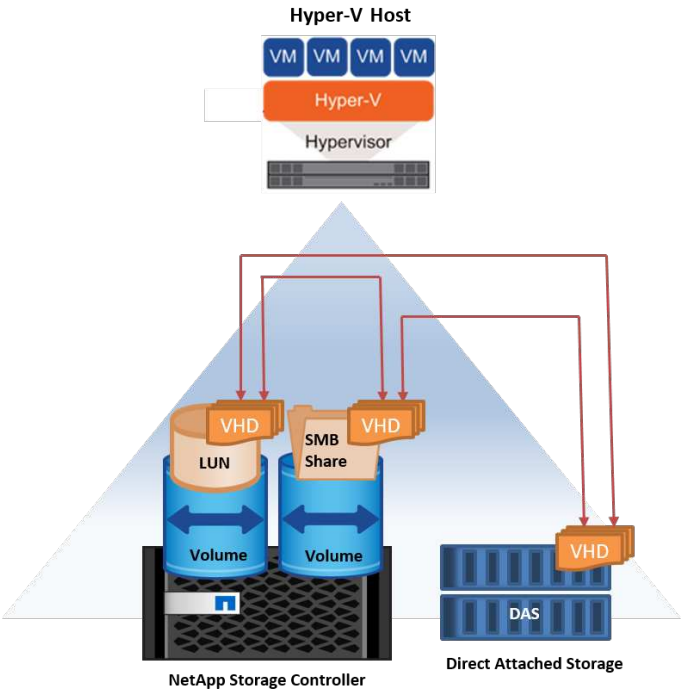
- **IntraSVM.** The data is owned by the same SVM:
- **Intravolume, intranode.** The source and destination files or LUNs reside within the same volume. The copy is performed with FlexClone file technology, which provides additional remote copy performance benefits.
- **Intervolume, intranode.** The source and destination files or LUNs are on different volumes that are on the same node.
- **Intervolume, internodes.** The source and destination files or LUNs are on different volumes that are located on different nodes.
- **InterSVM.** The data is owned by different SVMs.
- **Intervolume, intranode.** The source and destination files or LUNs are on different volumes that are on the same node.
- **Intervolume, internodes.** The source and destination files or LUNs are on different volumes that are on different nodes.
- **Intercluster.** Beginning with ONTAP 9.0, ODX is also supported for intercluster LUN transfers in SAN environments. Intercluster ODX is supported for SAN protocols only, not for SMB.

After the migration is complete, the backup and replication policies must be reconfigured to reflect the new volume holding the VMs. Any previous backups that were taken cannot be used.

VM storage (VHD/VHDX) can be migrated between the following storage types:

- DAS and the SMB share
- DAS and LUN
- An SMB share and a LUN
- Between LUNs

- Between SMB shares

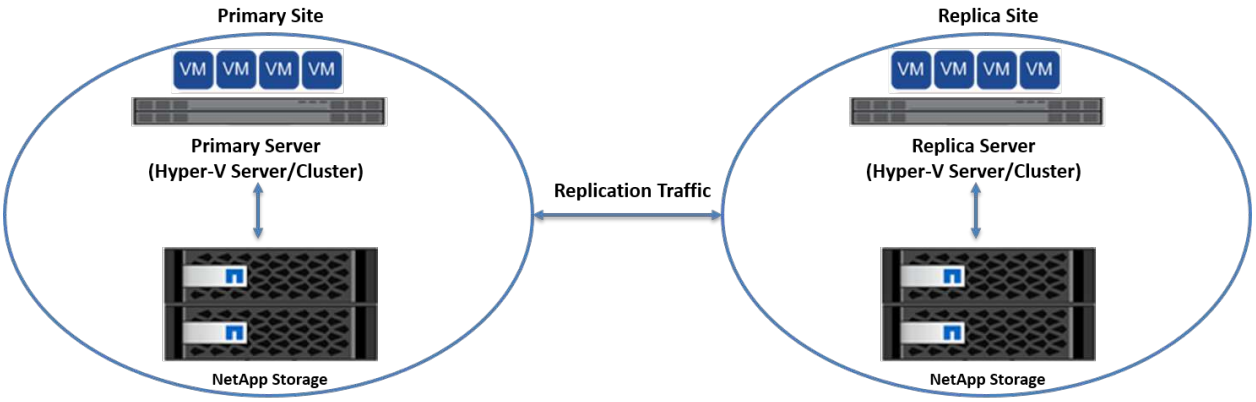


Further reading

For information about deploying storage live migration, see Appendix E: Deploy Hyper-V Storage Live Migration.

Hyper-V Replica: Disaster recovery for virtual machines

Hyper-V Replica replicates the Hyper-V VMs from a primary site to replica VMs on a secondary site, asynchronously providing disaster recovery for the VMs. The Hyper-V server at the primary site hosting the VMs is known as the primary server; the Hyper-V server at the secondary site that receives replicated VMs is known as the replica server. A Hyper-V Replica example scenario is shown in the following figure. You can use Hyper-V Replica for VMs between Hyper-V servers that are part of a failover cluster or between independent Hyper-V servers that are not part of any cluster.



Replication

After Hyper-V Replica is enabled for a VM on the primary server, initial replication creates an identical VM on the replica server. After the initial replication, Hyper-V Replica maintains a log file for the VHDs of the VM. The

log file is replayed in reverse order to the replica VHD in accordance with the replication frequency. This log and the use of reverse order make sure that the latest changes are stored and replicated asynchronously. If replication does not occur in line with the expected frequency, an alert is issued.

Extended replication

Hyper-V Replica supports extended replication in which a secondary replica server can be configured for disaster recovery. A secondary replica server can be configured for the replica server to receive the changes on the replica VMs. In an extended replication scenario, the changes on the primary VMs on the primary server are replicated to the replica server. Then the changes are replicated to the extended replica server. The VMs can be failed over to the extended replica server only when both primary and replica servers go down.

Failover

Failover is not automatic; the process must be manually triggered. There are three types of failover:

- **Test failover.** This type is used to verify that a replica VM can start successfully on the replica server and is initiated on the replica VM. This process creates a duplicate test VM during failover and does not affect regular production replication.
- **Planned failover.** This type is used to fail over VMs during planned downtime or expected outages. This process is initiated on the primary VM, which must be turned off on the primary server before a planned failover is run. After the machine fails over, Hyper-V Replica starts the replica VM on the replica server.
- **Unplanned failover.** This type is used when unexpected outages occur. This process is initiated on the replica VM and should be used only if the primary machine fails.

Recovery

When you configure replication for a VM, you can specify the number of recovery points. Recovery points represent points in time from which data can be recovered from a replicated machine.

Further reading

- For information about deploying Hyper-V Replica outside a clustered environment, see the section "Deploy Hyper-V Replica Outside of a Clustered Environment."
- For information about deploying Hyper-V Replica in a clustered environment, see the section "Deploy Hyper-V Replica in a Clustered Environment."

Storage efficiency

ONTAP provides industry leading storage efficiency for virtualized environments including Microsoft Hyper-V. NetApp also offers storage efficiency guarantee programs.

NetApp deduplication

NetApp deduplication works by removing duplicate blocks at the storage volume level, storing only one physical copy, regardless of how many logical copies are present. Therefore, deduplication creates the illusion that there are numerous copies of that block. Deduplication automatically removes duplicate data blocks on a 4KB block level across an entire volume. This process reclaims storage to achieve space and potential performance savings by reducing the number of physical writes to the disk. Deduplication can provide more than 70% space savings in Hyper-V environments.

Thin provisioning

Thin provisioning is an efficient way to provision storage because the storage is not preallocated up front. In other words, when a volume or LUN is created using thin provisioning, the space on the storage system is unused. The space remains unused until the data is written to the LUN or volume and only the necessary space to store the data is used. NetApp recommends enabling thin provisioning on the volume and disabling LUN reservation.

Quality of Service

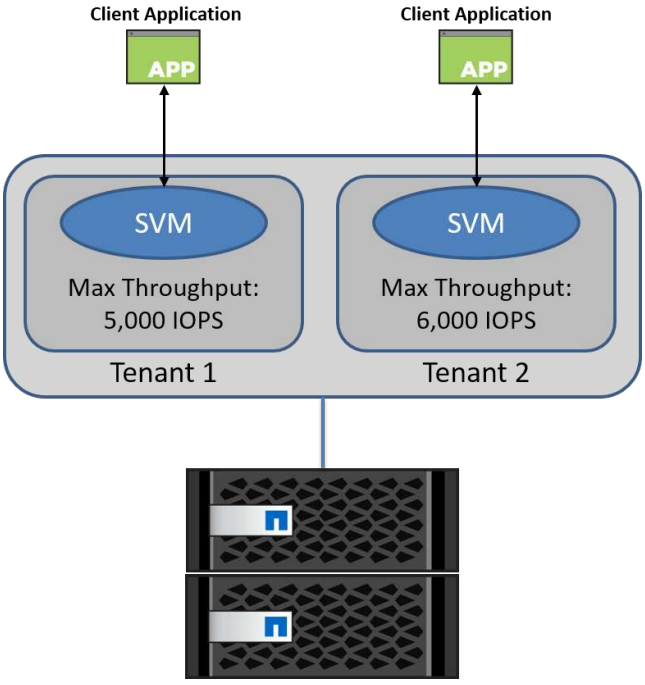
Storage QoS in clustered ONTAP enables you to group storage objects and set throughput limits on the group. Storage QoS can be used to limit the throughput to workloads and to monitor workload performance. With this ability, a storage administrator can separate workloads by organization, application, business unit, or production or development environments.

In enterprise environments, storage QoS helps to achieve the following:

- Prevents user workloads from affecting each other.
- Protects critical applications that have specific response times that must be met in IT-as-a-service (ITaaS) environments.
- Prevents tenants from affecting each other.
- Avoids performance degradation with the addition of each new tenant.

QoS allows you to limit the amount of I/O sent to an SVM, a flexible volume, a LUN, or a file. I/O can be limited by the number of operations or the raw throughput.

The following figure illustrates SVM with its own QoS policy enforcing a maximum throughput limit.



To configure an SVM with its own QoS policy and monitor policy group, run the following commands on your ONTAP cluster:

```
# create a new policy group pg1 with a maximum throughput of 5,000 IOPS
cluster::> qos policy-group create pg1 -vserver vs1 -max-throughput
5000iops
```

```
# create a new policy group pg2 without a maximum throughput
cluster::> qos policy-group create pg2 -vserver vs2
```

```
# monitor policy group performance
cluster::> qos statistics performance show
```

```
# monitor workload performance
cluster::> qos statistics workload performance show
```

Security

ONTAP provides a secure storage system for the Windows operating system.

Windows Defender Antivirus

Windows Defender is antimalware software installed and enabled on Windows Server by default. This software actively protects Windows Server against known malware and can regularly update antimalware definitions through Windows Update. NetApp LUNs and SMB shares can be scanned using Windows Defender.

Further reading

For further information, see the Windows Defender Overview.

BitLocker

BitLocker drive encryption is a data protection feature continued from Windows Server 2012. This encryption protects physical disks, LUNs, and CSVs.

Best practice

Before enabling BitLocker, the CSV must be put into maintenance mode. Therefore, NetApp recommends that decisions pertaining to BitLocker-based security be made before creating VMs on the CSV to avoid downtime.

Deploy Nano server

Learn about deploying Microsoft Windows Nano Server.

Deployment

To deploy a Nano Server as a Hyper-V host, complete the following steps:

1. Log in to Windows Server as a member of the administrator group.
2. Copy the NanoServerImageGenerator folder from the \NanoServer folder in the Windows Server ISO to the local hard drive.
3. To create a Nano Server VHD/VHDX, complete the following steps:
   a. Start Windows PowerShell as an administrator, navigate to the copied NanoServerImageGenerator folder on the local hard drive, and run the following cmdlet:

   ```
   Set-ExecutionPolicy RemoteSigned
   Import-Module .\NanoServerImageGenerator -Verbose
   ```

   b. Create a VHD for the Nano Server as a Hyper-V host by running the following PowerShell cmdlet. This command prompts you for an administrator password for the new VHD.

   ```
   New-NanoServerImage -Edition Standard -DeploymentType Guest
   -MediaPath <"input the path to the root of the contents of Windows
   Server 2016 ISO"> -TargetPath <"input the path, including the
   filename and extension where the resulting VHD/VHDX will be created">
   -ComputerName <"input the name of the nano server computer you are
   about to create"> -Compute
   ```

   c. In the following example, we create a Nano Server VHD with the feature Hyper-V host with failover clustering enabled. This example creates a Nano Server VHD from an ISO mounted at f:\. The newly created VHD is placed in a folder named NanoServer in the folder from where the cmdlet is run. The computer name is NanoServer and the resulting VHD contains the standard edition of Windows Server.

   ```
   New-NanoServerImage -Edition Standard -DeploymentType Guest
   -MediaPath f:\ -TargetPath .\NanoServer.vhd -ComputerName NanoServer
   -Compute -Clustering
   ```

   d. With the cmdlet New-NanoServerImage, configure parameters that set the IP address, the subnet mask, the default gateway, the DNS server, the domain name, and so on.
4. Use the VHD in a VM or physical host to deploy Nano Server as a Hyper-V host:
   a. For deployment on a VM, create a new VM in Hyper-V Manager and use the VHD created in Step 3.
   b. For deployment on a physical host, copy the VHD to the physical computer and configure it to boot from this new VHD. First, mount the VHD, run bcdboot e:\windows (where the VHD is mounted under E:\), unmount the VHD, restart the physical computer, and boot to the Nano Server.
5. Join the Nano Server to a domain (optional):
   a. Log in to any computer in the domain and create a data blob by running the following PowerShell cmdlet:

```
$domain = "<input the domain to which the Nano Server is to be
joined>"
$nanoserver = "<input name of the Nano Server>"
```

```
djoin.exe /provision /domain $domain /machine $nanoserver /savefile
C:\temp\odjblob /reuse
```

b. Copy the odjblob file to the Nano Server by running the following PowerShell cmdlets on a remote machine:

```
$nanoserver = "<input name of the Nano Server>"
$nanouname = ""<input username of the Nano Server>"
$nanopwd = ""<input password of the Nano Server>"
```

```
$filePath = 'c:\temp\odjblob'
$fileContents = Get-Content -Path $filePath -Encoding Unicode
```

```
$securenanopwd = ConvertTo-SecureString -AsPlainText -Force $nanopwd
$nanosecurecred = new-object management.automation.pscredential
$nanouname, $securenanopwd
```

```
Invoke-Command -VMName $nanoserver -Credential $nanosecurecred
-ArgumentList @($filePath,$fileContents) -ScriptBlock \{
    param($filePath,$data)
    New-Item -ItemType directory -Path c:\temp
    Set-Content -Path $filePath -Value $data -Encoding Unicode
    cd C:\temp
    djoin /requestodj /loadfile c:\temp\odjblob /windowspath
c:\windows /localos
}
```

c. Reboot the Nano Server.

Connect to Nano Server

To connect to the Nano Server remotely using PowerShell, complete the following steps:

1. Add the Nano Server as a trusted host on the remote computer by running the following cmdlet on the remote server:

```
Set-Item WSMan:\LocalHost\Client\TrustedHosts "<input IP Address of the
Nano Server>"
```

2. If the environment is safe and if you want to set all the hosts to be added as trusted hosts on a server, run the following command:

```
Set-Item WSMan:\LocalHost\Client\TrustedHosts *
```

3. Start the remote session by running the following cmdlet on the remote server. Provide the password for the Nano Server when prompted.

```
Enter-PSSession -ComputerName "<input IP Address of the Nano Server>"
-Credential ~\Administrator
```

To connect to the Nano Server remotely using GUI management tools from a remote Windows Server, complete the following commands:

4. Log in to the Windows Server as a member of the administrator group.
5. Start Server Manager**.**
6. To manage a Nano Server remotely from Server Manager, right-click All Servers, click Add Servers, provide the Nano Server's information, and add it. You can now see the Nano Server listed in the server list. Select the Nano Server, right-click it, and start managing it with the various options provided.
7. To manage services on a Nano Server remotely, complete the following steps:
   a. Open Services from the Tools section of Server Manager.
   b. Right-click Services (Local).
   c. Click Connect to Server.
   d. Provide the Nano Server details to view and manage the services on the Nano Server.
8. If the Hyper-V role is enabled on the Nano Server, complete the following steps to manage it remotely from Hyper-V Manager:
   a. Open Hyper-V Manager from the Tools section of Server Manager.
   b. Right-click Hyper-V Manager.
   c. Click Connect to Server and provide the Nano Server details. Now the Nano Server can be managed as a Hyper-V server to create and manage VMs on top of it.
9. If the failover clustering role is enabled on the Nano Server, complete the following steps to manage it remotely from the failover cluster manager:
   a. Open Failover Cluster Manager from the Tools section of Server Manager.
   b. Perform clustering-related operations with the Nano Server.

Deploy Hyper-V cluster

This appendix describes deploying a Hyper-V cluster.

Prerequisites

- At least two Hyper-V servers exist connected to each other.
- At least one virtual switch is configured on each Hyper-V server.
- The failover cluster feature is enabled on each Hyper-V server.
- SMB shares or CSVs are used as shared storage to store VMs and their disks for Hyper-V clustering.
- Storage should not be shared between different clusters. You should have only one CSV/CIFS share per cluster.
- If the SMB share is used as shared storage, then permissions on the SMB share must be configured to grant access to the computer accounts of all the Hyper-V servers in the cluster.

Deployment

1. Log in to one of the Windows Hyper-V servers as a member of the administrator group.
2. Start Server Manager**.**
3. In the Tools section, click Failover Cluster Manager.
4. Click the Create Cluster from Actions menu.
5. Provide details for the Hyper-V server that is part of this cluster.
6. Validate the cluster configuration. Select Yes when prompted for cluster configuration validation and select the tests required to validate whether the Hyper-V servers pass the prerequisites to be part of the cluster.
7. After validation succeeds, the Create Cluster wizard is started. In the wizard, provide the cluster name and the cluster IP address for the new cluster. A new failover cluster is then created for the Hyper-V server.
8. Click the newly created cluster in Failover Cluster Manager and manage it.
9. Define shared storage for the cluster to use. It can be either an SMB share or a CSV.
10. Using an SMB share as shared storage requires no special steps.
    - Configure a CIFS share on a NetApp storage controller. To do so, see the section "Provisioning in SMB Environments".
11. To use a CSV as shared storage, complete the following steps:
    a. Configure LUNs on a NetApp storage controller. To do so, see the section "Provisioning in SAN Environments."
    b. Make sure that all the Hyper-V servers in the failover cluster can see the NetApp LUNs. To do this for all the Hyper-V servers that are part of the failover cluster, make sure that their initiators are added to the initiator group on NetApp storage. Also be sure that their LUNs are discovered and make sure that MPIO is enabled.
    c. On any one of the Hyper-V servers in the cluster, complete the following steps:
        i. Take the LUN online, initialize the disk, create a new simple volume, and format it using NTFS or ReFS.
        ii. In Failover Cluster Manager, expand the cluster, expand Storage, right-click Disks, and then click Add Disks. Doing so opens the Add Disks to a Cluster wizard showing the LUN as a disk. Click OK to add the LUN as a disk.
        iii. Now the LUN is named Clustered Disk and is shown as Available Storage under Disks.
    d. Right-click the LUN (Clustered Disk) and click Add to Cluster Shared Volumes. Now the LUN is shown as a CSV.

e. The CSV is simultaneously visible and accessible from all the Hyper-V servers of the failover cluster at its local location C:\ClusterStorage\.

12. Create a highly available VM:
    a. In Failover Cluster Manager, select and expand the cluster created previously.
    b. Click Roles and then click Virtual Machines in Actions. Click New Virtual Machine.
    c. Select the node from the cluster where the VM should reside.
    d. In the Virtual Machine Creation wizard, provide the shared storage (SMB share or CSV) as the path to store the VM and its disks.
    e. Use Hyper-V Manager to set the shared storage (SMB share or CSV) as the default path to store the VM and its disks for a Hyper-V server.
13. Test planned failover. Move VMs to another node using live migration, quick migration, or storage migration (move). Review Live Migration in a Clustered Environment for more details.
14. Test unplanned failover. Stop cluster service on the server owning the VM.

Deploy Hyper-V Live Migration in a clustered environment

This appendix describes deploying live migration in a clustered environment.

Prerequisites

To deploy live migration, you need to have Hyper-V servers configured in a failover cluster with shared storage. Review Deploy Hyper-V Cluster for more details.

Deployment

To use live migration in a clustered environment, complete the following steps:

1. In Failover Cluster Manager, select and expand the cluster. If the cluster is not visible, then click Failover Cluster Manager, click Connect to Cluster, and provide the cluster name.
2. Click Roles, which lists all the VMs available in a cluster.
3. Right-click on the VM and click Move. Doing so provides you with three options:
    - **Live migration.** You can select a node manually or allow the cluster to select the best node. In live migration, the cluster copies the memory used by the VM from the current node to another node. Therefore, when the VM is migrated to another node, the memory and state information needed by the VM are already in place for the VM. This migration method is nearly instantaneous, but only one VM can be live migrated at a time.
    - **Quick migration.** You can select a node manually or allow the cluster to select the best node. In quick migration, the cluster copies the memory used by a VM to a disk in storage. Therefore, when the VM is migrated to another node, the memory and state information needed by the VM can be quickly read from the disk by the other node. With quick migration, multiple VMs can be migrated simultaneously.
    - **Virtual machine storage migration.** This method uses the Move Virtual Machine Storage wizard. With this wizard, you can select the VM disk along with other files to be moved to another location, which can be a CSV or an SMB share.

Deploy Hyper-V Live Migration outside a clustered environment

This section describes the deployment of Hyper-V live migration outside a clustered

environment.

Prerequisites

- Standalone Hyper-V servers with independent storage or shared SMB storage.
- The Hyper-V role installed on both the source and destination servers.
- Both Hyper-V servers belong to the same domain or to domains that trust each other.

Deployment

To perform live migration in a non-clustered environment, configure source and destination Hyper-V servers so that they can send and receive live migration operations. On both Hyper-V servers, complete the following steps:

1. Open Hyper-V Manager from the Tools section of Server Manager.
2. In Actions, click Hyper-V Settings.
3. Click Live Migrations and select Enable Incoming and Outgoing Live Migrations.
4. Choose whether to allow live migration traffic on any available network or only on specific networks.
5. Optionally, you can configure the authentication protocol and performance options from the Advanced section of Live Migrations.
6. If CredSSP is used as the authentication protocol, then make sure to log in to the source Hyper-V server from the destination Hyper-V server before moving the VM.
7. If Kerberos is used as the authentication protocol, then configure the constrained delegation. Doing so requires Active Directory domain controller access. To configure the delegation, complete the following steps:
   a. Log in to the Active Directory domain controller as the administrator.
   b. Start Server Manager.
   c. In the Tools section, click Active Directory Users and Computers.
   d. Expand the domain and click Computers.
   e. Select the source Hyper-V server from the list, right-click it, and click Properties.
   f. In the Delegation tab, select Trust This Computer for Delegation to Specified Services Only.
   g. Select Use Kerberos Only.
   h. Click Add, which opens the Add Services wizard.
   i. In Add Services, click Users and Computers, which opens Select Users or Computers**.**
   j. Provide the destination Hyper-V server name and click OK.
      - To move VM storage, select CIFS.
      - To move VMs, select the Microsoft Virtual System Migration service.
   k. In the Delegation tab, click OK.
   l. From the Computers folder, select the destination Hyper-V server from the list and repeat the process. In Select Users or Computers, provide the source Hyper-V server name.
8. Move the VM.
   a. Open Hyper-V Manager.

b. Right-click a VM and click Move.
c. Choose Move the Virtual Machine.
d. Specify the destination Hyper-V server for the VM.
e. Choose the move options. For Shared Live Migration, choose Move Only the Virtual Machine. For Shared Nothing Live Migration, choose any of the other two options based on your preferences.
f. Provide the location for the VM on the destination Hyper-V server based on your preferences.
g. Review the summary and click OK to move the VM.

Deploy Hyper-V torage Live Migration

Learn how to configure Hyper-V storage live migration

Prerequisites

- You must have a standalone Hyper-V server with independent storage (DAS or LUN) or SMB storage (local or shared among other Hyper-V servers).
- The Hyper-V server must be configured for live migration. Review the section on deployment in Live Migration Outside of a Clustered Environment.

Deployment

1. Open Hyper-V Manager.
2. Right-click a VM and click Move.
3. Select Move the Virtual Machine's Storage.
4. Select options for moving the storage based on your preferences.
5. Provide the new location for the VM's items.
6. Review the summary and click OK to move the VM's storage.

Deploy Hyper-V Replica outside a clustered environment

This appendix describes deploying Hyper-V Replica outside a clustered environment.

Prerequisites

- You need standalone Hyper-V servers located in the same or separate geographical locations serving as primary and replica servers.
- If separate sites are used, then the firewall at each site must be configured to allow communication between the primary and replica servers.
- The replica server must have enough space to store the replicated workloads.

Deployment

1. Configure the replica server.
   a. So that the inbound firewall rules allow incoming replication traffic, run the following PowerShell cmdlet:

```
Enable-Netfirewallrule -displayname "Hyper-V Replica HTTP Listener
(TCP-In)"
```

   b. Open Hyper-V Manager from the Tools section of Server Manager.
   c. Click Hyper-V Settings from Actions.
   d. Click Replication Configuration and select Enable This Computer as a Replica Server.
   e. In the Authentication and Ports section, select the authentication method and port.
   f. In the Authorization and Storage section, specify the location to store the replicated VMs and files.
2. Enable VM replication for VMs on the primary server. VM replication is enabled on a per-VM basis and not for the entire Hyper-V server.
   a. In Hyper-V Manager, right-click a VM and click Enable Replication to open the Enable Replication wizard.
   b. Provide the name of the replica server where the VM must be replicated.
   c. Provide the authentication type and the replica server port that was configured to receive replication traffic on the replica server.
   d. Select the VHDs to be replicated.
   e. Choose the frequency (duration) at which the changes are sent to the replica server.
   f. Configure recovery points to specify the number of recovery points to maintain on the replica server.
   g. Choose Initial Replication Method to specify the method to transfer the initial copy of the VM data to the replica server.
   h. Review the summary and click Finish.
   i. This process creates a VM replica on the replica server.

Replication

1. Run a test failover to make sure that the replica VM functions properly on the replica server. The test creates a temporary VM on the replica server.
   a. Log in to the replica server.
   b. In Hyper-V Manager, right-click a replica VM, click Replication, and click Test Failover.
   c. Choose the recovery point to use.
   d. This process creates a VM of the same name appended with -Test.
   e. Verify the VM to make sure that everything works well.
   f. After failover, the replica test VM is deleted if you select Stop Test Failover for it.
2. Run a planned failover to replicate the latest changes on the primary VM to the replica VM.
   a. Log in to the primary server.
   b. Turn off the VM to be failed over.
   c. In Hyper-V Manager, right-click the turned-off VM, click Replication, and click Planned Failover.
   d. Click Failover to transfer the latest VM changes to the replica server.
3. Run an unplanned failover in the case of primary VM failure.
   a. Log in to the replica server.

b. In Hyper-V Manager, right-click a replica VM, click Replication, and click Failover.
c. Choose the recovery point to use.
d. Click Failover to fail over the VM.

Deploy Hyper-V replica in a clustered environment

Learn how to deploy and configure Hyper-V replica with Windows Server Failover Cluster.

Prerequisites

- You need to have Hyper-V clusters located in the same or in separate geographical locations serving as primary and replica clusters. Review Deploy Hyper-V Cluster for more details.
- If separate sites are used, then the firewall at each site must be configured to allow communication between the primary and replica clusters.
- The replica cluster must have enough space to store the replicated workloads.

Deployment

1. Enable firewall rules on all the nodes of a cluster. Run the following PowerShell cmdlet with admin privileges on all the nodes in both the primary and replica clusters.

```
# For Kerberos authentication
get-clusternode | ForEach-Object \{Invoke-command -computername $_.name
-scriptblock \{Enable-Netfirewallrule -displayname "Hyper-V Replica HTTP
Listener (TCP-In)"}}
```

```
# For Certificate authentication
get-clusternode | ForEach-Object \{Invoke-command -computername $_.name
-scriptblock \{Enable-Netfirewallrule -displayname "Hyper-V Replica
HTTPS Listener (TCP-In)"}}
```

2. Configure the replica cluster.
   a. Configure the Hyper-V Replica broker with a NetBIOS name and IP address to use as the connection point to the cluster that is used as the replica cluster.
      i. Open Failover Cluster Manager.
      ii. Expand the cluster, click Roles, and click the Configure Role from Actions pane.
      iii. Select Hyper-V Replica Broker in the Select Role page.
      iv. Provide the NetBIOS name and IP address to be used as the connection point to the cluster (client access point).
      v. This process creates a Hyper-V Replica broker role. Verify that it comes online successfully.
   b. Configure replication settings.
      i. Right-click the replica broker created in the previous steps and click Replication Settings.
      ii. Select Enable This Cluster as a Replica Server.

iii. In the Authentication and Ports section, select the authentication method and port.

iv. In the authorization and storage section, select the servers allowed to replicate VMs to this cluster. Also, specify the default location where the replicated VMs are stored.

Replication

Replication is similar to the process described in the section Replica Outside a Clustered Environment.

Where to find additional information

Additional resources for Microsoft Windows and Hyper-V

- ONTAP concepts
  https://docs.netapp.com/us-en/ontap/concepts/introducing-data-management-software-concept.html
- Best practices for modern SAN
  https://www.netapp.com/media/10680-tr4080.pdf
- NetApp All-SAN Array Data Availability and Integrity with the NetApp ASA
  https://www.netapp.com/pdf.html?item=/media/85671-tr-4968.pdf
- SMB protocol best practices
  https://www.netapp.com/pdf.html?item=/media/10678-tr-4543pdf.pdf
- Getting Started with Nano Server
  https://technet.microsoft.com/library/mt126167.aspx
- What's New in Hyper-V on Windows Server
  https://technet.microsoft.com/windows-server-docs/compute/hyper-v/what-s-new-in-hyper-v-on-windows

Microsoft SQL Server

Microsoft SQL Server on ONTAP

ONTAP delivers an enterprise-class security and performance solution for your Microsoft SQL Server databases while also providing world-class tools to manage your environment.

> This documentation replaces the previously published technical report *TR-4590: Best practice guide for Microsoft SQL Server with ONTAP*

NetApp assumes that the reader has working knowledge of the following:

- ONTAP software
- NetApp SnapCenter as backup software, which includes:
  - SnapCenter Plug-in for Microsoft Windows
  - SnapCenter Plug-in for SQL Server
- Microsoft SQL Server architecture and administration

The scope of this best practices section is limited to technical design based on the design principles and preferred standards that NetApp recommends for storage infrastructure. The end-to-end implementation is out of the scope.

For configuration compatibility across the NetApp products, see the NetApp Interoperability Matrix Tool (IMT).

Microsoft SQL Server workloads

Before deploying SQL Server, you must understand the database workload requirements of the applications that your SQL Server instances support. Each application has different requirements for capacity, performance, and availability, and therefore each database should be designed to optimally support those requirements. Many organizations classify databases into multiple management tiers, using application requirements to define SLAs. SQL Server workloads can be described as follows:

- OLTP databases are often also the most critical databases in an organization. These databases usually back customer-facing applications and are considered essential to the company's core operations. Mission-critical OLTP databases and the applications they support often have SLAs that require high levels of performance and are sensitive to performance degradation and availability. They might also be candidates for Always On Failover Clusters or Always On Availability Groups. The I/O mix of these types of databases is usually characterized by 75% to 90% random read and 25% to 10% write.
- Decision support system (DSS) databases can be also referred to as data warehouses. These databases are mission critical in many organizations that rely on analytics for their business. These databases are sensitive to CPU utilization and read operations from disk when queries are being run. In many organizations, DSS databases are the most critical during the month, quarter, and year end. This workload typically has a 100% read I/O mix.

Database configuration

Microsoft SQL Server CPU configuration

To improve system performance, you need modify SQL Server settings and server configuration to use appropriate number of processors for execution.

Hyperthreading

Hyperthreading is Intel's proprietary simultaneous multithreading (SMT) implementation, which improves parallelization of computations (multitasking) performed on x86 microprocessors.

Hardware that uses hyperthreading allows the logical hyperthread CPUs to appear as physical CPUs to the operating system. SQL Server then sees the physical CPUs, which the operating system presents, and can use the hyperthreaded processors. This improves performance by increasing parallelization.

The caveat here is that each SQL Server version has its own limitations on the compute power it can use. For more information, see Compute Capacity Limits by Edition of SQL Server.

There are two options for licensing SQL Server. The first is known as a server + client access license (CAL) model; the second is the per processor core model. Although you can access all the product features available in SQL Server with the server + CAL strategy, there is a hardware limit of 20 CPU cores per socket. Even if you have SQL Server Enterprise Edition + CAL for a server with more than 20 CPU cores per socket, the application cannot use all those cores at a time on that instance.

The figure below shows the SQL Server log message after startup indicating the enforcement of the core limit.

Log entries indicate number of cores being used after SQL Server startup.

```
2017-01-11 07:16:30.71 Server      Microsoft SQL Server 2016
(RTM) - 13.0.1601.5 (X64)
     Apr 29 2016 23:23:58
     Copyright (c) Microsoft Corporation
     Enterprise Edition (64-bit) on Windows Server 2016
Datacenter 6.3 <X64> (Build 14393: )

2017-01-11 07:16:30.71 Server      UTC adjustment: -8:00
2017-01-11 07:16:30.71 Server      (c) Microsoft Corporation.
2017-01-11 07:16:30.71 Server      All rights reserved.
2017-01-11 07:16:30.71 Server      Server process ID is 10176.
2017-01-11 07:16:30.71 Server      System Manufacturer:
'FUJITSU', System Model: 'PRIMERGY RX2540 M1'.
2017-01-11 07:16:30.71 Server      Authentication mode is MIXED.
2017-01-11 07:16:30.71 Server      Logging SQL Server messages
in file 'C:\Program Files\Microsoft SQL Server
\MSSQL13.MSSQLSERVER\MSSQL\Log\ERRORLOG'.
2017-01-11 07:16:30.71 Server      The service account is 'SEA-
TM\FUJIA2R30$'. This is an informational message; no user action
is required.
2017-01-11 07:16:30.71 Server      Registry startup parameters:
      -d C:\Program Files\Microsoft SQL Server
\MSSQL13.MSSQLSERVER\MSSQL\DATA\master.mdf
      -e C:\Program Files\Microsoft SQL Server
\MSSQL13.MSSQLSERVER\MSSQL\Log\ERRORLOG
      -l C:\Program Files\Microsoft SQL Server
\MSSQL13.MSSQLSERVER\MSSQL\DATA\mastlog.ldf
      -T 3502
      -T 834
2017-01-11 07:16:30.71 Server      Command Line Startup
Parameters:
      -s "MSSQLSERVER"
2017-01-11 07:16:30.72 Server      SQL Server detected 2 sockets
with 18 cores per socket and 36 logical processors per socket,
72 total logical processors; using 40 logical processors based
on SQL Server licensing. This is an informational message; no
user action is required.
2017-01-11 07:16:30.72 Server      SQL Server is starting at
```

Therefore, to use all CPUs, you should use the per-processor core license. For detailed information about SQL Server licensing, see SQL Server 2022: Your modern data platform.

CPU affinity

You are unlikely to need to alter the processor affinity defaults unless you encounter performance problems, but it is still worth understanding what they are and how they work.

SQL Server supports processor affinity by two options:

- CPU affinity mask
- Affinity I/O mask

SQL Server uses all CPUs available from the operating system (if the per-processor core license is chosen). It creates schedulers on all the CPUs to make best use of the resources for any given workload. When multitasking, the operating system or other applications on the server can switch process threads from one processor to another. SQL Server is a resource-intensive application, and performance can be affected when this occurs. To minimize the impact, you can configure the processors so that all of the SQL Server load is directed to a preselected group of processors. This is achieved by using the CPU affinity mask.

The affinity I/O mask option binds SQL Server disk I/O to a subset of CPUs. In SQL Server OLTP environments, this extension can enhance the performance of SQL Server threads issuing I/O operations.

Max Degree of Parallelism (MAXDOP)

By default, SQL Server uses all available CPUs during query execution if the per-processor core license chosen.

Although this is helpful for large queries, it can cause performance problems and limit concurrency. A better approach is to limit parallelism to the number of physical cores in a single CPU socket. For example, on a server with two physical CPU sockets with 12 cores per socket, regardless of hyperthreading, MAXDOP should be set to 12. MAXDOP cannot restrict or dictate which CPU is to be used. Instead, it restricts the number of CPUs that can be used by a single batch query.

NetApp recommends for DSS such as data warehouses, start with MAXDOP at 50 and explore tuning up or down if required. Make sure you measure the critical queries in your application when making changes.

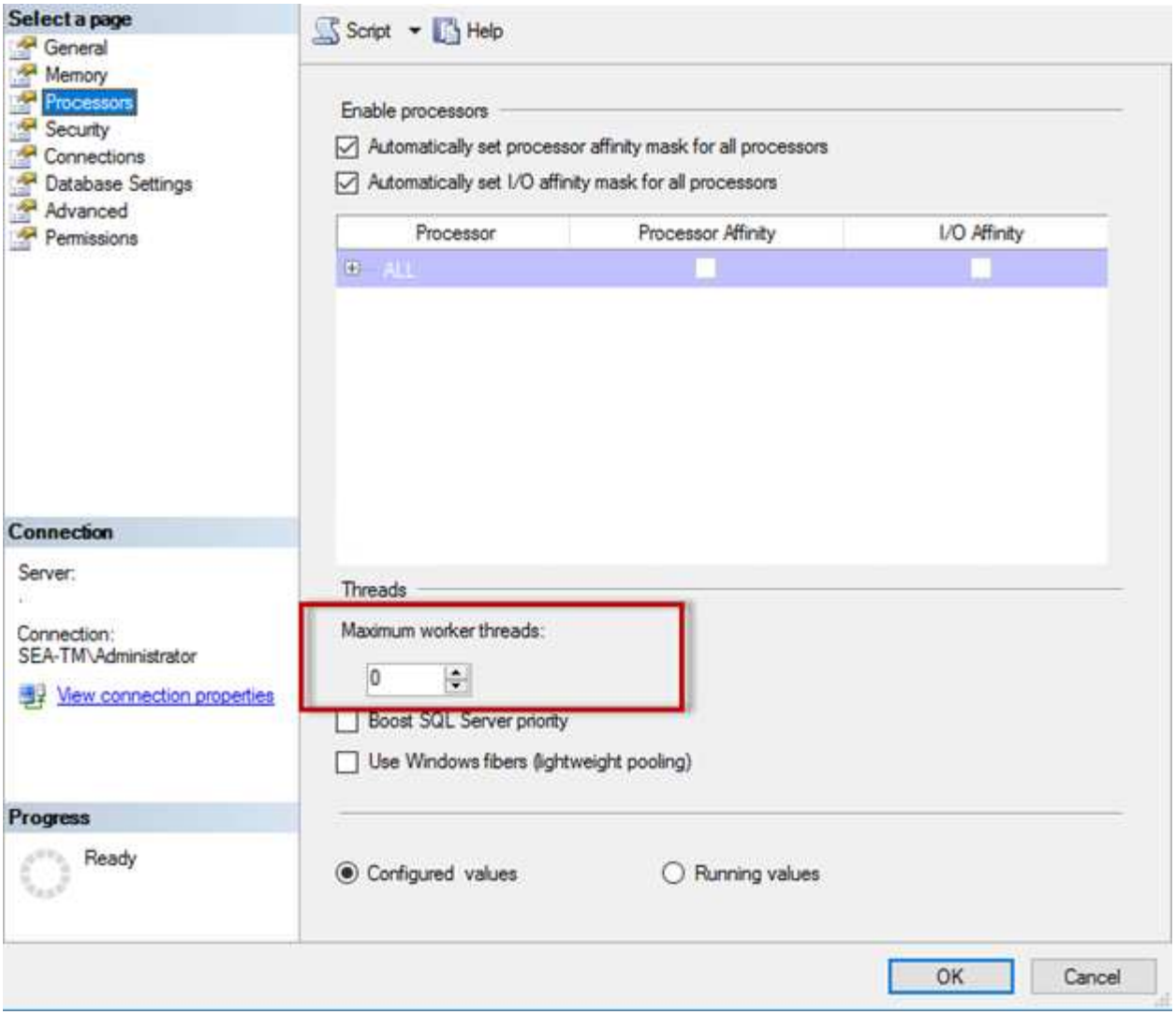
Max worker threads

The max worker threads option helps to optimize performance when large numbers of clients are connected to SQL Server.

Normally, a separate operating system thread is created for each query request. If hundreds of simultaneous connections are made to SQL Server, then one thread per query request consumes large amounts of system resources. The max worker threads option helps improve performance by enabling SQL Server to create a pool of worker threads to service a larger number of query requests.

The default value is 0, which allows SQL Server to automatically configure the number of worker threads at startup. This works for most systems. Max worker threads is an advanced option and should not be altered without assistance from an experienced database administrator (DBA).

When should you configure SQL Server to use more worker threads? If the average work queue length for each scheduler is above 1, you might benefit from adding more threads to the system, but only if the load is not CPU-bound or experiencing any other heavy waits. If either of those is happening, adding more threads does not help because they end up waiting for other system bottlenecks. For more information about max worker threads, see Configure the max worker threads Server Configuration Option.



Configuring max worker threads using SQL Server Management Studio.

```
The following example shows how to configure the max work threads option
using T-SQL.

EXEC sp_configure 'show advanced options', 1;
GO
RECONFIGURE ;
GO
EXEC sp_configure 'max worker threads', 900 ;
GO
RECONFIGURE;
GO
```

Microsoft SQL Server memory configuration

The following section explain configuring SQL server memory settings to optimize database performance.

Max server memory

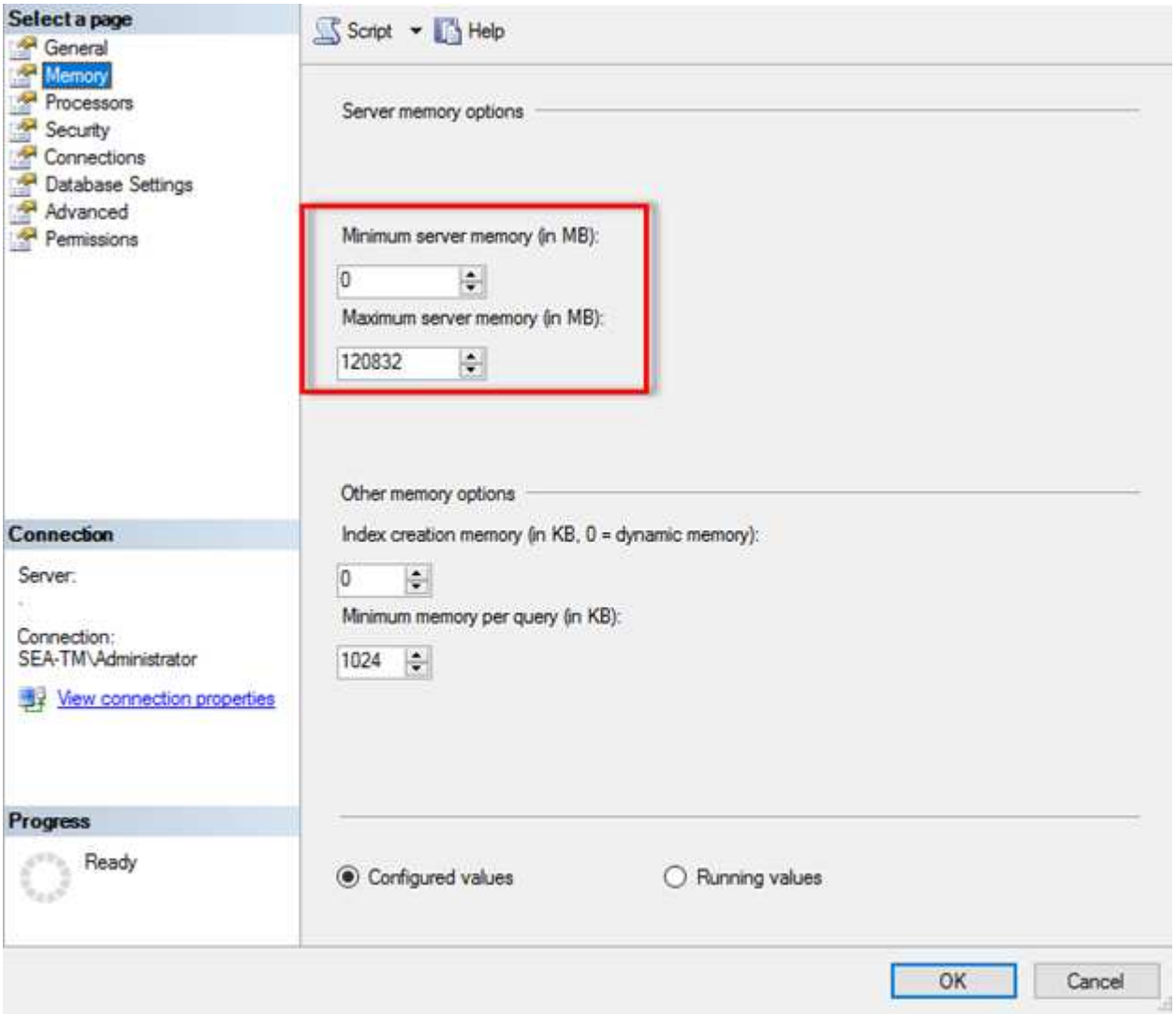
The max server memory option sets the maximum amount of memory that the SQL Server instance can use.

It is generally used if multiple applications are running on the same server where SQL Server is running and you want to guarantee that these applications have sufficient memory to function properly.

Some applications only use whatever memory is available when they start and do not request more even if needed. That is where the max server memory setting comes into play.

On a SQL Server cluster with several SQL Server instances, each instance could be competing for resources. Setting a memory limit for each SQL Server instance can help guarantee best performance for each instance.

NetApp recommends leaving at least 4GB to 6GB of RAM for the operating system to avoid performance issues.



Adjusting minimum and maximum server memory using SQL Server Management Studio.

Using SQL Server Management Studio to adjust minimum or maximum server memory requires a restart of the SQL Server service. You can adjust server memory using transact SQL (T-SQL) using this code:

```
EXECUTE sp_configure 'show advanced options', 1
GO
EXECUTE sp_configure 'min server memory (MB)', 2048
GO
EXEC sp_configure 'max server memory (MB)', 120832
GO
RECONFIGURE WITH OVERRIDE
```

Nonuniform memory access

Nonuniform memory access (NUMA) is a memory-access optimization method that helps increase processor speed without increasing the load on the processor bus.

If NUMA is configured on the server where SQL Server is installed, no additional configuration is required because SQL Server is NUMA aware and performs well on NUMA hardware.

Index create memory

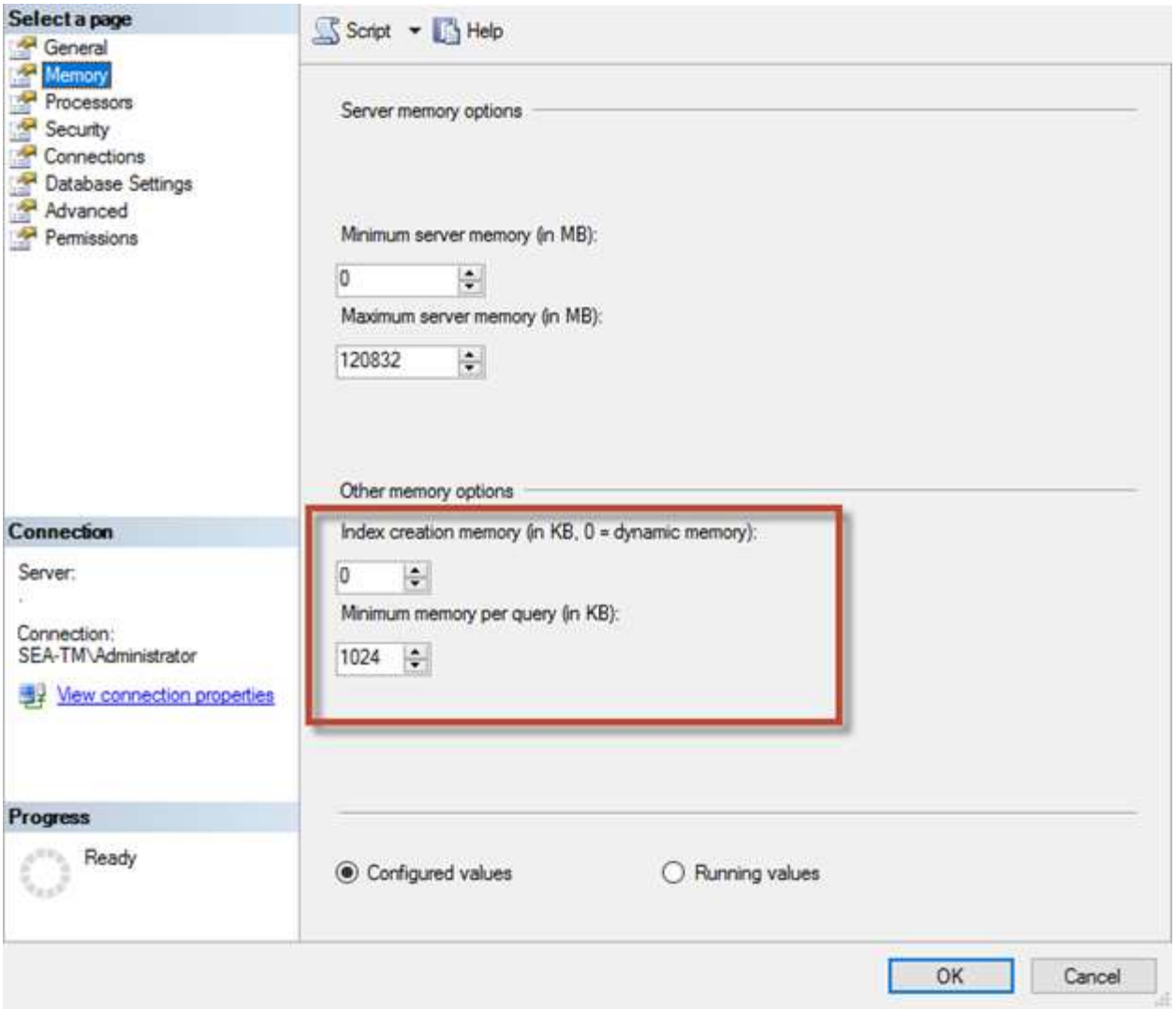
The index create memory option is another advanced option that you should not usually change.

It controls the maximum amount of RAM initially allocated for creating indexes. The default value for this option is 0, which means that it is managed by SQL Server automatically. However, if you encounter difficulties creating indexes, consider increasing the value of this option.

Min memory per query

When a query is run, SQL Server tries to allocate the optimum amount of memory to run efficiently.

By default, the min memory per query setting allocates >= to 1024KB for each query to run. It is a best practice is to leave this setting at the default value of 0 to allow SQL Server to dynamically manage the amount of memory allocated for index creation operations. However, if SQL Server has more RAM than it needs to run efficiently, the performance of some queries can be boosted if you increase this setting. Therefore, as long as memory is available on the server that is not being used by SQL Server, any other applications, or the operating system, then boosting this setting can help overall SQL Server performance. If no free memory is available, increasing this setting might hurt overall performance.



Buffer pool extensions

The buffer pool extension provides seamless integration of an NVRAM extension with the database engine buffer pool to significantly improve I/O throughput.

The buffer pool extension is not available in every SQL Server edition. It is available only with the 64-bit SQL Server Standard, Business Intelligence, and Enterprise editions.

The buffer pool extension feature extends the buffer pool cache with nonvolatile storage (usually SSDs). The extension allows the buffer pool to accommodate a larger database working set, forcing the paging of I/O between the RAM and the SSDs and effectively offloading small random I/Os from mechanical disks to SSDs. Because of the lower latency and better random I/O performance of SSDs, the buffer pool extension significantly improves I/O throughput.

The buffer pool extension feature offers the following benefits:

- Increased random I/O throughput
- Reduced I/O latency
- Increased transaction throughput
- Improved read performance with a larger hybrid buffer pool

- A caching architecture that can take advantage of existing and future low-cost memory

> **NetApp recommends** configuring the buffer pool extensions to:
>
> - Make sure that an SSD-backed LUN (such as NetApp AFF) is presented to the SQL Server host so that it can be used as a buffer pool extension target disk.
> - The extension file must be the same size as or larger than the buffer pool.

The following example shows a T-SQL command to set up a buffer pool extension of 32GB.

```
USE master
GO
ALTER SERVER CONFIGURATION
SET BUFFER POOL EXTENSION ON
  (FILENAME = 'P:\BUFFER POOL EXTENSION\SQLServerCache.BUFFER POOL
EXTENSION', SIZE = 32 GB);
GO
```

Microsoft SQL Server shared instance versus dedicated instance

Multiple SQL Server can be configured as a single instance per server or as multiple instances. The right decision usually depends on factors such as whether the server is to be used for production or development, whether the instance is considered critical to business operations and performance goals.

Shared instance configurations may be initially easier to configure, but it can lead to problems where resources become divided or locked, which in turn causes performance issues for other apps that have databases hosted on the shared SQL Server instance.

Troubleshooting performance issues can be complicated because you must figure out which instance is the root cause. This question is weighed against the costs of operating system licenses and SQL Server licenses. If application performance is paramount, then a dedicated instance is highly recommended.

Microsoft licenses SQL Server per core at the server level and not per instance. For this reason, database administrators are tempted to install as many SQL Server instances as the server can handle to save on licensing costs, which can lead to major performance issues later.

> **NetApp recommends** choosing dedicated SQL Server instances whenever possible to obtain optimal performance.

Storage configuration

Microsoft SQL Server storage considerations

The combination of ONTAP storage solutions and Microsoft SQL Server enables the creation of enterprise-level database storage designs that can meet today's most demanding application requirements.

To optimize both technologies, it is vital to understand the SQL Server I/O pattern and characteristics. A well-designed storage layout for a SQL Server database supports the performance of SQL Server and the management of the SQL Server infrastructure. A good storage layout also allows the initial deployment to be successful and the environment to grow smoothly over time as the business grows.

Data storage design

For SQL Server databases that do not use SnapCenter to perform backups, Microsoft recommends placing the data and log files on separate drives. For applications that simultaneously update and request data, the log file is write intensive, and the data file (depending on your application) is read/write intensive. For data retrieval, the log file is not needed. Therefore, requests for data can be satisfied from the data file placed on its own drive.

When you create a new database, Microsoft recommends specifying separate drives for the data and logs. To move files after the database is created, the database must be taken offline. For more Microsoft recommendations, see Place Data and Log Files on Separate Drives.

Aggregates

Aggregates are the lowest level storage containers for NetApp storage configurations. Some legacy documentation exists on the internet that recommends separating IO onto different sets of underlying drives. This is not recommended with ONTAP. NetApp has performed various I/O workload characterization tests using shared and dedicated aggregates with data files and transaction log files separated. The tests show that one large aggregate with more RAID groups and drives optimizes and improves storage performance and is easier for administrators to manage for two reasons:

- One large aggregate makes the I/O capabilities of all drives available to all files.
- One large aggregate enables the most efficient use of disk space.

For high availability (HA), place the SQL Server Always On Availability Group secondary synchronous replica on a separate storage virtual machine (SVM) in the aggregate. For disaster recovery purposes, place the asynchronous replica on an aggregate that is part of a separate storage cluster in the DR site, with content replicated by using NetApp SnapMirror technology. NetApp recommends having at least 10% free space available in an aggregate for optimal storage performance.

Volumes

NetApp FlexVol volumes are created and reside inside aggregates. This term sometimes causes confusion because an ONTAP volume is not a LUN. An ONTAP volume is a management container for data. A volume could contain files, LUNs or even S3 objects. A volume does not take up space, it is only used for management of the contained data.

Volume design considerations

Before you create a database volume design, it is important to understand how the SQL Server I/O pattern and characteristics vary depending on the workload and on the backup and recovery requirements. See the following NetApp recommendations for flexible volumes:

- Avoid sharing volumes between hosts. For example, while it would be possible to create 2 LUNs in a single volume and share each LUN to a different host, this should be avoided because it can complicate management.
- Use NTFS mount points instead of drive letters to surpass the 26-drive-letter limitation in Windows. When using volume mount points, it is a general recommendation to give the volume label the same name as the mount point.

- When appropriate, configure a volume autosize policy to help prevent out-of-space conditions. 17 Best practice guide for Microsoft SQL Server with ONTAP © 2022 NetApp, Inc. All rights reserved.
- If you install SQL Server on an SMB share, make sure that Unicode is enabled on the SMB/CIFS volumes for creating folders.
- Set the snapshot reserve value in the volume to zero for ease of monitoring from an operational perspective.
- Disable snapshot schedules and retention policies. Instead, use SnapCenter to coordinate Snapshot copies of the SQL Server data volumes.
- Place the SQL Server system databases on a dedicated volume.
- tempdb is a system database used by SQL Server as a temporary workspace, especially for I/O intensive DBCC CHECKDB operations. Therefore, place this database on a dedicated volume with a separate set of spindles. In large environments in which volume count is a challenge, you can consolidate tempdb into fewer volumes and store it in the same volume as other system databases after careful planning. Data protection for tempdb is not a high priority because this database is recreated every time SQL Server is restarted.
- Place user data files (.mdf) on separate volumes because they are random read/write workloads. It is common to create transaction log backups more frequently than database backups. For this reason, place transaction log files (.ldf) on a separate volume or VMDK from the data files so that independent backup schedules can be created for each. This separation also isolates the sequential write I/O of the log files from the random read/write I/O of data files and significantly improves SQL Server performance.

LUNs

- Make sure that the user database files and the log directory to store log backup are on separate volumes to prevent the retention policy from overwriting snapshots when these are used with SnapVault technology.
- Make sure that SQL Server databases reside on LUNs separate from LUNs that have non-database files, such as full-text search-related files.
- Placing database secondary files (as part of a filegroup) on separate volumes improves the performance of the SQL Server database. This separation is valid only if the database's .mdf file does not share its LUN with any other .mdf files.
- If you create LUNs with DiskManager or other tools, make sure that the allocation unit size is set to 64K for partitions when formatting the LUNs.
- See the Microsoft Windows and native MPIO under ONTAP best practices for modern SAN to apply multipathing support on Windows to iSCSI devices in the MPIO properties.

Microsoft SQL Server database files and filegroups

Proper SQL Server database file placement on ONTAP is critical during initial deployment stage. This ensures optimal performance, space management, backup and restore times that can be configured to match your business requirements.

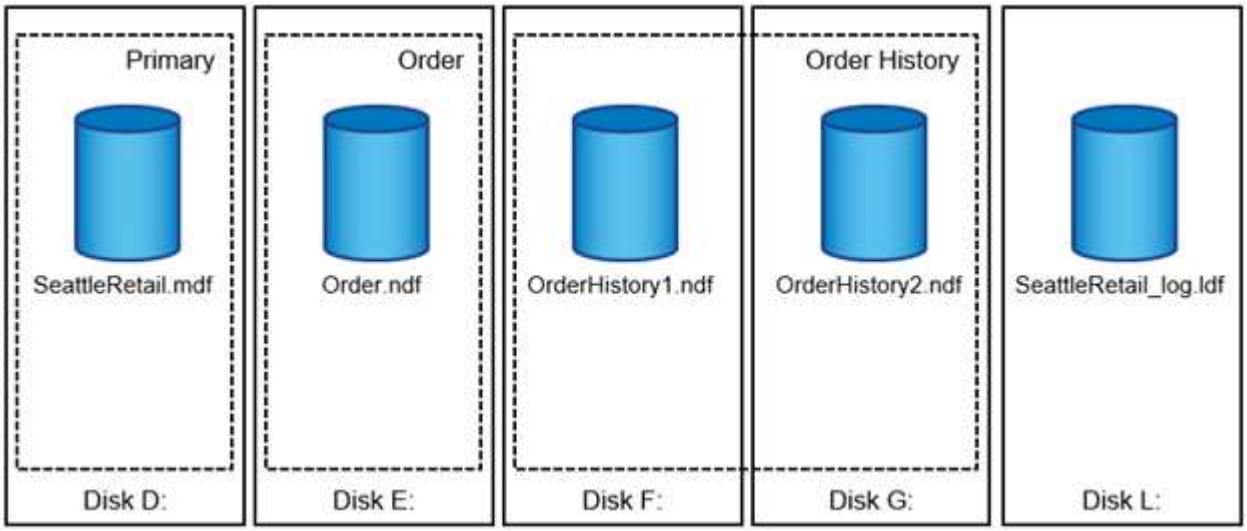
In theory, SQL Server (64-bit) supports 32,767 databases per instance and 524,272TB of database size, although the typical installation usually has several databases. However, the number of the databases SQL Server can handle depends on the load and hardware. It is not unusual to see SQL Server instances hosting dozens, hundreds, or even thousands of small databases.

Each database consists of one or more data files and one or more transaction log files. The transaction log stores the information about database transactions and all data modifications made by each session. Every time the data is modified, SQL Server stores enough information in the transaction log to undo (roll back) or

redo (replay) the action. A SQL Server transaction log is an integral part of SQL Server's reputation for data integrity and robustness. The transaction log is vital to the atomicity, consistency, isolation, and durability (ACID) capabilities of SQL Server. SQL Server writes to the transaction log as soon as any change to the data page happens. Every Data Manipulation Language (DML) statement (for example, select, insert, update, or delete) is a complete transaction, and the transaction log makes sure that the entire set-based operation takes place, making sure of the atomicity of the transaction.

Each database has one primary data file, which, by default, has the .mdf extension. In addition, each database can have secondary database files. Those files, by default, have .ndf extensions.

All database files are grouped into filegroups. A filegroup is the logical unit, which simplifies database administration. They allow the separation between logical object placement and physical database files. When you create the database objects tables, you specify in what filegroup they should be placed without worrying about the underlying data file configuration.



The ability to put multiple data files inside the filegroup allows you to spread the load across different storage devices, which helps to improve the I/O performance of the system. The transaction log in contrast does not benefit from the multiple files because SQL Server writes to the transaction log in a sequential manner.

The separation between logical object placement in the filegroups and physical database files allows you to fine-tune the database file layout, getting the most from the storage subsystem. For example, independent software vendors (ISVs) who are deploying their products to different customers can adjust the number of database files based on the underlying I/O configuration and the expected amount of data during the deployment stage. Those changes are transparent to the application developers, who are placing the database objects in the filegroups rather than database files.

NetApp recommends avoiding the use of the primary filegroup for anything but system objects. Creating a separate filegroup or set of filegroups for the user objects simplifies database administration and disaster recovery, especially in the case of large databases.

You can specify initial file size and autogrowth parameters at the time when you create the database or add new files to an existing database. SQL Server uses a proportional fill algorithm when choosing which data file it should write data into. It writes an amount of data proportionally to the free space available in the files. The more free space in the file, the more writes it handles.

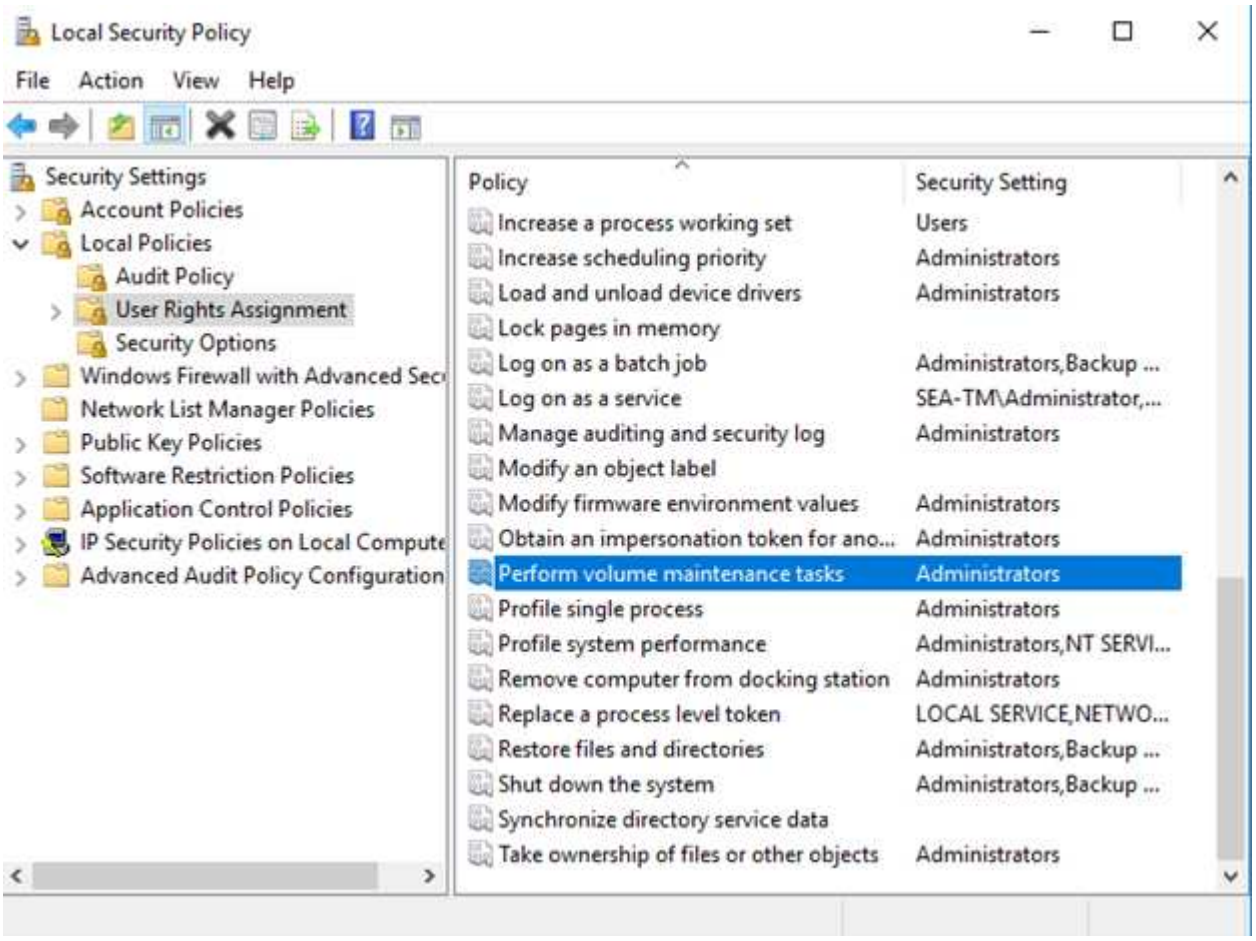
NetApp recommends that all files in the single filegroup have the same initial size and autogrowth parameters, with the grow size defined in megabytes rather than percentages. This helps the proportional fill algorithm evenly balance write activities across data files.

Every time SQL Server grows files, it fills newly allocated space with zeros. That process blocks all sessions that need to write to the corresponding file or, in case of transaction log growth, generate transaction log records.

SQL Server always zeroes out the transaction log, and that behavior cannot be changed. However, you can control whether data files are zeroing out by enabling or disabling instant file initialization. Enabling instant file initialization helps to speed up data file growth and reduces the time required to create or restore the database.

A small security risk is associated with instant file initialization. When this option is enabled, unallocated parts of the data file can contain information from previously deleted OS files. Database administrators can examine such data.

You can enable instant file initialization by adding the SA_MANAGE_VOLUME_NAME permission, also known as “perform volume maintenance task,” to the SQL Server startup account. You can do this under the local security policy management application (secpol.msc), as shown in the following figure. Open the properties for the “perform volume maintenance task” permission and add the SQL Server startup account to the list of users there.



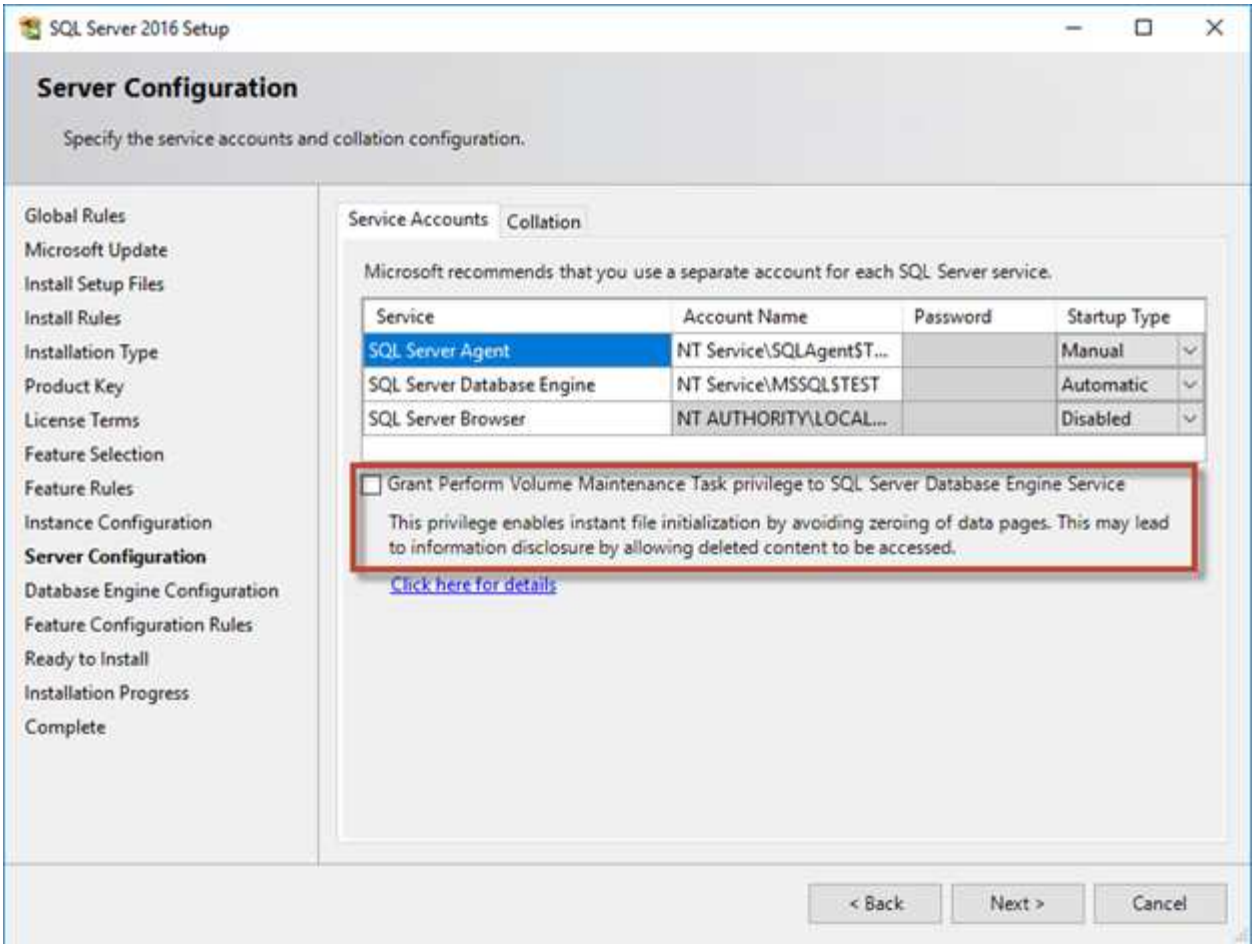
To check if the permission is enabled, you can use the code from the following example. This code sets two trace flags that force SQL Server to write additional information to the error log, create a small database, and read the content of the log.

```
DBCC TRACEON(3004,3605,-1)
GO
CREATE DATABASE DelMe
GO
EXECUTE sp_readerrorlog
GO
DROP DATABASE DelMe
GO
DBCC TRACEOFF(3004,3605,-1)
GO
```

When instant file initialization is not enabled, the SQL Server error log shows that SQL Server is zeroing the mdf data file in addition to zeroing the ldf log file, as shown in the following example. When instant file initialization is enabled, it displays only zeroing of the log file.

	LogDate	ProcessInfo	Text
365	2017-02-09 08:10:07.660	spid53	Ckpt dbid 3 flush delta counts.
366	2017-02-09 08:10:07.660	spid53	Ckpt dbid 3 logging active xact info.
367	2017-02-09 08:10:07.750	spid53	Ckpt dbid 3 phase 1 ended (8)
368	2017-02-09 08:10:07.750	spid53	About to log Checkpoint end.
369	2017-02-09 08:10:07.880	spid53	Ckpt dbid 3 complete
370	2017-02-09 08:10:08.130	spid53	Starting up database 'DelMe'.
371	2017-02-09 08:10:08.150	spid53	FixupLogTail(progress) zeroing C:\Program Files\Micros
372	2017-02-09 08:10:08.160	spid53	Zeroing C:\Program Files\Microsoft SQL Server\MSSQ
373	2017-02-09 08:10:08.170	spid53	Zeroing completed on C:\Program Files\Microsoft SQL
374	2017-02-09 08:10:08.710	spid53	Ckpt dbid 6 started
375	2017-02-09 08:10:08.710	spid53	About to log Checkpoint begin.

The perform volume maintenance task is simplified in SQL Server 2016 and is later provided as an option during the installation process. This figure displays the option to grant the SQL Server database engine service the privilege to perform the volume maintenance task.



Another important database option that controls the database file sizes is autoshrink. When this option is enabled, SQL Server regularly shrinks the database files, reduces their size, and releases space to the operating system. This operation is resource intensive and is rarely useful because the database files grow again after some time when new data comes into the system. Autoshrink must never be enabled on the database.

Microsoft SQL Server log directory

The log directory is specified in SQL Server to store transaction log backup data at the host level. If you are using SnapCenter to backup log files then each SQL Server host used by SnapCenter must have a host log directory configured to perform log backups. SnapCenter has a database repository, so metadata related to backup, restore, or cloning operations is stored in a central database repository.

The sizes of the host log directory is calculated as follows:
Size of host log directory = ((maximum DB LDF size x daily log change rate %) x (snapshot retention) ÷ (1 - LUN overhead space %)
The host log directory sizing formula assumes a 10% LUN overhead space

Place the log directory on a dedicated volume or LUN. The amount of data in the host log directory depends on the size of the backups and the number of days that backups are retained. SnapCenter allows only one host log directory per SQL Server host. You can configure the host log directories at SnapCenter -→ Host -→ Configure Plug-in.

NetApp recommends the following for a host log directory:

- Make sure that the host log directory is not shared by any other type of data that can potentially corrupt the backup snapshot data.
- Do not place user databases or system databases on a LUN that hosts mount points.
- Create the host log directory on the dedicated FlexVol volume to which SnapCenter copies transaction logs.
- Use SnapCenter wizards to migrate databases to NetApp storage so that the databases are stored in valid locations, enabling successful SnapCenter backup and restore operations. Keep in mind that the migration process is disruptive and can cause the databases to go offline while the migration is in progress.
- The following conditions must be in place for failover cluster instances (FCIs) of SQL Server:
  - If you are using a failover cluster instance, the host log directory LUN must be a cluster disk resource in the same cluster group as the SQL Server instance being backed up SnapCenter.
  - If you are using a failover cluster instance, user databases must be placed on shared LUNs that are physical disk cluster resources assigned to the cluster group associated with the SQL Server instance.

Microsoft SQL Server tempdb files

Tempdb database can be heavily utilized. In addition to optimal placement of user database files on ONTAP, alter tempdb datafiles to reduce allocation contention

Page contention can occur on lobal allocation map (GAM), shared global allocation map (SGAM), or page free space (PFS) pages when SQL Server must write to special system pages to allocate new objects. Latches protect (lock) these pages in memory. On a busy SQL Server instance, it can take a long time to get a latch on a system page in tempdb. This results in slower query run times and is known as latch contention. See the following best practices for creating tempdb data files:

- For < or = to 8 cores: tempdb data files = number of cores
- For > 8 cores: 8 tempdb data files

The following example script modifies tempdb by creating eight tempdb files and moving tempdb to the mount point `C:\MSSQL\tempdb` for SQL Server 2012 and later.

```
use master

go

-- Change logical tempdb file name first since SQL Server shipped with
logical file name called tempdev

alter database tempdb modify file (name = 'tempdev', newname =
'tempdev01');
```

```
-- Change location of tempdev01 and log file

alter database tempdb modify file (name = 'tempdev01', filename =
'C:\MSSQL\tempdb\tempdev01.mdf');

alter database tempdb modify file (name = 'templog', filename =
'C:\MSSQL\tempdb\templog.ldf');

GO

-- Assign proper size for tempdev01

ALTER DATABASE [tempdb] MODIFY FILE ( NAME = N'tempdev01', SIZE = 10GB );

ALTER DATABASE [tempdb] MODIFY FILE ( NAME = N'templog', SIZE = 10GB );

GO

-- Add more tempdb files

ALTER DATABASE [tempdb] ADD FILE ( NAME = N'tempdev02', FILENAME =
N'C:\MSSQL\tempdb\tempdev02.ndf' , SIZE = 10GB , FILEGROWTH = 10%);

ALTER DATABASE [tempdb] ADD FILE ( NAME = N'tempdev03', FILENAME =
N'C:\MSSQL\tempdb\tempdev03.ndf' , SIZE = 10GB , FILEGROWTH = 10%);

ALTER DATABASE [tempdb] ADD FILE ( NAME = N'tempdev04', FILENAME =
N'C:\MSSQL\tempdb\tempdev04.ndf' , SIZE = 10GB , FILEGROWTH = 10%);

ALTER DATABASE [tempdb] ADD FILE ( NAME = N'tempdev05', FILENAME =
N'C:\MSSQL\tempdb\tempdev05.ndf' , SIZE = 10GB , FILEGROWTH = 10%);

ALTER DATABASE [tempdb] ADD FILE ( NAME = N'tempdev06', FILENAME =
N'C:\MSSQL\tempdb\tempdev06.ndf' , SIZE = 10GB , FILEGROWTH = 10%);

ALTER DATABASE [tempdb] ADD FILE ( NAME = N'tempdev07', FILENAME =
N'C:\MSSQL\tempdb\tempdev07.ndf' , SIZE = 10GB , FILEGROWTH = 10%);

ALTER DATABASE [tempdb] ADD FILE ( NAME = N'tempdev08', FILENAME =
N'C:\MSSQL\tempdb\tempdev08.ndf' , SIZE = 10GB , FILEGROWTH = 10%);

GO
```

Beginning with SQL Server 2016, the number of CPU cores visible to the operating system is automatically detected during installation and, based on that number, SQL Server calculates and configures the number of tempdb files required for optimum performance.

Microsoft SQL Server and storage efficiency

ONTAP storage efficiency is optimized to store and manage SQL Server data in a way that consumes the least amount of storage space with little or no effect on the overall performance of the system.

Storage efficiency is a combination of RAID, provisioning (overall layout and utilization), mirroring, and other data protection technologies. NetApp technologies including snapshots, thin provisioning, and cloning optimizes existing storage in the infrastructure and deferring or avoiding future storage expenditures. The more you use these technologies together, the larger the savings.

Thin provisioning

Thin provisioning comes in many forms and is integral to many features that ONTAP offers to an enterprise application environment. Thin provisioning is also closely related to efficiency technologies for the same reason: efficiency features allow more logical data to be stored than what technically exists on the storage system.

Almost any use of snapshots involves thin provisioning. For example, a typical 10TB database on NetApp storage includes around 30 days of snapshots. This arrangement results in approximately 10TB of data visible in the active file system and 300TB dedicated to snapshots. The total 310TB of storage usually resides on approximately 12TB to 15TB of space. The active database consumes 10TB, and the remaining 300TB of data only requires 2TB to 5TB of space because only the changes to the original data are stored.

Cloning is also an example of thin provisioning. A major NetApp customer created 40 clones of an 80TB database for use by development. If all 40 developers using these clones overwrote every block in every datafile, over 3.2PB of storage would be required. In practice, turnover is low, and the collective space requirement is closer to 40TB because only changes are stored on the drives.

Space management

Some care must be taken with thin provisioning an application environment because data change rates can increase unexpectedly. For example, space consumption due to snapshots can grow rapidly if database tables are reindexed, or wide-scale patching is applied to VMware guests. A misplaced backup can write a large amount of data in a very short time. Finally, it can be difficult to recover some applications if a file system runs out of free space unexpectedly.

Fortunately, these risks can be addressed with careful configuration of `volume-autogrow` and `snapshot-autodelete` policies. As their names imply, these options enable a user to create policies that automatically clear space consumed by snapshots or grow a volume to accommodate additional data. Many options are available and needs vary by customer.

See the logical storage management documentation for a complete discussion of these features.

LUN thin provisioning

The efficiency of thin provisioning of active LUNs in a file system environment can be lost over time as data is deleted. Unless that deleted data is either overwritten with zeros or the space is released with TRIM/UNMAP space reclamation, the "erased" data occupies more and more unallocated whitespace in the file system. Furthermore, thin provisioning of active LUNs is of limited use in many database environments because datafiles are initialized to their full size at the time of creation.

Careful planning of LVM configuration can improve efficiency and minimize the need for storage provisioning and LUN resizing. When an LVM such as Veritas VxVM or Oracle ASM is used, the underlying LUNs are

divided into extents that are only used when needed. For example, if a dataset begins at 2TB in size but could grow to 10TB over time, this dataset could be placed on 10TB of thin-provisioned LUNs organized in an LVM diskgroup. It would occupy only 2TB of space at the time of creation and would only claim additional space as extents are allocated to accommodate data growth. This process is safe as long as space is monitored.

Fractional reservations

Fractional reserve refers to the behavior of a LUN in a volume with respect to space efficiency. When the option `fractional-reserve` is set to 100%, all data in the volume can experience 100% turnover with any data pattern without exhausting space on the volume.

For example, consider a database on a single 250GB LUN in a 1TB volume. Creating a snapshot would immediately result in the reservation of an additional 250GB of space in the volume to guarantee that the volume does not run out of space for any reason. Using fractional reserves is generally wasteful because it is extremely unlikely that every byte in the database volume would need to be overwritten. There is no reason to reserve space for an event that never happens. Still, if a customer cannot monitor space consumption in a storage system and must be certain that space never runs out, 100% fractional reservations would be required to use snapshots.

Compression and deduplication

Compression and deduplication are both forms of thin provisioning. For example, a 50TB data footprint might compress to 30TB, resulting in a savings of 20TB. For compression to yield any benefits, some of that 20TB must be used for other data, or the storage system must be purchased with less than 50TB. The result is storing more data than is technically available on the storage system. From the data point of view, there is 50TB of data, even though it occupies only 30TB on the drives.

There is always a possibility that the compressibility of a dataset changes, which would result in increased consumption of real space. This increase in consumption means that compression must be managed as with other forms of thin provisioning in terms of monitoring and using `volume-autogrow` and `snapshot-autodelete`.

Compression and deduplication are discussed in further detail in the section xref:./mssql/efficiency.html

Compression and fractional reservations

Compression is a form of thin provisioning. Fractional reservations affect the use of compression, with one important note; space is reserved in advance of the snapshot creation. Normally, fractional reserve is only important if a snapshot exists. If there is no snapshot, fractional reserve is not important. This is not the case with compression. If a LUN is created on a volume with compression, ONTAP preserves space to accommodate a snapshot. This behavior can be confusing during configuration, but it is expected.

As an example, consider a 10GB volume with a 5GB LUN that has been compressed down to 2.5GB with no snapshots. Consider these two scenarios:

- Fractional reserve = 100 results in 7.5GB utilization
- Fractional reserve = 0 results in 2.5GB utilization

The first scenario includes 2.5GB of space consumption for current data and 5GB of space to account for 100% turnover of the source in anticipation of snapshot use. The second scenario reserves no extra space.

Although this situation might seem confusing, it is unlikely to be encountered in practice. Compression implies thin provisioning, and thin provisioning in a LUN environment requires fractional reservations. It is always possible for compressed data to be overwritten by something uncompressible, which means a volume must be

thin provisioned for compression to result in any savings.

> **NetApp recommends** the following reserve configurations:
>
> - Set `fractional-reserve` to 0 when basic capacity monitoring is in place along with `volume-autogrow` and `snapshot-autodelete`.
> - Set `fractional-reserve` to 100 if there is no monitoring ability or if it is impossible to exhaust space under any circumstance.

Efficiency

Space efficiency features, such as compression, compaction, and deduplication are designed to increase the amount of logical data that fits on a given amount of physical storage. The result is lower costs and management overhead.

At a high level, compression is a mathematical process whereby patterns in data are detected and encoded in a way that reduces space requirements. In contrast, deduplication detects actual repeated blocks of data and removes the extraneous copies. Compaction allows multiple logical blocks of data to share the same physical block on media.

> See the sections below on thin provisioning for an explanation of the interaction between storage efficiency and fractional reservation.

SQL Server itself also has features to compress and efficiently manage data. SQL Server currently supports two types of data compression: row compression and page compression.

Row compression changes the data storage format. For example, it changes integers and decimals to the variable-length format instead of their native fixed-length format. It also changes fixed-length character strings to the variable-length format by eliminating blank spaces. Page compression implements row compression and two other compression strategies (prefix compression and dictionary compression). You can find more details about page compression in Page Compression Implementation.

Data compression is currently supported in the Enterprise, Developer, and Evaluation editions of SQL Server 2008 and later. Although compression can be performed by the database itself, this is rarely observed in a SQL Server environment.

Here are the recommendation for managing space for SQL Server data files

- Use thin provisioning in SQL Server environments to improve space utilization and to reduce the overall storage requirements when the space guarantee functionality is used.
- Use autogrow for most common deployment configurations because the storage admin only needs to monitor space usage in the aggregate.
- Advice not to enable deduplication on any volumes containing SQL Server data files unless the volume is known to contain multiple copies of the same data, such as restoring database from backups to a single volume.

Space reclamation

Space reclamation can be initiated periodically to recover unused space in a LUN. With SnapCenter, you can use the following PowerShell command to start space reclamation.

```
Invoke-SdHostVolumeSpaceReclaim -Path drive_path
```

If you need to run space reclamation, this process should be run during periods of low activity because it initially consumes cycles on the host.

Microsoft SQL Server data protection with NetApp management software

Planning database backup is based on business requirements. By combining ONTAP's NetApp Snapshot technology and leveraging Microsoft SQL Server API's, you can quickly take application consistent backup irrespective of size of user databases. For more advanced or scale-out data management requirements, NetApp offers SnapCenter.

SnapCenter

SnapCenter is the NetApp data protection software for enterprise applications. SQL Server databases can be quickly and easily protected with the SnapCenter Plug-in for SQL Server and with OS operations managed by the SnapCenter Plug-in for Microsoft Windows.

SQL Server instance can be a standalone setup, failover cluster instance or it can be always on availability group. The result is that from single-pane-of-glass, databases can be protected, cloned and restored from primary or secondary copy. SnapCenter can manage SQL Server databases both on-premises, in the cloud, and hybrid configurations.Database copies can also be created in few minutes on the orignal or alternate host for development or for reporting purpose.

> **NetApp recommends** using SnapCenter to create Snapshot copies. The T-SQL method described below also works, but SnapCenter offers complete automation over the backup, restore, and cloning process. It also performs discovery to ensure the correct snapshots are being created. No pre-configuration is required.
> …
> SQL Server also requires coordination between the OS and the storage to ensure the correct data is present in snapshots at the time of creation. In most cases, the only safe method to do this is with SnapCenter or T-SQL. Snapshots created without this additional coordination may not be reliably recoverable.

For more details about the SQL Server Plug-in for SnapCenter, see TR-4714: Best practice guide for SQL Server using NetApp SnapCenter.

Protecting database using T-SQL snapshots

In SQL Server 2022, Microsoft introduced T-SQL snapshots that offers a path to scripting and automation of backups operations. Rather than performing full-sized copies, you can prepare the database for snapshots. Once the database is ready for backup, you can leveraging ONTAP REST API's to create snapshots..

The following is a sample backup workflow:

1. Freeze a database with ALTER command. This prepares the database for a consistent snapshot on the underlying storage. After the freeze you can thaw the database and record the snapshot with BACKUP command.

2. Perform snapshots of multiple databases on the storage volumes simultaneously with the new BACKUP GROUP and BACKUP SERVER commands.
3. Perform FULL backups or COPY_ONLY FULL backups. These backups are recorded in msdb as well.
4. Perform point-in-time recovery using log backups taken with the normal streaming approach after the snapshot FULL backup. Streaming differential backups are also supported if desired.

To learn more, see Microsoft documentation to know about the T-SQL snapshots.

Microsoft SQL Server disaster recovery with ONTAP

Enterprise databases and application infrastructures often require replication to protect from natural disaster or unexpected business disruption with minimal downtime.

The SQL Server Always-On availability group replication feature can be an excellent option, and NetApp offers optiosn to integrate data protection with Always-On. In some cases, however, you might want to consider ONTAP replication technology. ONTAP replication options, including MetroCluster and SnapMirror, can scale better with minimal performance impact, protect non-SQL data, and generally provide a full-infrastructure replication and DR solution.

SnapMirror asynchronous

SnapMirror technology offers a fast and flexible asynchronous enterprise solution for replicating data over LANs and WANs. SnapMirror technology transfers only changed data blocks to the destination after the initial mirror is created, significantly reducing network bandwidth requirements.

The following are recommendations for SnapMirror for SQL Server:

- If CIFS is used, the destination SVM must be a member of the same Active Directory domain of which the source SVM is a member so that the access control lists (ACLs) stored within NAS files are not broken during recovery from a disaster.
- Using destination volume names that are the same as the source volume names is not required but can make the process of mounting destination volumes into the destination simpler to manage. If CIFS is used, you must make the destination NAS namespace identical in paths and directory structure to the source namespace.
- For consistency purposes, do not schedule SnapMirror updates from the controllers. Instead, enable SnapMirror updates from SnapCenter to update SnapMirror after either full or log backup is completed.
- Distribute volumes that contain SQL Server data across different nodes in the cluster to allow all cluster nodes to share SnapMirror replication activity. This distribution optimizes the use of node resources.

For more information about SnapMirror, see TR-4015: SnapMirror Configuration and Best Practices Guide for ONTAP 9.

Securing Microsoft SQL Server on ONTAP

Securing a SQL Server database environment is a multdimensional effort that goes beyond managing the database itself. ONTAP offers several unique features designed to secure the storage aspect of your database infrastructure.

Snapshot copies

Storage snapshots are point-in-time replicas of the target data. ONTAP's implementation includes the abilities to set various policies and store up to 1024 snapshots per volume. Snapshots in ONTAP are space-efficient. Space is only consumed as the original dataset changes. They are also read-only. A snapshot can be deleted, but it cannot be changed.

In some cases, snapshots can be scheduled directly on ONTAP. In other cases, software such as SnapCenter may be required to orchestrate application or OS operations before creating snapshots. Whichever approach is best for your workload, an aggressive snapshot strategy can provide data security through frequent, easily-accessible access to backups of everything from boot LUNs to mission-critical databases.

Note: An ONTAP Flexible Volume, or more simply, a volume is not synonymous with a LUN. Volumes are management containers for data such as files or LUNs. For example, a database might be placed on an 8-LUN stripe set, with all LUNs contained in a single volume.

For more information on snapshots, click here.

Tamperproof snapshots

Beginning with ONTAP 9.12.1, snapshots are not just read-only, they can also be protected from accidental or intentional deletion. The feature is called Tamperproof Snapshots. A retention period can be set and enforced via snapshot policy. The resulting snapshots cannot be deleted until they have reached their expiration date. There are no administrative or support center overrides.

This ensures that an intruder, a malicious insider, or even a ransomware attack is unable to compromise the backups even if it resulted in access to the ONTAP system itself. When combined with an frequent snapshot schedule, the result is extremely powerful data protection with a very low RPO.

For more information on Tamperproof Snapshots, click here.

SnapMirror replication

Snapshots can also be replicated to a remote system. This includes Tamperproof Snapshots, where the retention period is applied and enforced on the remote system. The result is the same data protection benefits as local snapshots, but the data is located on a second storage array. This ensures that destruction of the original array does not compromise the backups.

A second system also opens new options for administrative security. For example, some NetApp customers segregate authentication credentials for the primary and secondary storage systems. No single administrative user has access to both systems, which means a malicious administrator cannot delete all copies of data.

For more information on SnapMirror, click here.

Storage Virtual Machines

A newly configured ONTAP storage system is similar to a newly provisioned VMware ESX server because neither of them can support any users until a virtual machine is created. With ONTAP, you create a Storage Virtual Machine (SVM) which becomes the most basic unit of storage management. Each SVM has its own storage resources, protocol configurations, IP addresses, and FCP WWNs. This is the foundation of ONTAP mult-tenancy.

For example, you might configure one SVM for critical production workloads, and a second SVM on a different network segment for development activities. You could then restrict access to the production SVM to certain

administrators, while granting developers more expansive control over the storage resources in the development SVM. You might also need to provide a third SVM to your financial and HR teams to store especially critical eyes-only data.

For more information about SVMs, click here.

Administrative RBAC

ONTAP offers powerful role-based access control (RBAC) for administrative logins. Some admins might need full cluster access, while others might only need access to certain SVMs. Advanced helpdesk personnel might need the ability to increase volumes sizes. The result is you can grant administrative users the access required to perform their job responsibilities, and nothing more. Furthermore, you can secure these logins using PKI from various vendors, restrict access to ssh keys only, and enforce failed login attempt lockouts.

For more information on administrative access control, click here.

Multfactor authentication

ONTAP and certain other NetApp products now support multifactor authentication (MFA) using a variety of methods. The result is a compromised username/password alone is not a security thread without the data from the second factor, such as a FOB or a smartphone app.

For more information, click here.

API RBAC

Automation requires API calls, but not all tools require full administrative access. To help secure automation systems, RBAC is also available at the API level. You can limit the automation user accounts to the API calls required. For example, monitoring software does not need change access, it only requires read access. Workflows that provision storage do not need the ability to delete storage.

To learn more, start chttps://docs.netapp.com/us-en/ontap-automation/rest/rbac_overview.html[here.]

Multi-admin verification (MAV)

Multi "factor" authentication can be taken even further by requiring two different administrators, each with their own credentials, to approve certain activities. This includes changing login permissions, running diagnostic commands, and deleting data.

For more information on multi-admin verification (MAV), click here

MySQL

MySQL databases on ONTAP

MySQL and its variants, including MariaDb and Percona MySQL, is the world's most popular database.

> This documentation on ONTAP and the MySQL database replaces the previously published *TR-4722: MySQL database on ONTAP best practices.*

ONTAP is an ideal platform for MySQL database because ONTAP is literally designed for databases. Numerous features such as random IO latency optimizations to advanced quality of service (QoS) to basic FlexClone functionality were created specifically to address the needs of database workloads.

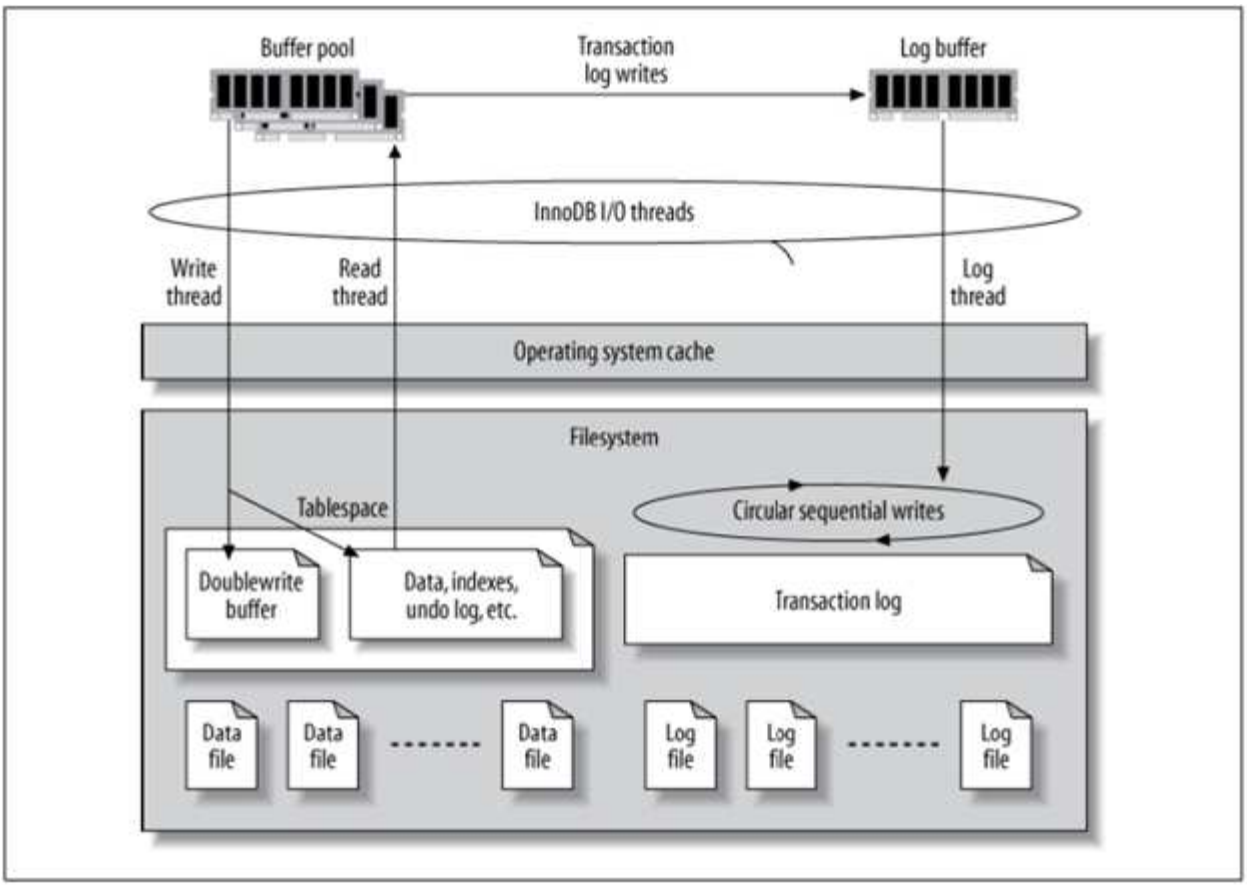
Additional features such as nondisruptive upgrades, (including storage replacement) ensure that your critical databases remain available. You can also have instant disaster recovery for large environments through MetroCluster, or select databases using SnapMirror active sync.

Most importantly, ONTAP delivers unmatched performance with the ability to size the solution for your unique needs. Our high-end systems can deliver over 1M IOPS with latencies measured in microseconds, but if you only need 100K IOPS you can right-size your storage solution with a smaller controller that still runs the exact same storage operating system.

Database configuration

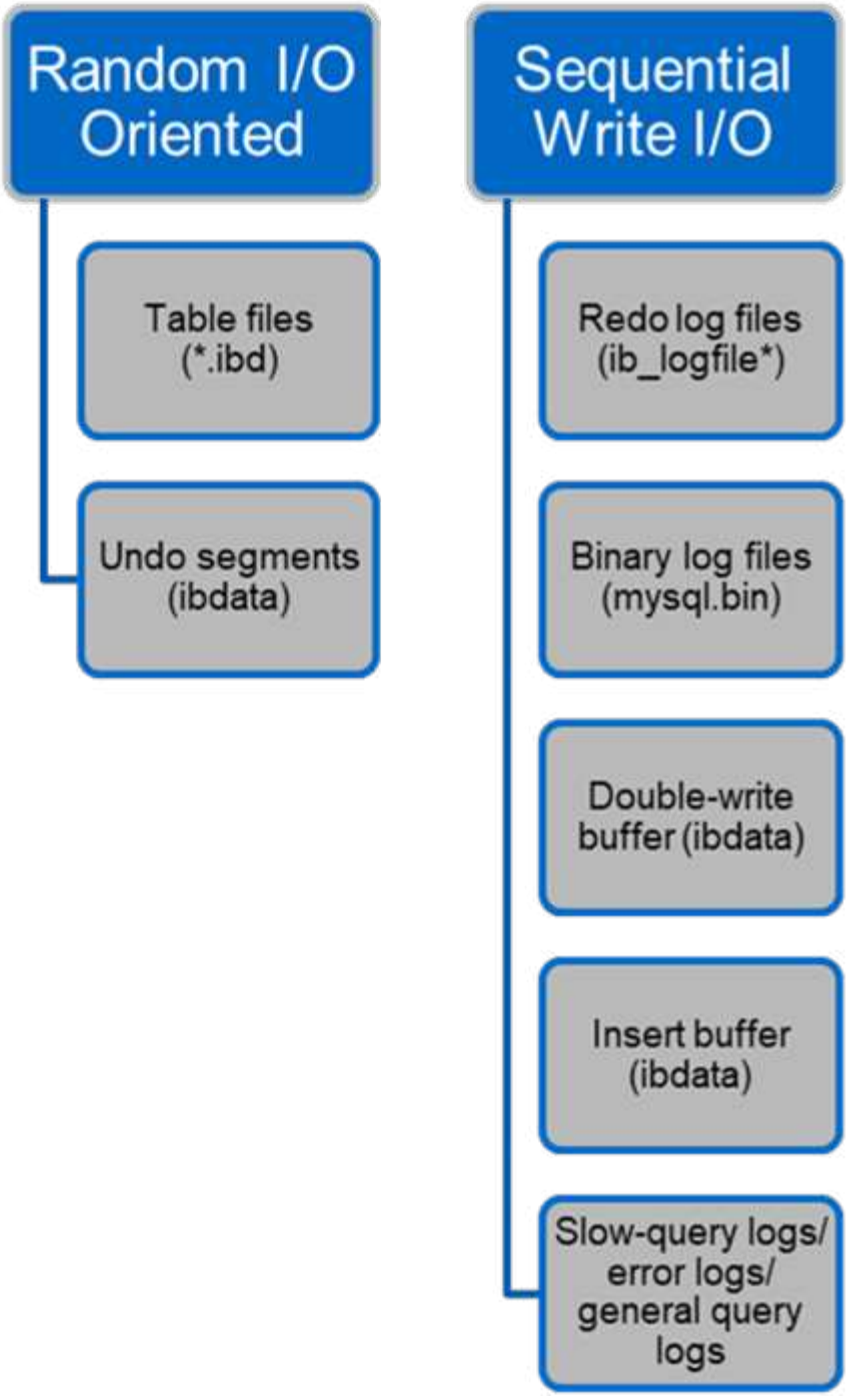
MySQL and InnoDB

InnoDB acts as the middle layer between storage and the MySQL server, it stores the data to the drives.



MySQL I/O is categorized into two types:

- Random file I/O
- Sequential file I/O



Data files are randomly read and overwritten, which results in high IOPS. Therefore, SSD storage is recommended.

Redo log files and binary log files are transactional logs. They are sequentially written, so you can get good performance on HDD with the write cache. A sequential read happens on recovery, but it rarely causes a performance problem, because log file size is usually smaller than data files, and sequential reads are faster than random reads (occurring on data files).

The double-write buffer is a special feature of InnoDB. InnoDB first writes flushed pages to the double-write buffer and then writes the pages to their correct positions on the data files. This process prevents page corruption. Without the double-write buffer, the page might become corrupted if a power failure occurs during

the write-to-drives process. Because writing to the double-write buffer is sequential, it is highly optimized for HDDs. Sequential reads occur on recovery.

Because ONTAP NVRAM already provides write protection, double-write buffering is not required. MySQL has a parameter, `skip_innodb_doublewrite`, to disable the double-write buffer. This feature can substantially improve performance.

The insert buffer is also a special feature of InnoDB. If non-unique secondary index blocks are not in memory, InnoDB inserts entries into the insert buffer to avoid random I/O operations. Periodically, the insert buffer is merged into the secondary index trees in the database. The insert buffer reduces the number of I/O operations by merging I/O requests to the same block; random I/O operations can be sequential. The insert buffer is also highly optimized for HDDs. Both sequential writes and reads occur during normal operations.

Undo segments are random I/O oriented. To guarantee multi-version concurrency (MVCC), InnoDB must register old images in the undo segments. Reading previous images from the undo segments requires random reads. If you run a long transaction with repeatable reads (such as mysqldump—single transaction) or run a long query, random reads can occur. Therefore, storing undo segments on SSDs is better in this instance. If you run only short transactions or queries, the random reads are not an issue.

NetApp recommends the following storage design layout because of the InnoDB I/O characteristics.

- One volume to store random and sequential I/O-oriented files of MySQL
- Another volume to store purely sequential I/O-oriented files of MySQL

This layout also helps you design data protection policies and strategies.

MySQL configuration parameters

NetApp recommends a few important MySQL configuration parameters to obtain optimal performance.

Parameters	Values
innodb_log_file_size	256M
innodb_flush_log_at_trx_commit	2
innodb_doublewrite	0
innodb_flush_method	fsync
innodb_buffer_pool_size	11G
innodb_io_capacity	8192
innodb_buffer_pool_instances	8
innodb_lru_scan_depth	8192
open_file_limit	65535

To set the parameters described in this section, you must change them in the MySQL configuration file (my.cnf). The NetApp best practices are a result of tests performed in-house.

innodb_log_file_size

Selecting the right size for the InnoDB log file size is important for the write operations and for having a decent recovery time after a server crash.

Because so many transactions are logged in to the file, the log file size is important for write operations. When records are modified, the change is not immediately written back to the tablespace. Instead, the change is recorded at the end of the log file and the page is marked as dirty. InnoDB uses its log to convert the random I/O into sequential I/O.

When the log is full, the dirty page is written out to the tablespace in sequence to free up space in the log file. For example, suppose a server crashes in the middle of a transaction, and the write operations are only recorded in the log file. Before the server can go live again, it must go through a recovery phase in which the changes recorded in the log file are replayed. The more entries that are in the log file, the longer it takes for the server to recover.

In this example, the log file size affects both the recovery time and the write performance. When choosing the right number for the log file size, balance the recovery time against write performance. Typically, anything between 128M and 512M is a good value.

innodb_flush_log_at_trx_commit

When there is a change to the data, the change is not immediately written to storage.

Instead, the data is recorded in a log buffer, which is a portion of memory that InnoDB allocates to buffer changes that are recorded in the log file. InnoDB flushes the buffer to the log file when a transaction is committed, when the buffer gets full, or once per second, whichever event happens first. The configuration variable that controls this process is innodb_flush_log_at_trx_commit. The value options include:

- When you set `innodb_flush_log_trx_at_commit=0`, InnoDB writes the modified data (in the InnoDB buffer pool) to the log file (ib_logfile) and flushes the log file (write to storage) every second. However, it does not do anything when the transaction is committed. If there is a power failure or system crash, none of the unflushed data is recoverable because it is not written to either the log file or drives.
- When you set `innodb_flush_log_trx_commit=1`, InnoDB writes the log buffer to the transaction log and flushes to durable storage for every transaction. For example, for all transaction commits, InnoDB writes to the log and then writes to storage. Slower storage negatively affects performance; for example, the number of InnoDB transactions per second is reduced.
- When you set `innodb_flush_log_trx_commit=2`, InnoDB writes the log buffer to the log file at every commit; however, it doesn’t write data to storage. InnoDB flushes data once every second. Even if there is a power failure or system crash, option 2 data is available in the log file and is recoverable.

If performance is the main goal, set the value to 2. Since InnoDB writes to the drives once per second, not for every transaction commit, performance improves dramatically. If a power failure or crash occurs, data can be recovered from the transaction log.

If data safety is the main goal, set the value to 1 so that for every transaction commit, InnoDB flushes to the drives. However, performance might be affected.

NetApp recommends set the innodb_flush_log_trx_commit value to 2 for better performance.

innodb_doublewrite

When `innodb_doublewrite` is enabled (the default), InnoDB stores all data twice: first to the double-write buffer and then to the actual data files.

You can turn off this parameter with `--skip-innodb_doublewrite` for benchmarks or when you're more concerned with top performance than data integrity or possible failures. InnoDB uses a file flush technique called double-write. Before it writes pages to the data files, InnoDB writes them to a contiguous area called the double-write buffer. After the write and the flush to the double-write buffer are complete, InnoDB writes the pages to their proper positions in the data file. If the operating system or a mysqld process crashes during a page write, InnoDB can later find a good copy of the page from the double-write buffer during crash recovery.

NetApp recommends disabling the double-write buffer. ONTAP NVRAM serves the same function. Double-buffering will unnecessarily damage performance.

innodb_buffer_pool_size

The InnoDB buffer pool is the most important part of any tuning activity.

InnoDB relies heavily on the buffer pool for caching indexes and rowing the data, the adaptive hash index, the insert buffer, and many other data structures used internally. The buffer pool also buffers changes to data so that write operations don't have to be performed immediately to storage, thus improving performance. The buffer pool is an integral part of InnoDB and its size must be adjusted accordingly. Consider the following factors when setting the buffer pool size:

- For a dedicated InnoDB-only machine, set the buffer pool size to 80% or more of available RAM.
- If it's not a MySQL dedicated server, set the size to 50% of RAM.

innodb_flush_method

The innodb_flush_method parameter specifies how InnoDB opens and flushes the log and data files.

Optimizations

In InnoDB optimization, setting this parameter tweaks the database performance when applicable.

The following options are for flushing the files through InnoDB:

- `fsync`. InnoDB uses the `fsync()` system call to flush both the data and log files. This option is the default setting.
- `O_DSYNC`. InnoDB uses the `O_DSYNC` option to open and flush the log files and fsync() to flush the data files. InnoDB does not use `O_DSYNC` directly, because there have been problems with it on many varieties of UNIX.
- `O_DIRECT`. InnoDB uses the `O_DIRECT` option (or `directio()` on Solaris) to open the data files and uses `fsync()` to flush both the data and log files. This option is available on some GNU/Linux versions, FreeBSD, and Solaris.
- `O_DIRECT_NO_FSYNC`. InnoDB uses the `O_DIRECT` option during flushing I/O; however, it skips the `fsync()` system call afterward. This option is unsuitable for some types of file systems (for example, XFS). If you are not sure if your file system requires an `fsync()` system call—for example, to preserve all

file metadata—use the `O_DIRECT` option instead.

Observation

In the NetApp lab tests, the `fsync` default option was used on NFS and SAN, and it was a great performance improviser compared to `O_DIRECT`. While using the flush method as `O_DIRECT` with ONTAP, we observed that the client writes a lot of single-byte writes at the border of the 4096 block in serial fashion. These writes increased latency over the network and degraded performance.

innodb_io_capacity

In the InnoDB plug-in, a new parameter called innodb_io_capacity was added from MySQL 5.7.

It controls the maximum number of IOPS that InnoDB performs (which includes the flushing rate of dirty pages as well as the insert buffer [ibuf] batch size). The innodb_io_capacity parameter sets an upper limit on IOPS by InnoDB background tasks, such as flushing pages from the buffer pool and merging data from the change buffer.

Set the innodb_io_capacity parameter to the approximate number of I/O operations that the system can perform per second. Ideally, keep the setting as low as possible, but not so low that background activities slow down. If the setting is too high, data is removed from the buffer pool and insert buffer too quickly for caching to provide a significant benefit.

NetApp recommends that if using this setting over NFS, analyzing the test result of IOPS (SysBench/Fio) and set the parameter accordingly. Use the smallest value possible for flushing and purging to keep up unless you see more modified or dirty pages than you want in the InnoDB buffer pool.

Do not use extreme values such as 20,000 or more unless you've proved that lower values are not sufficient for your workload.

The Innodb_IO_capacity parameter regulates flushing rates and related I/O.

You can seriously harm performance by setting this parameter or the innodb_io_capacity_max parameter too high and wasting I/O operations with premature flushing.

innodb_lru_scan_depth

The `innodb_lru_scan_depth` parameter influences the algorithms and heuristics of the flush operation for the InnoDB buffer pool.

This parameter is primarily of interest to performance experts tuning I/O-intensive workloads. For each buffer pool instance, this parameter specifies how far down in the least recently used (LRU) page list the page cleaner thread should continue scanning, looking for dirty pages to flush. This background operation is performed once per second.

You can adjust the value up or down to minimize the number of free pages. Don't set the value much higher than needed, because the scans can have a significant performance cost. Also, consider adjusting this parameter when changing the number of buffer pool instances, because `innodb_lru_scan_depth * innodb_buffer_pool_instances` defines the amount of work performed by the page cleaner thread each

second.

A setting smaller than the default is suitable for most workloads. Consider increasing the value only if you have spare I/O capacity under a typical workload. Conversely, if a write-intensive workload saturates your I/O capacity, decrease the value, especially if you have a large buffer pool.

open_file_limits

The `open_file_limits` parameter determines the number of files that the operating system permits mysqld to open.

The value of this parameter at run time is the real value permitted by the system and might be different from the value you specify at server startup. The value is 0 on systems where MySQL cannot change the number of open files. The effective `open_files_limit` value is based on the value that is specified at the system startup (if any) and the values of `max_connections` and `table_open_cache` by using these formulas:

- 10 + `max_connections` + (`table_open_cache` x 2)
- `max_connections` x 5
- Operating system limit if positive
- If the operating system limit is infinity: `open_files_limit` value is specified at startup; 5,000 if none

The server attempts to obtain the number of file descriptors using the maximum of these four values. If that many descriptors cannot be obtained, the server attempts to obtain as many as the system will permit.

Host configuration

MySQL containerzation

Containerization of MySQL databases is becoming more prevalent.

Low-level container management is almost always performed through Docker. Container management platforms such as OpenShift and Kubernetes make management of large container environments even simpler. The benefits of containerization include lower costs, because there is no need to license a hypervisor. Also, containers allow multiple databases to run isolated from one another while sharing the same underlying kernel and operating system. Containers can be provisioned in microseconds.

NetApp offers Astra Trident to provide advanced management capabilities of storage. For example, Astra Trident allows a container created in Kubernetes to automatically provision its storage on the appropriate tier, apply export policies, set snapshot policies, and even clone one container to another. For additional information, see the Astra Trident documentation.

MySQL and NFSv3 slot tables

NFSv3 performance on Linux depends on a parameter called `tcp_max_slot_table_entries`.

TCP slot tables are the NFSv3 equivalent of host bus adapter (HBA) queue depth. These tables control the number of NFS operations that can be outstanding at any one time. The default value is usually 16, which is far too low for optimum performance. The opposite problem occurs on newer Linux kernels, which can automatically increase the TCP slot table limit to a level that saturates the NFS server with requests.

For optimum performance and to prevent performance problems, adjust the kernel parameters that control the TCP slot tables.

Run the `sysctl -a | grep tcp.*.slot_table` command, and observe the following parameters:

```
# sysctl -a | grep tcp.*.slot_table
sunrpc.tcp_max_slot_table_entries = 128
sunrpc.tcp_slot_table_entries = 128
```

All Linux systems should include `sunrpc.tcp_slot_table_entries`, but only some include `sunrpc.tcp_max_slot_table_entries`. They should both be set to 128.

Caution
Failure to set these parameters may have significant effects on performance. In some cases, performance is limited because the linux OS is not issuing sufficient I/O. In other cases, I/O latencies increases as the linux OS attempts to issue more I/O than can be serviced.

I/O schedulers and MySQL

The Linux kernel allows low-level control over the way that I/O to block devices is scheduled.

The defaults on various distributions of Linux vary considerably. MySQL recommends that you use `NOOP` or a `deadline` I/O scheduler with native asynchronous I/O (AIO) on Linux. In general, NetApp customers and internal testing show better results with NoOps.

MySQL's InnoDB storage engine uses the asynchronous I/O subsystem (native AIO) on Linux to perform read-ahead and write requests for data file pages. This behavior is controlled by the `innodb_use_native_aio` configuration option, which is enabled by default. With native AIO, the type of I/O scheduler has greater influence on I/O performance. Conduct benchmarks to determine which I/O scheduler provides the best results for your workload and environment.

See the relevant Linux and MySQL documentation for instructions on configuring the I/O scheduler.

MySQL file descriptors

To run, the MySQL server needs file descriptors, and the default values are not sufficient.

It uses them to open new connections, store tables in the cache, create temporary tables to resolve complicated queries, and access persistent ones. If mysqld is not able to open new files when needed, it can stop functioning correctly. A common symptom of this issue is error 24, "Too many open files." The number of file descriptors mysqld can open simultaneously is defined by the `open_files_limit` option set in the configuration file (`/etc/my.cnf`). But `open_files_limit` also depends on the limits of the operating system. This dependency makes setting the variable more complicated.

MySQL cannot set its `open_files_limit` option higher than what is specified under `ulimit 'open files'`. Therefore, you need to explicitly set these limits at the operating system level to allow MySQL to open files as needed. There are two ways to check the file limit in Linux:

- The `ulimit` command quickly gives you a detailed description of the parameters being allowed or locked. The changes made by running this command are not permanent and will erase after a system reboot.
- Changes to the `/etc/security/limit.conf` file are permanent and are not affected by a system reboot.

Make sure to change both the hard and soft limits for user mysql. The following excerpts are from the configuration:

```
mysql hard nofile 65535
mysql soft nofile 65353
```

In parallel, update the same configuration in `my.cnf` to fully use the open file limits.

Storage configuration

MySQL with NFS

The MySQL documentation recommends that you use NFSv4 for NAS deployments.

ONTAP NFS transfer sizes

By default, ONTAP will limit NFS IO sizes to 64K. Random IO with an MySQL database uses a much smaller block size which is well below the 64K maximum. Large-block IO is usually parallelized, so the 64K maximum is also not a limitation.

There are some workloads where the 64K maximum does create a limitation. In particular, single-threaded operations such as full table scan backup operations will run faster and more efficiently if the database can perform fewer but larger IO's. The optimum IO handling size for ONTAP with database workloads is 256K. The NFS mount options listed for specific operating systems below have been updated from 64K to 256K accordingly.

The maximum transfer size for a given ONTAP SVM can be changed as follows:

```
Cluster01::> set advanced

Warning: These advanced commands are potentially dangerous; use them only
when directed to do so by NetApp personnel.

Do you want to continue? {y|n}: y

Cluster01::*> nfs server modify -vserver vserver1 -tcp-max-xfer-size
262144
```

Never decrease the maximum allowable transfer size on ONTAP below the value of rsize/wsize of currently mounted NFS filesystems. This can create hangs or even data corruption with some operating systems. For example, if NFS clients are currently set at a rsize/wsize of 65536, then the ONTAP maximum transfer size could be adjusted between 65536 and 1048576 with no effect because the clients themselves are limited. Reducing the maximum transfer size below 65536 can damage availability or data.

NetApp recommends

Setting the following NFSv4 fstab (/etc/fstab) setting:

```
nfs4 rw,
hard,nointr,bg,vers=4,proto=tcp,noatime,rsize=262144,wsize=262144
```

A common issue with NFSv3 was the locked InnoDB log files after a power outage. Using time or switching log files solved this issue. However, NFSv4 has locking operations and keeps track of open files and delegations.

MySQL with SAN

There are two options to configure MySQL with SAN using the usual two-volume model.

Smaller databases can be placed on a pair of standard LUNs as long as the I/O and capacity demands are within the limits of a single LUN file system. For example, a database that requires approximately 2K random IOPS can be hosted on a single file system on a single LUN. Likewise, a database that is only 100GB in size would fit on a single LUN without creating a management problem.

Larger databases require multiple LUNs. For example, a database that requires 100K IOPS would most likely need at least eight LUNs. A single LUN would become a bottleneck because of the inadequate number of SCSI channels to drives. A 10TB database would similarly be difficult to manage on a single 10TB LUN. Logical volume managers are designed to bond the performance and capacity capabilities of multiple LUNs together to improve performance and manageability.

In both cases, a pair of ONTAP volumes should be sufficient. With a simple configuration, the data file LUN would be placed in a dedicated volume, as would the log LUN. With a logical volume manager configuration, all the LUNs in the data file volume group would be in a dedicated volume, and the LUNs of the log volume group would be in a second dedicated volume.

NetApp recommends using two file systems for MySQL deployments on SAN:

- The first file system stores all MySQL data including tablespace, data, and index.
- The second file system stores all logs (binary logs, slow logs, and transaction logs).

There are multiple reasons for separating data in this manner, including:

- The I/O patterns of data files and log files differ. Separating them would allow more options with QoS controls.
- Optimal use of Snapshot technology requires the ability to independently restore the data files. Commingling data files with log files interferes with data file restoration.
- NetApp SnapMirror technology can be used to provide a simple, low-RPO disaster recovery capability for a database; however, it requires different replication schedules for the data files and logs.

Use this basic two-volume layout to future-proof the solution so that all ONTAP features can be used if needed.

NetApp recommends formatting your drive with the ext4 file system because of the following features:

- Extended approach to block management features used in the journaling file system (JFS) and delayed allocation features of the extended file system (XFS).
- Ext4 permits file systems of up to 1 exbibyte (2^60 bytes) and files of up to 16 tebibytes (16 * 2^40 bytes). In contrast, the ext3 file system supports only a maximum file system size of 16TB and a maximum file size of 2TB.
- In ext4 file systems, multiblock allocation (mballoc) allocates multiple blocks for a file in a single operation, instead of allocating them one by one, as in ext3. This configuration reduces the overhead of calling the block allocator several times, and it optimizes the allocation of memory.
- Although XFS is the default for many Linux distributions, it manages metadata differently and is not suitable for some MySQL configurations.

NetApp recommends using 4k block size options with the mkfs utility to align with existing block LUN size.

```
mkfs.ext4 -b 4096
```

NetApp LUNs store data in 4KB physical blocks, which yields eight 512-byte logical blocks.

If you do not set up the same block size, I/O will not be aligned with physical blocks correctly and could write in two different drives in a RAID group, resulting in latency.

It is important that you align I/O for smooth read/write operations. However, when the I/O begins at a logical block that is not at the start of a physical block, the I/O is misaligned. I/O operations are aligned only when they begin at a logical block—the first logical block in a physical block.

Oracle Database

Oracle databases on ONTAP

ONTAP is designed for Oracle databases. For decades, ONTAP has been optimized for the unique demands of relational database I/O and multiple ONTAP features were created specifically to service the needs of Oracle databases and even at the request of Oracle Inc. itself.

> This documentation replaces these previously published technical reports *TR-3633: Oracle databases on ONTAP; TR-4591: Oracle data protection: Backup, recovery, replication; TR-4592: Oracle on MetroCluster; and TR-4534: Migration of Oracle Databases to NetApp Storage Systems*

In addition to the many possible ways ONTAP brings value to your database environment, there is also a wide variety of user requirements, including database size, performance requirements, and data protection needs. Known deployments of NetApp storage include everything from a virtualized environment of approximately 6,000 databases running under VMware ESX to a single-instance data warehouse currently sized at 996TB and growing. As a result, there are few clear best practices for configuring an Oracle database on NetApp storage.

The requirements for operating an Oracle database on NetApp storage are addressed in two ways. First, when a clear best practice exists, it will be called out specifically. At a high level, many design considerations that must be addressed by architects of Oracle storage solutions based on their specific business requirements will be explained.

ONTAP configuration

RAID and Oracle databases

RAID refers to the use of redundancy to protect data against the loss of a drive.

Questions occasionally arise concerning RAID levels in the configuration of NetApp storage used for Oracle databases and other enterprise applications. Many legacy Oracle best practices regarding storage array configuration contain warnings about using RAID mirroring and/or avoiding certain types of RAID. Although they raise valid points, these sources do not apply to RAID 4 and the NetApp RAID DP and RAID-TEC technologies used in ONTAP.

RAID 4, RAID 5, RAID 6, RAID DP, and RAID-TEC all use parity to ensure drive failure does not result in data loss. These RAID options offer much better storage utilization in comparison to mirroring, but most RAID implementations have a drawback that affects write operations. Completion of a write operation on other RAID implementations may require multiple drive reads to regenerate the parity data, a process commonly called the RAID penalty.

ONTAP, however, does not incur this RAID penalty. This is because of the integration of NetApp WAFL (Write Anywhere File Layout) with the RAID layer. Write operations are coalesced in RAM and prepared as a complete RAID stripe, including parity generation. ONTAP does not need to perform a read in order to complete a write, which means that ONTAP and WAFL avoid the RAID penalty. Performance for latency-critical operations, such as redo logging, is unimpeded, and random data-file writes do not incur any RAID penalty resulting from a need to regenerate parity.

With respect to statistical reliability, even RAID DP offers better protection than RAID mirroring. The primary problem is the demand made on drives during a RAID rebuild. With a mirrored RAID set, the risk of data loss from a drive failing while rebuilding to its partner in the RAID set is much greater than the risk of a triple- drive failure in a RAID DP set.

Oracle databases and storage capacity management

Managing a database or other enterprise application with predictable, manageable, high performance enterprise storage requires some free space on the drives for data and metadata management. The amount of free space required depends on the type of drive used, and business processes.

Free space is defined as any space that is not used for actual data and includes unallocated space on the aggregate itself and unused space within the constituent volumes. Thin provisioning must also be considered. For example, a volume might contain a 1TB LUN of which only 50% is utilized by real data. In a thin provisioned environment, this would correctly appear to be consuming 500GB of space. However, in a fully provisioned environment, the full capacity of 1TB appears to be in use. The 500GB of unallocated space is hidden. This space is unused by actual data and should therefore be included in the calculation of total free space.

NetApp recommendations for storage systems used for enterprise applications are as follows:

SSD aggregates, including AFF systems

NetApp recommends a minimum of 10% free space. This includes all unused space, including free space within the aggregate or a volume and any free space that is allocated due to the use of full provisioning but is not used by actual data. Logical space is unimportant, the question is how much actual free physical space is available for data storage.

The recommendation of 10% free space is very conservative. SSD aggregates can support workloads at even higher levels of utilization without any effect on performance. However, as the utilization of the aggregate increases, the risk of running out of space also increases if utilization is not monitored carefully. Furthermore, while running a system at 99% capacity may not incur a performance penalty, but it would likely incur management effort trying to keep it from filling up completely while additional hardware is ordered, and it may take some time to procure and install additional drives.

HDD aggregates, including Flash Pool aggregates

NetApp recommends a minimum of 15% free space when spinning drives are used. This includes all unused space, including free space within the aggregate or a volume and any free space that is allocated due to the use of full provisioning but is not used by actual data. Performance will be impacted at free speech approaches 10%.

Oracle databases and Storage Virtual Machines

Oracle database storage management is centralized on a Storage Virtual Machine (SVM)

An SVM, known as a vserver at the ONTAP CLI, is a basic functional unit of storage, and it is useful to compare an SVM to a guest on a VMware ESX server.

When first installed, ESX has no pre-configured capabilities, such as hosting a guest OS or supporting an end-user application. It is an empty container until a virtual machine (VM) is defined. ONTAP is similar. When

ONTAP is first installed, it has no data-serving capabilities until an SVM is created. It is the SVM personality that defines the data services.

As with other aspects of storage architecture, the best options for SVM and logical interface (LIF) design depend heavily on scaling requirements and business needs.

SVMs

There is no official best practice for provisioning SVMs for ONTAP. The right approach depends on management and security requirements.

Most customers operate one primary SVM for most of their day-to-day requirements but then create a small number of SVMs for special needs. For example, you might wish to create:

- An SVM for a critical business database managed by a specialist team
- An SVM for a development group to whom complete administrative control has been given so that they can manage their own storage independently
- An SVM for sensitive business data, such as human resources or financial reporting data, for which the administrative team must be limited

In a multi-tenant environment, each tenant's data can be given a dedicated SVM. The limit for the number of SVMs and LIFs per cluster, HA pair, and node are dependant on the protocol being used, the node model, and the version of ONTAP. Consult the NetApp Hardware Universe for these limits.

Oracle database performance management with ONTAP QoS

Safely and efficiently managing multiple Oracle databases requires an effective QoS strategy. The reason is the ever-increasing performance capabilities of a modern storage system.

Specifically, the increased adoption of all-flash storage has enabled the consolidation of workloads. Storage arrays relying on spinning media tended to support only a limited number of I/O-intensive workloads because of the limited IOPS capabilities of older rotational drive technology. One or two highly active databases would saturate the underlying drives long before the storage controllers reached their limits. This has changed. The performance capability of a relatively small number of SSD drives can saturate even the most powerful storage controllers. This means the full capabilities of the controllers can be leveraged without the fear of sudden collapse of performance as spinning media latency spikes.

As a reference example, a simple two-node HA AFF A800 system is capable of servicing up to one million random IOPS before latency climbs above one millisecond. Very few single workloads would be expected to reach such levels. Fully utilizing this AFF A800 system array will involve hosting multiple workloads, and doing this safely while ensuring predictability requires QoS controls.

There are two types of quality of service (QoS) in ONTAP: IOPS and bandwidth. QoS controls can be applied to SVMs, volumes, LUNs, and files.

IOPS QoS

An IOPS QoS control is obviously based on the total IOPS of a given resource, but there are a number of aspects of IOPS QoS that might not be intuitive. A few customers were initially puzzled by the apparent increase in latency when an IOPS threshold is reached. Increasing latency is the natural result of limiting IOPS. Logically, it functions similarly to a token system. For example, if a given volume containing datafiles has a 10K IOPS limit, each I/O that arrives must first receive a token to continue processing. So long as no more than

10K tokens have been consumed in a given second, no delays are present. If IO operations must wait to receive their token, this wait appears as additional latency. The harder a workload pushes up against the QoS limit, the longer each IO must wait in the queue for its turn to be processed, which appears to the user as higher latency.

> Be cautious when applying QoS controls to database transaction/redo log data. While the performance demands of redo logging are normally much, much lower than datafiles, the redo log activity is bursty. The IO happens in brief pulses, and a QoS limit that appears appropriate for average redo IO levels may be too low for the actual requirements. The result can be severe performance limitations as QoS engages with each redo log burst. In general, redo and archive logging should not be limited by QoS.

Bandwidth QoS

Not all I/O sizes are the same. For example, a database might be performing a large number of small block reads which would result in the IOPS threshold being reached, but databases might also be performing a full table scan operation which would consist of a very small number of large block reads, consuming a very large amount of bandwidth but relatively few IOPS.

Likewise, a VMware environment might drive a very high number of random IOPS during boot-up, but would perform fewer but larger IOs during an external backup.

Sometimes effectively managing performance require either IOPS or bandwidth QoS limits, or even both.

Minimum/guaranteed QoS

Many customers seek a solution that includes guaranteed QoS, which is more difficult to achieve than it might seem and is potentially quite wasteful. For example, placing 10 databases with a 10K IOPS guarantee requires sizing a system for a scenario in which all 10 databases are simultaneously running at 10K IOPS, for a total of 100K.

The best use for minimum QoS controls is to protect critical workloads. For example, consider an ONTAP controller with a maximum possible IOPS of 500K and a mix of production and development workloads. You should apply maximum QoS policies to development workloads to prevent any given database from monopolizing the controller. You would then apply minimum QoS policies to production workloads to make sure that they always have the required IOPS available when needed.

Adaptive QoS

Adaptive QoS refers to the ONTAP feature where the QoS limit is based on the capacity of the storage object. It is rarely used with databases because there is not usually any link between the size of a database and its performance requirements. Large databases can be nearly inert, while smaller databases can be the most IOPS-intensive.

Adaptive QoS can be very useful with virtualization datastores because the IOPS requirements of such datasets do tend to correlate to the total size of the database. A newer datastore containing 1TB of VMDK files is likely to need about half the performance as a 2TB datastore. Adaptive QoS allows you to grow the QoS limits automatically as the datastore becomes populated with data.

Oracle databases and ONTAP efficiency features

ONTAP space efficiency features are optimized for Oracle databases. In almost all cases, the best approach is to leave the defaults in place with all efficiency features enabled.

Space efficiency features, such as compression, compaction, and deduplication are designed to increase the amount of logical data that fits on a given amount of physical storage. The result is lower costs and management overhead.

At a high level, compression is a mathematical process whereby patterns in data are detected and encoded in a way that reduces space requirements. In contrast, deduplication detects actual repeated blocks of data and removes the extraneous copies. Compaction allows multiple logical blocks of data to share the same physical block on media.

> See the sections below on thin provisioning for an explanation of the interaction between storage efficiency and fractional reservation.

Compression

Prior to the availability of all-flash storage systems, array-based compression was of limited value because most I/O-intensive workloads required a very large number of spindles to provide acceptable performance. Storage systems invariably contained much more capacity than required as a side effect of the large number of drives. The situation has changed with the rise of solid-state storage. There is no longer a need to vastly overprovision drives purely to obtain good performance. The drive space in a storage system can be matched to actual capacity needs.

The increased IOPS capability of solid-state drives (SSDs) almost always yields cost savings compared to spinning drives, but compression can achieve further savings by increasing the effective capacity of solid-state media.

There are several ways to compress data. Many databases include their own compression capabilities, but this is rarely observed in customer environments. The reason is usually the performance penalty for a **change** to compressed data, plus with some applications there are high licensing costs for database-level compression. Finally, there is the overall performance consequences to database operations. It makes little sense to pay a high per-CPU license cost for a CPU that performs data compression and decompression rather than real database work. A better option is to offload the compression work on to the storage system.

Adaptive compression

Adaptive compression has been thoroughly tested with enterprise workloads with no observed effect on performance, even in an all-flash environment in which latency is measured in microseconds. Some customers have even reported a performance increase with the use of compression because the data remains compressed in cache, effectively increasing the amount of available cache in a controller.

ONTAP manages physical blocks in 4KB units. Adaptive compression uses a default compression block size of 8KB, which means data is compressed in 8KB units. This matches the 8KB block size most often used by relational databases. Compression algorithms become more efficient as more data is compressed as a single unit. A 32KB compression block size would be more space-efficient than an 8KB compression block unit. This does mean that adaptive compression using the default 8KB block size does lead to slightly lower efficiency rates, but there is also a significant benefit to using a smaller compression block size. Database workloads include a large amount of overwrite activity. Overwriting a 8KB of a compressed 32KB block of data requires reading back the entire 32KB of logical data, decompressing it, updating the required 8KB region, recompressing, and then writing the entire 32KB back to the drives. This is a very expensive operation for a storage system and is the reason some competing storage arrays based on larger compression block sizes also incur a significant performance penalty with database workloads.

The block size used by adaptive compression can be increased up to 32KB. This may improve storage efficiency and should be considered for quiescent files such as transaction logs and backup files when a substantial amount of such data is stored on the array. In some situations, active databases that use a 16KB or 32KB block size may also benefit from increasing the block size of adaptive compression to match. Consult a NetApp or partner representative for guidance on whether this is appropriate for your workload.

Compression block sizes larger than 8KB should not be used alongside deduplication on streaming backup destinations. The reason is small changes to the backed-up data affect the 32KB compression window. If the window shifts, the resulting compressed data differs across the entire file. Deduplication occurs after compression, which means the deduplication engine sees each compressed backup differently. If deduplication of streaming backups is required, only 8KB block adaptive compression should be used. Adaptive compression is preferable, because it works at a smaller block size and does not disrupt deduplication efficiency. For similar reasons, host-side compression also interferes with deduplication efficiency.

Compression alignment

Adaptive compression in a database environment requires some consideration of compression block alignment. Doing so is only a concern for data that is subject to random overwrites of very specific blocks. This approach is similar in concept to overall file system alignment, where the start of a filesystem must be aligned to a 4K device boundary and the blocksize of a filesystem must be a multiple of 4K.

For example, an 8KB write to a file is compressed only if it aligns with an 8KB boundary within the file system itself. This point means that it must fall on the first 8KB of the file, the second 8KB of the file, and so forth. The simplest way to ensure correct alignment is to use the correct LUN type, any partition created should have an offset from the start of the device that is a multiple of 8K, and use a filesystem block size that is a multiple of the database block size.

Data such as backups or transaction logs are sequentially written operations that span multiple blocks, all of which are compressed. Therefore, there is no need to consider alignment. The only I/O pattern of concern is the random overwrites of files.

Data compaction

Data compaction is a technology that improves compression efficiency. As stated previously, adaptive compression alone can provide at best 2:1 savings because it is limited to storing an 8KB I/O in a 4KB WAFL block. Compression methods with larger block sizes deliver better efficiency. However, they are not suitable for data that is subject to small block overwrites. Decompressing 32KB units of data, updating an 8KB portion, recompressing, and writing back to the drives creates overhead.

Data compaction works by allowing multiple logical blocks to be stored within physical blocks. For example, a database with highly compressible data such as text or partially full blocks may compress from 8KB to 1KB. Without compaction, that 1KB of data would still occupy an entire 4KB block. Inline data compaction allows that 1KB of compressed data to be stored in just 1KB of physical space alongside other compressed data. It is not a compression technology; it is simply a more efficient way of allocating space on the drives and therefore should not create any detectable performance effect.

The degree of savings obtained vary. Data that is already compressed or encrypted cannot generally be further compressed, and therefore such datasets do not benefit from compaction. In contrast, newly initialized datafiles that contain little more than block metadata and zeros compress up to 80:1.

Temperature sensitive storage efficiency

Temperature sensitive storage efficiency (TSSE) is a available in ONTAP 9.8 and later that relies on block access heat maps to identify infrequently accessed blocks and compress them with greater efficiency.

Deduplication

Deduplication is the removal of duplicate block sizes from a dataset. For example, if the same 4KB block existed in 10 different files, deduplication would redirect that 4KB block within all 10 files to the same 4KB physical block. The result would be a 10:1 improvement in efficiency for that data.

Data such as VMware guest boot LUNs usually deduplicate extremely well because they consist of multiple copies of the same operating system files. Efficiency of 100:1 and greater have been observed.

Some data does not contain duplicate data. For example, an Oracle block contains a header that is globally unique to the database and a trailer that is nearly unique. As a result, deduplication of an Oracle database rarely delivers more than 1% savings. Deduplication with MS SQL databases is slightly better, but unique metadata at the block level is still a limitation.

Space savings of up to 15% in databases with 16KB and large block sizes have been observed in a few cases. The initial 4KB of each block contains the globally unique header, and the final 4KB block contains the nearly unique trailer. The internal blocks are candidates for deduplication, although in practice this is almost entirely attributed to the deduplication of zeroed data.

Many competing arrays claim the ability to deduplicate databases based on the presumption that a database is copied multiple times. In this respect, NetApp deduplication could also be used, but ONTAP offers a better option: NetApp FlexClone technology. The end result is the same; multiple copies of a database that share most of the underlying physical blocks are created. Using FlexClone is much more efficient than taking the time to copy database files and then deduplicating them. It is, in effect, nonduplication rather than deduplication, because a duplicate is never created in the first place.

Efficiency and thin provisioning

Efficiency features are forms of thin provisioning. For example, a 100GB LUN occupying a 100GB volume might compress down to 50GB. There are no actual savings realized yet because the volume is still 100GB. The volume must first be reduced in size so that the space saved can be used elsewhere on the system. If later changes to the 100GB LUN result in the data becoming less compressible, then the LUN grows in size and the volume could fill up.

Thin provisioning is strongly recommended because it can simplify management while delivering a substantial improvement in usable capacity with associated cost savings. The reason is simple - database environments frequently include a lot of empty space, a large number of volumes and LUNs, and compressible data. Thick provisioning results in the reservation of space on storage for volumes and LUNs just in case they someday become 100% full and contain 100% uncompressible data. That is unlikely to ever occur. Thin provisioning allows that space to be reclaimed and used elsewhere and allows capacity management to be based on the storage system itself rather than many smaller volumes and LUNs.

Some customers prefer to use thick provisioning, either for specific workloads or generally based on established operational and procurement practices.

Caution: If a volume is thick provisioned, care must be taken to completely disable all efficiency features for that volume, including decompression and the removal of deduplication using the `sis undo` command. The volume should not appear in `volume efficiency show` output. If it does, the volume is still partially configured for efficiency features. As a result, overwrite guarantees work differently, which increases the chance that configuration oversights cause the volume to unexpectedly run out of space, resulting in database

I/O errors.

Efficiency best practices

NetApp recommends the following:

AFF defaults

Volumes created on ONTAP running on an all-flash AFF system are thin provisioned with all inline efficiency features enabled. Although databases generally do not benefit from deduplication and may include uncompressible data, the default settings are nevertheless appropriate for almost all workloads. ONTAP is designed to efficiently process all types of data and I/O patterns, whether or not they result in savings. Defaults should only be changed if the reasons are fully understood and there is a benefit to deviating.

General recommendations

- If volumes and/or LUNs are not thin provisioned, you should must disable all efficiency settings because using these features provides no savings and the combination of thick provisioning with space efficiency enabled can cause unexpected behavior, including out-of-space errors.
- If data is not subject to overwrites, such as with backups or database transaction logs, you can achieve greater efficiency by enabling TSSE with a low cooling period.
- Some files might contain a significant amount of uncompressible data, for example when compression is already enabled at the application level of files are encrypted. If any of these scenarios are true, consider disabling compression to allow more efficient operation on other volumes containing compressible data.
- Do not use both 32KB compression and deduplication with database backups. See the section Adaptive compression for details.

Thin provisioning with Oracle databases

Thin provisioning for an Oracle database requires careful planning because the result is configuring more space on a storage system than is necessarily physically available. It is very much worth the effort because, when done correctly, the result is significant cost savings and improvements in manageability.

Thin provisioning comes in many forms and is integral to many features that ONTAP offers to an enterprise application environment. Thin provisioning is also closely related to efficiency technologies for the same reason: efficiency features allow more logical data to be stored than what technically exists on the storage system.

Almost any use of snapshots involves thin provisioning. For example, a typical 10TB database on NetApp storage includes around 30 days of snapshots. This arrangement results in approximately 10TB of data visible in the active file system and 300TB dedicated to snapshots. The total 310TB of storage usually resides on approximately 12TB to 15TB of space. The active database consumes 10TB, and the remaining 300TB of data only requires 2TB to 5TB of space because only the changes to the original data are stored.

Cloning is also an example of thin provisioning. A major NetApp customer created 40 clones of an 80TB database for use by development. If all 40 developers using these clones overwrote every block in every datafile, over 3.2PB of storage would be required. In practice, turnover is low, and the collective space requirement is closer to 40TB because only changes are stored on the drives.

Space management

Some care must be taken with thin provisioning an application environment because data change rates can increase unexpectedly. For example, space consumption due to snapshots can grow rapidly if database tables are reindexed, or wide-scale patching is applied to VMware guests. A misplaced backup can write a large amount of data in a very short time. Finally, it can be difficult to recover some applications if a file system runs out of free space unexpectedly.

Fortunately, these risks can be addressed with careful configuration of `volume-autogrow` and `snapshot-autodelete` policies. As their names imply, these options enable a user to create policies that automatically clear space consumed by snapshots or grow a volume to accommodate additional data. Many options are available and needs vary by customer.

See the logical storage management documentation for a complete discussion of these features.

LUN thin provisioning

The efficiency of thin provisioning of active LUNs in a file system environment can be lost over time as data is deleted. Unless that deleted data is either overwritten with zeros or the space is released with TRIM/UNMAP space reclamation, the "erased" data occupies more and more unallocated whitespace in the file system. Furthermore, thin provisioning of active LUNs is of limited use in many database environments because datafiles are initialized to their full size at the time of creation.

Careful planning of LVM configuration can improve efficiency and minimize the need for storage provisioning and LUN resizing. When an LVM such as Veritas VxVM or Oracle ASM is used, the underlying LUNs are divided into extents that are only used when needed. For example, if a dataset begins at 2TB in size but could grow to 10TB over time, this dataset could be placed on 10TB of thin-provisioned LUNs organized in an LVM diskgroup. It would occupy only 2TB of space at the time of creation and would only claim additional space as extents are allocated to accommodate data growth. This process is safe as long as space is monitored.

Fractional reservations

Fractional reserve refers to the behavior of a LUN in a volume with respect to space efficiency. When the option `fractional-reserve` is set to 100%, all data in the volume can experience 100% turnover with any data pattern without exhausting space on the volume.

For example, consider a database on a single 250GB LUN in a 1TB volume. Creating a snapshot would immediately result in the reservation of an additional 250GB of space in the volume to guarantee that the volume does not run out of space for any reason. Using fractional reserves is generally wasteful because it is extremely unlikely that every byte in the database volume would need to be overwritten. There is no reason to reserve space for an event that never happens. Still, if a customer cannot monitor space consumption in a storage system and must be certain that space never runs out, 100% fractional reservations would be required to use snapshots.

Compression and deduplication

Compression and deduplication are both forms of thin provisioning. For example, a 50TB data footprint might compress to 30TB, resulting in a savings of 20TB. For compression to yield any benefits, some of that 20TB must be used for other data, or the storage system must be purchased with less than 50TB. The result is storing more data than is technically available on the storage system. From the data point of view, there is 50TB of data, even though it occupies only 30TB on the drives.

There is always a possibility that the compressibility of a dataset changes, which would result in increased consumption of real space. This increase in consumption means that compression must be managed as with other forms of thin provisioning in terms of monitoring and using `volume-autogrow` and `snapshot-`

`autodelete`.

Compression and deduplication are discussed in further detail in the section xref:./oracle/efficiency.html

Compression and fractional reservations

Compression is a form of thin provisioning. Fractional reservations affect the use of compression, with one important note; space is reserved in advance of the snapshot creation. Normally, fractional reserve is only important if a snapshot exists. If there is no snapshot, fractional reserve is not important. This is not the case with compression. If a LUN is created on a volume with compression, ONTAP preserves space to accommodate a snapshot. This behavior can be confusing during configuration, but it is expected.

As an example, consider a 10GB volume with a 5GB LUN that has been compressed down to 2.5GB with no snapshots. Consider these two scenarios:

- Fractional reserve = 100 results in 7.5GB utilization
- Fractional reserve = 0 results in 2.5GB utilization

The first scenario includes 2.5GB of space consumption for current data and 5GB of space to account for 100% turnover of the source in anticipation of snapshot use. The second scenario reserves no extra space.

Although this situation might seem confusing, it is unlikely to be encountered in practice. Compression implies thin provisioning, and thin provisioning in a LUN environment requires fractional reservations. It is always possible for compressed data to be overwritten by something uncompressible, which means a volume must be thin provisioned for compression to result in any savings.

NetApp recommends the following reserve configurations:

- Set `fractional-reserve` to 0 when basic capacity monitoring is in place along with `volume-autogrow` and `snapshot-autodelete`.
- Set `fractional-reserve` to 100 if there is no monitoring ability or if it is impossible to exhaust space under any circumstance.

Oracle databases and ONTAP controller failover/switchover

An understanding of storage takeover and switchover functions is required to ensure that Oracle database operations are not disrupted by these operations. In addition, the arguments used by takeover and switchover operations can affect data integrity if used incorrectly.

- Under normal conditions, incoming writes to a given controller are synchronously mirrored to its partner. In a NetApp MetroCluster environment, writes are also mirrored to a remote controller. Until a write is stored in nonvolatile media in all locations, it is not acknowledged to the host application.
- The media storing the write data is called nonvolatile memory or NVMEM. It is also sometimes referred to nonvolatile random- access memory (NVRAM), and it can be thought of as a write cache although it functions as a journal. In a normal operation, the data from NVMEM is not read; it is only used to protect data in the event of a software or hardware failure. When data is written to drives, the data is transferred from the RAM in the system, not from NVMEM.
- During a takeover operation, one node in a high availability (HA) pair takes over the operations from its partner. A switchover is essentially the same, but it applies to MetroCluster configurations in which a remote node takes over the functions of a local node.

During routine maintenance operations, a storage takeover or switchover operation should be transparent, other than for a potential brief pause in operations as the network paths change. Networking can be complicated, however, and it is easy to make errors, so NetApp strongly recommends testing takeover and switchover operations thoroughly before putting a storage system into production. Doing so is the only way to be sure that all network paths are configured correctly. In a SAN environment, carefully check the output of the command `sanlun lun show -p` to make sure that all expected primary and secondary paths are available.

Care must be taken when issuing a forced takeover or switchover. Forcing a change to storage configuration with these options means that the state of the controller that owns the drives is disregarded and the alternative node forcibly takes control of the drives. Incorrect forcing of a takeover can result in data loss or corruption. This is because a forced takeover or switchover can discard the contents of NVMEM. After the takeover or switchover is complete, the loss of that data means that the data stored on the drives might revert to a slightly older state from the point of view of the database.

A forced takeover with a normal HA pair should rarely be required. In almost all failure scenarios, a node shut downs and informs the partner so that an automatic failover takes place. There are some edge cases, such as a rolling failure in which the interconnect between nodes is lost and then one controller is lost, in which a forced takeover is required. In such a situation, the mirroring between nodes is lost before the controller failure, which means that the surviving controller would have no longer has a copy of the writes in progress. The takeover then needs to be forced, which means that data potentially is lost.

The same logic applies to a MetroCluster switchover. In normal conditions, a switchover is nearly transparent. However, a disaster can result in a loss of connectivity between the surviving site and the disaster site. From the point of view of the surviving site, the problem could be nothing more than an interruption in connectivity between sites, and the original site might still be processing data. If a node cannot verify the state of the primary controller, only a forced switchover is possible.

NetApp recommends taking the following precautions:

- Be very careful to not accidentally force a takeover or a switchover. Normally, forcing should not be required, and forcing the change can cause data loss.
- If a forced takeover or switchover is required, make sure that the applications are shut down, all file systems are dismounted and logical volume manager (LVM) volume groups are varyoffed. ASM diskgroups must be unmounted.
- In the event of a forced MetroCluster switchover, fence off the failed node from all surviving storage resources. For more information, see the MetroCluster Management and Disaster Recovery Guide for the relevant version of ONTAP.

MetroCluster and multiple aggregates

MetroCluster is a synchronous replication technology that switches to asynchronous mode if connectivity is interrupted. This is the most common request from customers, because guaranteed synchronous replication means that interruption in site connectivity leads to a complete stall of database I/O, taking the database out of service.

With MetroCluster, aggregates rapidly resynchronize after connectivity is restored. Unlike other storage technologies, MetroCluster should never require a complete remirroring after site failure. Only delta changes must be shipped.

In datasets that span aggregates, there is a small risk that additional data recovery steps would be required in a rolling disaster scenario. Specifically, if (a) connectivity between sites is interrupted, (b) connectivity is restored, (c) the aggregates reach a state in which some are synchronized and some are not, and then (d) the primary site is lost, the result is a surviving site in which the aggregates are not synchronized with one another.

If this happens, parts of the dataset are synchronized with one another and it is not possible to bring up applications, databases, or datastores without recovery. If a dataset spans aggregates, NetApp strongly recommends leveraging snapshot-based backups with one of the many available tools to verify rapid recoverability in this unusual scenario.

Database configuration

Oracle database block sizes

ONTAP internally uses a variable block size, which means Oracle databases can be configured with any block size desired. However, filesystem block sizes can affect performance and in some cases a larger redo block size can improve performance.

Datafile block sizes

Some OSs offer a choice of file system block sizes. For file systems supporting Oracle datafiles, the block size should be 8KB when compression is used. When compression is not required, a block size of either 8KB or 4KB can be used.

If a datafile is placed on a file system with a 512-byte block, misaligned files are possible. The LUN and the file system might be properly aligned based on NetApp recommendations, but the file I/O would be misaligned. Such a misalignment would cause severe performance problems.

File systems supporting redo logs must use a block size that is a multiple of the redo block size. This generally requires that both the redo log file system and the redo log itself use a block size of 512 bytes.

Redo block sizes

At very high redo rates, it is possible that 4KB block sizes would perform better because high redo rates allow I/O to be performed in fewer and more efficient operations. If redo rates are greater than 50MBps, consider testing a 4KB block size.

A few customer problems have been identified with databases using redo logs with a 512-byte block size on a file system with a 4KB block size and many very small transactions. The overhead involved in applying multiple 512-byte changes to a single 4KB file system block led to performance problems that were resolved by changing the file system to use a block size of 512 bytes.

NetApp recommends that you do not change the redo block size unless advised by a relevant customer support or professional services organization or the change is based on official product documentation.

Oracle database parameters: db_file_multiblock_read_count

The `db_file_multiblock_read_count` parameter controls the maximum number of Oracle database blocks that Oracle reads as a single operation during sequential I/O.

This parameter does not, however, affect the number of blocks that Oracle reads during any and all read operations, nor does it affect random I/O. Only the block size of sequential I/O is affected.

Oracle recommends that the user leave this parameter unset. Doing so allows the database software to automatically set the optimum value. This generally means that this parameter is set to a value that yields an I/O size of 1MB. For example, a 1MB read of 8KB blocks would require 128 blocks to be read, and the default

value for this parameter would therefore be 128.

Most database performance problems observed by NetApp at customer sites involve an incorrect setting for this parameter. There were valid reasons to change this value with Oracle versions 8 and 9. As a result, the parameter might be unknowingly present in `init.ora` files because the database was upgraded in place to Oracle 10 and later. A legacy setting of 8 or 16, compared to a default value of 128, significantly damages sequential I/O performance.

> **NetApp recommends** setting the `db_file_multiblock_read_count` parameter should not be present in the `init.ora` file. NetApp has never encountered a situation in which changing this parameter improved performance, but there are many cases in which it caused clear damage to sequential I/O throughput.

Oracle database parameters: filesystemio_options

The Oracle initialization parameter `filesystemio_options` controls the use of asynchronous and direct I/O.

Contrary to common belief, asynchronous and direct I/O are not mutually exclusive. NetApp has observed that this parameter is frequently misconfigured in customer environments, and this misconfiguration is directly responsible for many performance problems.

Asynchronous I/O means that Oracle I/O operations can be parallelized. Before the availability of asynchronous I/O on various OSs, users configured numerous dbwriter processes and changed the server process configuration. With asynchronous I/O, the OS itself performs I/O on behalf of the database software in a highly efficient and parallel manner. This process does not place data at risk, and critical operations, such as Oracle redo logging, are still performed synchronously.

Direct I/O bypasses the OS buffer cache. I/O on a UNIX system ordinarily flows through the OS buffer cache. This is useful for applications that do not maintain an internal cache, but Oracle has its own buffer cache within the SGA. In almost all cases, it is better to enable direct I/O and allocate server RAM to the SGA rather than to rely on the OS buffer cache. The Oracle SGA uses the memory more efficiently. In addition, when I/O flows through the OS buffer, it is subject to additional processing, which increases latencies. The increased latencies are especially noticeable with heavy write I/O when low latency is a critical requirement.

The options for `filesystemio_options` are:

- **async.** Oracle submits I/O requests to the OS for processing. This process allows Oracle to perform other work rather than waiting for I/O completion and thus increases I/O parallelization.
- **directio.** Oracle performs I/O directly against physical files rather than routing I/O through the host OS cache.
- **none.** Oracle uses synchronous and buffered I/O. In this configuration, the choice between shared and dedicated server processes and the number of dbwriters are more important.
- **setall.** Oracle uses both asynchronous and direct I/O. In almost all cases, the use of `setall` is optimal.

> The `filesystemio_options` parameter has no effect in DNFS and ASM environments. The use of DNFS or ASM automatically results in the use of both asynchronous and direct I/O.

Some customers have encountered asynchronous I/O problems in the past, especially with previous Red Hat Enterprise Linux 4 (RHEL4) releases. Some out-of-date advice on the internet still suggests avoiding asynchronous IO because of out-of-date information. Asynchronous I/O is stable on all current OSs. There is

no reason to disable it, absent a known bug with the OS.

If a database has been using buffered I/O, a switch to direct I/O might also warrant a change in the SGA size. Disabling buffered I/O eliminates the performance benefit that the host OS cache provides for the database. Adding RAM back to the SGA repairs this problem. The net result should be an improvement in I/O performance.

Although it is almost always better to use RAM for the Oracle SGA than for OS buffer caching, it might be impossible to determine the best value. For example, it might be preferable to use buffered I/O with very small SGA sizes on a database server with many intermittently active Oracle instances. This arrangement allows the flexible use of the remaining free RAM on the OS by all running database instances. This is a highly unusual situation, but it has been observed at some customer sites.

NetApp recommends setting `filesystemio_options` to `setall`, but be aware that under some circumstances the loss of the host buffer cache might require an increase in the Oracle SGA.

Oracle Real Application Clusters (RAC) timeouts

Oracle RAC is a clusterware product with several types of internal heartbeat processes that monitor the health of the cluster.

The information in the misscount section includes critical information for Oracle RAC environments using networked storage, and in many cases the default Oracle RAC settings will need to be changed to ensure the RAC cluster survives network path changes and storage failover/switchover operations.

disktimeout

The primary storage-related RAC parameter is `disktimeout`. This parameter controls the threshold within which voting file I/O must complete. If the `disktimeout` parameter is exceeded, then the RAC node is evicted from the cluster. The default for this parameter is 200. This value should be sufficient for standard storage takeover and giveback procedures.

NetApp strongly recommends testing RAC configurations thoroughly before placing them into production because many factors affect a takeover or giveback. In addition to the time required for storage failover to complete, additional time is also required for Link Aggregation Control Protocol (LACP) changes to propagate. Also, SAN multipathing software must detect an I/O timeout and retry on an alternate path. If a database is extremely active, a large amount of I/O must be queued and retried before voting disk I/O is processed.

If an actual storage takeover or giveback cannot be performed, the effect can be simulated with cable pull tests on the database server.

NetApp recommends the following:

- Leaving the `disktimeout` parameter at the default value of 200.
- Always test a RAC configuration thoroughly.

misscount

The `misscount` parameter normally affects only the network heartbeat between RAC nodes. The default is 30 seconds. If the grid binaries are on a storage array or the OS boot drive is not local, this parameter might

become important. This includes hosts with boot drives located on an FC SAN, NFS-booted OSs, and boot drives located on virtualization datastores such as a VMDK file.

If access to a boot drive is interrupted by a storage takeover or giveback, it is possible that the grid binary location or the entire OS temporarily hangs. The time required for ONTAP to complete the storage operation and for the OS to change paths and resume I/O might exceed the `misscount` threshold. As a result, a node immediately evicts after connectivity to the boot LUN or grid binaries is restored. In most cases, the eviction and subsequent reboot occur with no logging messages to indicate the reason for the reboot. Not all configurations are affected, so test any SAN-booting, NFS-booting, or datastore-based host in a RAC environment so that RAC remains stable if communication to the boot drive is interrupted.

In the case of nonlocal boot drives or a nonlocal file system hosting `grid` binaries, the `misscount` will need to be changed to match `disktimeout`. If this parameter is changed, conduct further testing to also identify any effects on RAC behavior, such as node failover time.

NetApp recommends the following:

- Leave the `misscount` parameter at the default value of 30 unless one of the following conditions applies:
  - `grid` binaries are located on a network-attached drive, including NFS, iSCSI, FC, and datastore-based drives.
  - The OS is SAN booted.
- In such cases, evaluate the effect of network interruptions that affect access to OS or `GRID_HOME` file systems. In some cases, such interruptions cause the Oracle RAC daemons to stall, which can lead to a `misscount`-based timeout and eviction. The timeout defaults to 27 seconds, which is the value of `misscount` minus `reboottime`. In such cases, increase `misscount` to 200 to match `disktimeout`.

Host configuration

Oracle databases with IBM AIX

Configuration topics for Oracle database on IBM AIX with ONTAP.

Concurrent I/O

Achieving optimum performance on IBM AIX requires the use of concurrent I/O. Without concurrent I/O, performance limitations are likely because AIX performs serialized, atomic I/O, which incurs significant overhead.

Originally, NetApp recommended using the `cio` mount option to force the use of concurrent I/O on the file system, but this process had drawbacks and is no longer required. Since the introduction of AIX 5.2 and Oracle 10gR1, Oracle on AIX can open individual files for concurrent IO, as opposed to forcing concurrent I/O on the entire file system.

The best method for enabling concurrent I/O is to set the `init.ora` parameter `filesystemio_options` to `setall`. Doing so allows Oracle to open specific files for use with concurrent I/O.

Using `cio` as a mount option forces the use of concurrent I/O, which can have negative consequences. For example, forcing concurrent I/O disables readahead on file systems, which can damage performance for I/O occurring outside the Oracle database software, such as copying files and performing tape backups.

Furthermore, products such as Oracle GoldenGate and SAP BR*Tools are not compatible with using the `cio` mount option with certain versions of Oracle.

NetApp recommends the following:

- Do not use the `cio` mount option at the file system level. Rather, enable concurrent I/O through the use of `filesystemio_options=setall`.
- Only use the `cio` mount option should if it is not possible to set `filesystemio_options=setall`.

AIX NFS mount options

The following table lists the AIX NFS mount options for Oracle single instance databases.

File type	Mount options
ADR Home	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Controlfiles Datafiles Redo logs	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
`ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, intr`

The following table lists the AIX NFS mount options for RAC.

File type	Mount options
ADR Home	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Controlfiles Datafiles Redo logs	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac`
`CRS/Voting`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac`
Dedicated `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Shared `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr`

The primary difference between single-instance and RAC mount options is the addition of `noac` to the mount options. This addition has the effect of disabling the host OS caching that enables all instances in the RAC cluster to have a consistent view of the state of the data.

Although using the `cio` mount option and the `init.ora` parameter `filesystemio_options=setall` has

the same effect of disabling host caching, it is still necessary to use `noac`. `noac` is required for shared `ORACLE_HOME` deployments to facilitate the consistency of files such as Oracle password files and `spfile` parameter files. If each instance in a RAC cluster has a dedicated `ORACLE_HOME`, then this parameter is not required.

AIX jfs/jfs2 Mount Options

The following table lists the AIX jfs/jfs2 mount options.

File type	Mount options
ADR Home	Defaults
Controlfiles Datafiles Redo logs	Defaults
ORACLE_HOME	Defaults

Before using AIX `hdisk` devices in any environment, including databases, check the parameter `queue_depth`. This parameter is not the HBA queue depth; rather it relates to the SCSI queue depth of the individual `hdisk device. Depending on how the LUNs are configured, the value for` `queue_depth` might be too low for good performance. Testing has shown the optimum value to be 64.

Oracle databases with HP-UX

Configuration topics for Oracle database on HP-UX with ONTAP.

HP-UX NFS Mount Options

The following table lists the HP-UX NFS mount options for a single instance.

File type	Mount options
ADR Home	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`
Control files Datafiles Redo logs	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144,forc edirectio, nointr,suid`
`ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`

The following table lists the HP-UX NFS mount options for RAC.

File type	Mount options
ADR Home	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, noac,suid`

File type	Mount options
Control files Datafiles Redo logs	`rw, bg,hard, [vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,forcedirectio,suid`
CRS/Voting	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac, forcedirectio,suid`
Dedicated `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`
Shared `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,suid`

The primary difference between single-instance and RAC mount options is the addition of `noac` and `forcedirectio` to the mount options. This addition has the effect of disabling host OS caching, which enables all instances in the RAC cluster to have a consistent view of the state of the data. Although using the `init.ora` parameter `filesystemio_options=setall` has the same effect of disabling host caching, it is still necessary to use `noac` and `forcedirectio`.

The reason `noac` is required for shared `ORACLE_HOME` deployments is to facilitate consistency of files such as Oracle password files and spfiles. If each instance in a RAC cluster has a dedicated `ORACLE_HOME`, this parameter is not required.

HP-UX VxFS mount options

Use the following mount options for file systems hosting Oracle binaries:

```
delaylog,nodatainlog
```

Use the following mount options for file systems containing datafiles, redo logs, archive logs, and control files in which the version of HP-UX does not support concurrent I/O:

```
nodatainlog,mincache=direct,convosync=direct
```

When concurrent I/O is supported (VxFS 5.0.1 and later, or with the ServiceGuard Storage Management Suite), use these mount options for file systems containing datafiles, redo logs, archive logs, and control files:

```
delaylog,cio
```

> The parameter `db_file_multiblock_read_count` is especially critical in VxFS environments. Oracle recommends that this parameter remain unset in Oracle 10g R1 and later unless specifically directed otherwise. The default with an Oracle 8KB block size is 128. If the value of this parameter is forced to 16 or less, remove the `convosync=direct` mount option because it can damage sequential I/O performance. This step damages other aspects of performance and should only be taken if the value of `db_file_multiblock_read_count` must be changed from the default value.

Oracle databases with Linux

Configuration topics specific to the Linux OS.

Linux NFSv3 TCP slot tables

TCP slot tables are the NFSv3 equivalent of host bus adapter (HBA) queue depth. These tables control the number of NFS operations that can be outstanding at any one time. The default value is usually 16, which is far too low for optimum performance. The opposite problem occurs on newer Linux kernels, which can automatically increase the TCP slot table limit to a level that saturates the NFS server with requests.

For optimum performance and to prevent performance problems, adjust the kernel parameters that control the TCP slot tables.

Run the `sysctl -a | grep tcp.*.slot_table` command, and observe the following parameters:

```
# sysctl -a | grep tcp.*.slot_table
sunrpc.tcp_max_slot_table_entries = 128
sunrpc.tcp_slot_table_entries = 128
```

All Linux systems should include `sunrpc.tcp_slot_table_entries`, but only some include `sunrpc.tcp_max_slot_table_entries`. They should both be set to 128.

Caution
Failure to set these parameters may have significant effects on performance. In some cases, performance is limited because the linux OS is not issuing sufficient I/O. In other cases, I/O latencies increases as the linux OS attempts to issue more I/O than can be serviced.

Linux NFS mount options

The following table lists the Linux NFS mount options for a single instance.

File type	Mount options
ADR Home	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144`
Control files Datafiles Redo logs	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr`

File type	Mount options
ORACLE_HOME	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr`

The following table lists the Linux NFS mount options for RAC.

File type	Mount options
ADR Home	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,actimeo=0`
Control files Data files Redo logs	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,actimeo=0`
CRS/voting	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,actimeo=0`
Dedicated `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144`
Shared `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,actimeo=0`

The primary difference between single-instance and RAC mount options is the addition of `actimeo=0` to the mount options. This addition has the effect of disabling the host OS caching, which enables all instances in the RAC cluster to have a consistent view of the state of the data. Although using the `init.ora` parameter `filesystemio_options=setall` has the same effect of disabling host caching, it is still necessary to use `actimeo=0`.

The reason `actimeo=0` is required for shared `ORACLE_HOME` deployments is to facilitate consistency of files such as the Oracle password files and spfiles. If each instance in a RAC cluster has a dedicated `ORACLE_HOME`, then this parameter is not required.

Generally, nondatabase files should be mounted with the same options used for single-instance datafiles, although specific applications might have different requirements. Avoid the mount options `noac` and `actimeo=0` if possible because these options disable file system-level readahead and buffering. This can cause severe performance problems for processes such as extraction, translation, and loading.

ACCESS and GETATTR

Some customers have noted that an extremely high level of other IOPS such as ACCESS and GETATTR can dominate their workloads. In extreme cases, operations such as reads and writes can be as low as 10% of the total. This is normal behavior with any database that includes using `actimeo=0` and/or `noac` on Linux because these options cause the Linux OS to constantly reload file metadata from the storage system. Operations such as ACCESS and GETATTR are low-impact operations that are serviced from the ONTAP cache in a database environment. They should not be considered genuine IOPS, such as reads and writes, that create true demand on storage systems. These other IOPS do create some load, however, especially in RAC environments. To address this situation, enable DNFS, which bypasses the OS buffer cache and avoids these unnecessary metadata operations.

Linux Direct NFS

One additional mount option, called `nosharecache`, is required when (a) DNFS is enabled and (b) a source volume is mounted more than once on a single server (c) with a nested NFS mount. This configuration is seen primarily in environments supporting SAP applications. For example, a single volume on a NetApp system could have a directory located at `/vol/oracle/base` and a second at `/vol/oracle/home`. If `/vol/oracle/base` is mounted at `/oracle` and `/vol/oracle/home` is mounted at `/oracle/home`, the result is nested NFS mounts that originate on the same source.

The OS can detect the fact that `/oracle` and `/oracle/home` reside on the same volume, which is the same source file system. The OS then uses the same device handle for accessing the data. Doing so improves the use of OS caching and certain other operations, but it interferes with DNFS. If DNFS must access a file, such as the `spfile`, on `/oracle/home`, it might erroneously attempt to use the wrong path to the data. The result is a failed I/O operation. In these configurations, add the `nosharecache` mount option to any NFS file system that shares a source FlexVol volume with another NFS file system on that host. Doing so forces the Linux OS to allocate an independent device handle for that file system.

Linux Direct NFS and Oracle RAC

The use of DNFS has special performance benefits for Oracle RAC on the Linux OS because Linux does not have a method to force direct I/O, which is required with RAC for coherency across the nodes. As a workaround, Linux requires the use of the `actimeo=0` mount option, which causes file data to expire immediately from the OS cache. This option in turn forces the Linux NFS client to constantly reread attribute data, which damages latency and increases load on the storage controller.

Enabling DNFS bypasses the host NFS client and avoids this damage. Multiple customers have reported significant performance improvements on RAC clusters and significant decreases in ONTAP load (especially with respect to other IOPS) when enabling DNFS.

Linux Direct NFS and oranfstab file

When using DNFS on Linux with the multipathing option, multiple subnets must be used. On other OSs, multiple DNFS channels can be established by using the `LOCAL` and `DONTROUTE` options to configure multiple DNFS channels on a single subnet. However, this does not work properly on Linux and unexpected performance problems can result. With Linux, each NIC used for DNFS traffic must be on a different subnet.

I/O scheduler

The Linux kernel allows low-level control over the way that I/O to block devices is scheduled. The defaults on various distribution of Linux vary considerably. Testing shows that Deadline usually offers the best results, but on occasion NOOP has been slightly better. The difference in performance is minimal, but test both options if it is necessary to extract the maximum possible performance from a database configuration. CFQ is the default in many configurations, and it has demonstrated significant performance problems with database workloads.

See the relevant Linux vendor documentation for instructions on configuring the I/O scheduler.

Multipathing

Some customers have encountered crashes during network disruption because the multipath daemon was not running on their system. On recent versions of Linux, the installation process of the OS and the multipathing daemon might leave these OSs vulnerable to this problem. The packages are installed correctly, but they are not configured for automatic startup after a reboot.

For example, the default for the multipath daemon on RHEL5.5 might appear as follows:

```
[root@host1 iscsi]# chkconfig --list | grep multipath
multipathd      0:off   1:off   2:off   3:off   4:off   5:off   6:off
```

This can be corrected with the following commands:

```
[root@host1 iscsi]# chkconfig multipathd on
[root@host1 iscsi]# chkconfig --list | grep multipath
multipathd      0:off   1:off   2:on    3:on    4:on    5:on    6:off
```

ASM mirroring

ASM mirroring might require changes to the Linux multipath settings to allow ASM to recognize a problem and switch over to an alternate fail group. Most ASM configurations on ONTAP use external redundancy, which means that data protection is provided by the external array and ASM does not mirror data. Some sites use ASM with normal redundancy to provide two-way mirroring, normally across different sites.

The Linux settings shown in the NetApp Host Utilities documentation include multipath parameters that result in indefinite queuing of I/O. This means an I/O on a LUN device with no active paths waits as long as required for the I/O to complete. This is usually desirable because Linux hosts wait as long as needed for SAN path changes to complete, for FC switches to reboot, or for a storage system to complete a failover.

This unlimited queuing behavior causes a problem with ASM mirroring because ASM must receive an I/O failure for it to retry I/O on an alternate LUN.

Set the following parameters in the Linux `multipath.conf` file for ASM LUNs used with ASM mirroring:

```
polling_interval 5
no_path_retry 24
```

These settings create a 120-second timeout for ASM devices. The timeout is calculated as the `polling_interval` * `no_path_retry` as seconds. The exact value might need to be adjusted in some circumstances, but a 120 second timeout should be sufficient for most uses. Specifically, 120 seconds should allow a controller takeover or giveback to occur without producing an I/O error that would result in the fail group being taken offline.

A lower `no_path_retry` value can shorten the time required for ASM to switch to an alternate fail group, but this also increases the risk of an unwanted failover during maintenance activities such as a controller takeover. The risk can be mitigated by careful monitoring of the ASM mirroring state. If an unwanted failover occurs, the mirrors can be rapidly resynced if the resync is performed relatively quickly. For additional information, see the Oracle documentation on ASM Fast Mirror Resync for the version of Oracle software in use.

Linux xfs, ext3, and ext4 mount options

NetApp recommends using the default mount options.

Oracle databases with ASMLib/AFD (ASM Filter Driver)

Configuration topics specific to the Linux OS using AFD and ASMlib

ASMlib block sizes

ASMlib is an optional ASM management library and associated utilities. Its primary value is the capability to stamp a LUN or an NFS-based file as an ASM resource with a human-readable label.

Recent versions of ASMlib detect a LUN parameter called Logical Blocks Per Physical Block Exponent (LBPPBE). This value was not reported by the ONTAP SCSI target until recently. It now returns a value that indicates that a 4KB block size is preferred. This is not a definition of block size, but it is a hint to any application that uses LBPPBE that I/Os of a certain size might be handled more efficiently. ASMlib does, however, interpret LBPPBE as a block size and persistently stamps the ASM header when the ASM device is created.

This process can cause problems with upgrades and migrations in a number of ways, all based on the inability to mix ASMlib devices with different block sizes in the same ASM diskgroup.

For example, older arrays generally reported an LBPPBE value of 0 or did not report this value at all. ASMlib interprets this as a 512-byte block size. Newer arrays would be interpreted as having a 4KB block size. It is not possible to mix both 512-byte and 4KB devices in the same ASM diskgroup. Doing so would block a user from increasing the size of the ASM diskgroup using LUNs from two arrays or leveraging ASM as a migration tool. In other cases, RMAN might not permit the copying of files between an ASM diskgroup with a 512-byte block size and an ASM diskgroup with a 4KB block size.

The preferred solution is to patch ASMlib. The Oracle bug ID is 13999609, and the patch is present in oracleasm-support-2.1.8-1 and higher. This patch allows a user to set the parameter `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` to `true` in the `/etc/sysconfig/oracleasm` configuration file. Doing so blocks ASMlib from using the LBPPBE parameter, which means that LUNs on the new array are now recognized as 512-byte block devices.

> The option does not change the block size on LUNs that were previously stamped by ASMlib. For example, if an ASM diskgroup with 512-byte blocks must be migrated to a new storage system that reports a 4KB block, the option `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` must be set before the new LUNs are stamped with ASMlib. If devices have already been stamped by oracleasm, they must be reformatted before being restamped with a new block size. First, deconfigure the device with `oracleasm deletedisk`, and then clear the first 1GB of the device with `dd if=/dev/zero of=/dev/mapper/device bs=1048576 count=1024`. Finally, if the device had been previously partitioned, use the `kpartx` command to remove stale partitions or simply reboot the OS.

If ASMlib cannot be patched, ASMlib can be removed from the configuration. This change is disruptive and requires the unstamping of ASM disks and making sure that the `asm_diskstring` parameter is set correctly. This change does not, however, require the migration of data.

ASM Filter Drive (AFD) block sizes

AFD is an optional ASM management library which is becoming the replacement for ASMlib. From a storage point of view, it is very similar to ASMlib, but it includes additional features such as the ability to block non-Oracle I/O to reduce the chances of user or application errors that could corrupt data.

Device block sizes

Like ASMlib, AFD also reads the LUN parameter Logical Blocks Per Physical Block Exponent (LBPPBE) and by default uses the physical block size, not the logical block size.

This could create a problem if AFD is added to an existing configuration where the ASM devices are already formatted as 512 byte block devices. The AFD driver would recognize the LUN as a 4K device and the mismatch between the ASM label and the physical device would prevent access. Likewise, migrations would be affected because it is not possible to mix both 512-byte and 4KB devices in the same ASM diskgroup. Doing so would block a user from increasing the size of the ASM diskgroup using LUNs from two arrays or leveraging ASM as a migration tool. In other cases, RMAN might not permit the copying of files between an ASM diskgroup with a 512-byte block size and an ASM diskgroup with a 4KB block size.

The solution is simple - AFD includes a parameter to control whether it uses the logical or physical block sizes. This is a global parameter affecting all devices on the system. To force AFD to use the logical block size, set `options oracleafd oracleafd_use_logical_block_size=1` in the `/etc/modprobe.d/oracleafd.conf` file.

Multipath transfer sizes

Recent linux kernel changes enforce I/O size restrictions sent to multipath devices, and AFD does not honor these restrictions. The I/Os are then rejected, which causes the LUN path to go offline. The result is an inability to install Oracle Grid, configure ASM, or create a database.

The solution is to manually specify the maximum transfer length in the multipath.conf file for ONTAP LUNs:

```
devices {
            device {
                vendor "NETAPP"
                product "LUN.*"
                max_sectors_kb 4096
        }
    }
```

> Even if no problems currently exist, this parameter should be set if AFD is used to ensure that a future linux upgrade does not unexpectedly cause problems.

Oracle databases with Microsoft Windows

Configuration topics for Oracle database on Microsoft Windows with ONTAP..

NFS

Oracle supports the use of Microsoft Windows with the direct NFS client. This capability offers a path to the management benefits of NFS, including the ability to view files across environments, dynamically resize volumes, and leverage a less expensive IP protocol. See the official Oracle documentation for information on installing and configuring a database on Microsoft Windows using DNFS. No special best practices exist.

SAN

For optimal compression efficiency, ensure the NTFS file system uses an 8K or larger allocation unit. Use of a

4K allocation unit, which is generally the default, negatively impacts compression efficiency.

Oracle databases with Solaris

Configuration topics specific to the Solaris OS.

Solaris NFS mount options

The following table lists the Solaris NFS mount options for a single instance.

File type	Mount options
ADR Home	`rw,bg,hard,[vers=3,vers=4.1], roto=tcp, timeo=600,rsize=262144,wsize=262144`
Controlfiles Datafiles Redo logs	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,llock,suid`
`ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`

The use of `llock` has been proven to dramatically improve performance in customer environments by removing the latency associated with acquiring and releasing locks on the storage system. Use this option with care in environments in which numerous servers are configured to mount the same file systems and Oracle is configured to mount these databases. Although this is a highly unusual configuration, it is used by a small number of customers. If an instance is accidentally started a second time, data corruption can occur because Oracle is unable to detect the lock files on the foreign server. NFS locks do not otherwise offer protection; as in NFS version 3, they are advisory only.

Because the `llock` and `forcedirectio` parameters are mutually exclusive, it is important that `filesystemio_options=setall` is present in the `init.ora` file so that `directio` is used. Without this parameter, host OS buffer caching is used and performance can be adversely affected.

The following table lists the Solaris NFS RAC mount options.

File type	Mount options
ADR Home	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, noac`
Control files Data files Redo logs	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,forcedirectio`
CRS/Voting	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,forcedirectio`
Dedicated `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`

File type	Mount options
Shared `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,suid`

The primary difference between single-instance and RAC mount options is the addition of `noac` and `forcedirectio` to the mount options. This addition has the effect of disabling the host OS caching, which enables all instances in the RAC cluster to have a consistent view of the state of the data. Although using the `init.ora` parameter `filesystemio_options=setall` has the same effect of disabling host caching, it is still necessary to use `noac` and `forcedirectio`.

The reason `actimeo=0` is required for shared `ORACLE_HOME` deployments is to facilitate consistency of files such as Oracle password files and spfiles. If each instance in a RAC cluster has a dedicated `ORACLE_HOME`, this parameter is not required.

Solaris UFS mount options

NetApp strongly recommends using the logging mount option so that data integrity is preserved in the case of a Solaris host crash or the interruption of FC connectivity. The logging mount option also preserves the usability of Snapshot backups.

Solaris ZFS

Solaris ZFS must be installed and configured carefully to deliver optimum performance.

mvector

Solaris 11 included a change in how it processes large I/O operations which can result in severe performance problems on SAN storage arrays. The problem is documented in detail in the NetApp bug report 630173, "Solaris 11 ZFS Performance Regression. " The solution is to change an OS parameter called `zfs_mvector_max_size`.

Run the following command as root:

```
[root@host1 ~]# echo "zfs_mvector_max_size/W 0t131072" |mdb -kw
```

If any unexpected problems arise from this change, it can be easily reversed by running the following command as root:

```
[root@host1 ~]# echo "zfs_mvector_max_size/W 0t1048576" |mdb -kw
```

Kernel

Reliable ZFS performance requires a Solaris kernel patched against LUN alignment problems. The fix was introduced with patch 147440-19 in Solaris 10 and with SRU 10.5 for Solaris 11. Only use Solaris 10 and later with ZFS.

LUN configuration

To configure a LUN, complete the following steps:

1. Create a LUN of type `solaris`.
2. Install the appropriate Host Utility Kit (HUK) specified by the NetApp Interoperability Matrix Tool (IMT).
3. Follow the instructions in the HUK exactly as described. The basic steps are outlined below, but refer to the latest documentation for the proper procedure.
   a. Run the `host_config` utility to update the `sd.conf/sdd.conf` file. Doing so allows the SCSI drives to correctly discover ONTAP LUNs.
   b. Follow the instructions given by the `host_config` utility to enable multipath input/output (MPIO).
   c. Reboot. This step is required so that any changes are recognized across the system.
4. Partition the LUNs and verify that they are properly aligned. See "Appendix B: WAFL Alignment Verification" for instructions on how to directly test and confirm alignment.

zpools

A zpool should only be created after the steps in the LUN Configuration are performed. If the procedure is not done correctly, it can result in serious performance degradation due to the I/O alignment. Optimum performance on ONTAP requires I/O to be aligned to a 4K boundary on a drive. The file systems created on a zpool use an effective block size that is controlled through a parameter called `ashift`, which can be viewed by running the command `zdb -C`.

The value of `ashift` defaults to 9, which means 2^9, or 512 bytes. For optimum performance, the `ashift` value must be 12 (2^12=4K). This value is set at the time the zpool is created and cannot be changed, which means that data in zpools with `ashift` other than 12 should be migrated by copying data to a newly created zpool.

After creating a zpool, verify the value of `ashift` before proceeding. If the value is not 12, the LUNs were not discovered correctly. Destroy the zpool, verify that all steps shown in the relevant Host Utilities documentation were performed correctly, and recreate the zpool.

zpools and Solaris LDOMs

Solaris LDOMs create an additional requirement for making sure that I/O alignment is correct. Although a LUN might be properly discovered as a 4K device, a virtual vdsk device on an LDOM does not inherit the configuration from the I/O domain. The vdsk based on that LUN defaults back to a 512-byte block.

An additional configuration file is required. First, the individual LDOM's must be patched for Oracle bug 15824910 to enable the additional configuration options. This patch has been ported into all currently used versions of Solaris. Once the LDOM is patched, it is ready for configuration of the new properly aligned LUNs as follows:

1. Identify the LUN or LUNs to be used in the new zpool. In this example, it is the c2d1 device.

```
[root@LDOM1 ~]# echo | format
Searching for disks...done
AVAILABLE DISK SELECTIONS:
  0. c2d0 <Unknown-Unknown-0001-100.00GB>
     /virtual-devices@100/channel-devices@200/disk@0
  1. c2d1 <SUN-ZFS Storage 7330-1.0 cyl 1623 alt 2 hd 254 sec 254>
     /virtual-devices@100/channel-devices@200/disk@1
```

2. Retrieve the vdc instance of the devices to be used for a ZFS pool:

```
[root@LDOM1 ~]#  cat /etc/path_to_inst
#
# Caution! This file contains critical kernel state
#
"/fcoe" 0 "fcoe"
"/iscsi" 0 "iscsi"
"/pseudo" 0 "pseudo"
"/scsi_vhci" 0 "scsi_vhci"
"/options" 0 "options"
"/virtual-devices@100" 0 "vnex"
"/virtual-devices@100/channel-devices@200" 0 "cnex"
"/virtual-devices@100/channel-devices@200/disk@0" 0 "vdc"
"/virtual-devices@100/channel-devices@200/pciv-communication@0" 0 "vpci"
"/virtual-devices@100/channel-devices@200/network@0" 0 "vnet"
"/virtual-devices@100/channel-devices@200/network@1" 1 "vnet"
"/virtual-devices@100/channel-devices@200/network@2" 2 "vnet"
"/virtual-devices@100/channel-devices@200/network@3" 3 "vnet"
"/virtual-devices@100/channel-devices@200/disk@1" 1 "vdc" << We want
this one
```

3. Edit `/platform/sun4v/kernel/drv/vdc.conf`:

```
block-size-list="1:4096";
```

This means that device instance 1 is assigned a block size of 4096.

As an additional example, assume vdsk instances 1 through 6 need to be configured for a 4K block size and `/etc/path_to_inst` reads as follows:

```
"/virtual-devices@100/channel-devices@200/disk@1" 1 "vdc"
"/virtual-devices@100/channel-devices@200/disk@2" 2 "vdc"
"/virtual-devices@100/channel-devices@200/disk@3" 3 "vdc"
"/virtual-devices@100/channel-devices@200/disk@4" 4 "vdc"
"/virtual-devices@100/channel-devices@200/disk@5" 5 "vdc"
"/virtual-devices@100/channel-devices@200/disk@6" 6 "vdc"
```

4. The final `vdc.conf` file should contain the following:

```
block-size-list="1:8192","2:8192","3:8192","4:8192","5:8192","6:8192";
```

Caution
The LDOM must be rebooted after vdc.conf is configured and the vdsk is created. This step cannot be avoided. The block size change only takes effect after a reboot. Proceed with zpool configuration and ensure that ashift is properly set to 12 as described previously.

ZFS Intent Log (ZIL)

Generally, there is no reason to locate the ZFS Intent Log (ZIL) on a different device. The log can share space with the main pool. The primary use of a separate ZIL is when using physical drives that lack the write caching features in modern storage arrays.

logbias

Set the `logbias` parameter on ZFS file systems hosting Oracle data.

```
zfs set logbias=throughput <filesystem>
```

Using this parameter reduces overall write levels. Under the defaults, written data is committed first to the ZIL and then to the main storage pool. This approach is appropriate for a configuration using a plain drive configuration, which includes an SSD-based ZIL device and spinning media for the main storage pool. This is because it allows a commit to occur in a single I/O transaction on the lowest latency media available.

When using a modern storage array that includes its own caching capability, this approach is not generally necessary. Under rare circumstances, it might be desirable to commit a write with a single transaction to the log, such as a workload that consists of highly concentrated, latency-sensitive random writes. There are consequences in the form of write amplification because the logged data is eventually written to the main storage pool, resulting in a doubling of the write activity.

Direct I/O

Many applications, including Oracle products, can bypass the host buffer cache by enabling direct I/O. This strategy does not work as expected with ZFS file systems. Although the host buffer cache is bypassed, ZFS itself continues to cache data. This action can result in misleading results when using tools such as fio or sio to perform performance tests because it is difficult to predict whether I/O is reaching the storage system or whether it is being cached locally within the OS. This action also makes it very difficult to use such synthetic tests to compare ZFS performance to other file systems. As a practical matter, there is little to no difference in

file system performance under real user workloads.

Multiple zpools

Snapshot-based backups, restores, clones, and archiving of ZFS-based data must be performed at the level of the zpool and typically requires multiple zpools. A zpool is analogous to an LVM disk group and should be configured using the same rules. For example, a database is probably best laid out with the datafiles residing on `zpool1` and the archive logs, control files, and redo logs residing on `zpool2`. This approach permits a standard hot backup in which the database is placed in hot backup mode, followed by a snapshot of `zpool1`. The database is then removed from hot backup mode, the log archive is forced, and a snapshot of `zpool2` is created. A restore operation requires unmounting the zfs file systems and offlining the zpool in its entirety, following by a SnapRestore restore operation. The zpool can then be brought online again and the database recovered.

filesystemio_options

The Oracle parameter `filesystemio_options` works differently with ZFS. If `setall` or `directio` is used, write operations are synchronous and bypass the OS buffer cache, but reads are buffered by ZFS. This action causes difficulties in performance analysis because I/O is sometimes intercepted and serviced by the ZFS cache, making storage latency and total I/O less than it might appear to be.

Network configuration

Logical interface design for Oracle databases

Oracle databases need access to storage. Logical interfaces (LIFs) are the network plumbing that connects a storage virtual machine (SVM) to the network and in turn to the database. Proper LIF design is required to ensure sufficient bandwidth exists for each database workload, and failover does not result in a loss of storage services.

This section provides an overview of key LIF design principles. For more comprehensive documentation, see the ONTAP Network Management documentation. As with other aspects of database architecture, the best options for storage virtual machine (SVM, known as a vserver at the CLI) and logical interface (LIF) design depend heavily on scaling requirements and business needs.

Consider the following primary topics when building a LIF strategy:

- **Performance.** Is the network bandwidth sufficient?
- **Resiliency.** Are there any single points of failure in the design?
- **Manageability.** Can the network be scaled nondisruptively?

These topics apply to the end-to-end solution, from the host through the switches to the storage system.

LIF types

There are multiple LIF types. ONTAP documentation on LIF types provide more complete information on this topic, but from a functional perspective LIFs can be divided into the following groups:

- **Cluster and node management LIFs.** LIFs used to manage the storage cluster.
- **SVM management LIFs.** Interfaces that permit access to an SVM through the REST API or ONTAPI (also known as ZAPI) for functions such as snapshot creation or volume resizing. Products such as SnapManager for Oracle (SMO) must have access to an SVM management LIF.

- **Data LIFs.** Interfaces for FC, iSCSI, NVMe/FC, NVMe/TCP, NFS, or SMB/CIFS data.

> A data LIF used for NFS traffic can also be used for management by changing the firewall policy from `data` to `mgmt` or another policy that allows HTTP, HTTPS, or SSH. This change can simplify network configuration by avoiding the configuration of each host for access to both the NFS data LIF and a separate management LIF. It is not possible to configure an interface for both iSCSI and management traffic, despite the fact that both use an IP protocol. A separate management LIF is required in iSCSI environments.

SAN LIF design

LIF design in a SAN environment is relatively simple for one reason: multipathing. All modern SAN implementations allow a client to access data over multiple, independent, network paths and select the best path or paths for access. As a result, performance with respect to LIF design is simpler to address because SAN clients automatically load-balance I/O across the best available paths.

If a path becomes unavailable, the client automatically selects a different path. The resulting simplicity of design makes SAN LIFs generally more manageable. This does not mean that a SAN environment is always more easily managed, because there are many other aspects of SAN storage that are much more complicated than NFS. It simply means that SAN LIF design is easier.

Performance

The most important consideration with LIF performance in a SAN environment is bandwidth. For example, a two-node ONTAP AFF cluster with two 16Gb FC ports per node allows up to 32Gb of bandwidth to/from each node.

Resiliency

SAN LIFs do not fail over on an AFF storage system. If a SAN LIF fails because of controller failover, then the client's multipathing software detects the loss of a path and redirects I/O to a different LIF. With ASA storage systems, LIFs will be failed over after a short delay, but this does not interrupt IO because there are already active paths on the other controller. The failover process occurs in order to restore host access on all defined ports.

Manageability

LIF migration is a much more common task in an NFS environment because LIF migration is often associated with relocating volumes around the cluster. There is no need to migrate a LIF in a SAN environment when volumes are relocated within the HA pair. That is because, after the volume move has completed, ONTAP sends a notification to the SAN about a change in paths, and the SAN clients automatically reoptimize. LIF migration with SAN is primarily associated with major physical hardware changes. For example, if a nondisruptive upgrade of the controllers is required, a SAN LIF is migrated to the new hardware. If an FC port is found to be faulty, a LIF can be migrated to an unused port.

Design recommendations

NetApp makes the following recommendations:

- Do not create more paths than are required. Excessive numbers of paths make overall management more complicated and can cause problems with path failover on some hosts. Furthermore, some hosts have unexpected path limitations for configurations such as SAN booting.
- Very few configurations should require more than four paths to a LUN. The value of having more than two nodes advertising paths to LUNs is limited because the aggregate hosting a LUN is inaccessible if the

node that owns the LUN and its HA partner fail. Creating paths on nodes other than the primary HA pair is not helpful in such a situation.

- Although the number of visible LUN paths can be managed by selecting which ports are included in FC zones, it is generally easier to include all potential target points in the FC zone and control LUN visibility at the ONTAP level.
- In ONTAP 8.3 and later, the selective LUN mapping (SLM) feature is the default. With SLM, any new LUN is automatically advertised from the node that owns the underlying aggregate and the node's HA partner. This arrangement avoids the need to create port sets or configure zoning to limit port accessibility. Each LUN is available on the minimum number of nodes required for both optimal performance and resiliency. *In the event a LUN must be migrated outside of the two controllers, the additional nodes can be added with the `lun mapping add-reporting-nodes` command so that the LUNs are advertised on the new nodes. Doing so creates additional SAN paths to the LUNs for LUN migration. However, the host must perform a discovery operation to use the new paths.
- Do not be overly concerned about indirect traffic. It is best to avoid indirect traffic in a very I/O-intensive environment for which every microsecond of latency is critical, but the visible performance effect is negligible for typical workloads.

NFS LIF design

In contrast to SAN protocols, NFS has a limited ability to define multiple paths to data. The parallel NFS (pNFS) extensions to NFSv4 address this limitation, but as ethernet speeds have reached 100Gb and beyond there is rarely value in adding additional paths.

Performance and resiliency

Although measuring SAN LIF performance is primarily a matter of calculating the total bandwidth from all primary paths, determining NFS LIF performance requires taking a closer look at the exact network configuration. For example, two 10Gb ports can be configured as raw physical ports, or they can be configured as a Link Aggregation Control Protocol (LACP) interface group. If they are configured as an interface group, multiple load balancing policies are available that work differently depending on whether traffic is switched or routed. Finally, Oracle direct NFS (dNFS) offers load-balancing configurations that do not exist in any OS NFS clients at this time.

Unlike SAN protocols, NFS file systems require resiliency at the protocol layer. For example, a LUN is always configured with multipathing enabled, meaning that multiple redundant channels are available to the storage system, each of which uses the FC protocol. An NFS file system, on the other hand, depends on the availability of a single TCP/IP channel that can only be protected at the physical layer. This arrangement is why options such as port failover and LACP port aggregation exist.

In an NFS environment, both performance and resiliency are provided at the network protocol layer. As a result, both topics are intertwined and must be discussed together.

Bind LIFs to port groups

To bind a LIF to a port group, associate the LIF IP address with a group of physical ports. The primary method for aggregating physical ports together is LACP. The fault-tolerance capability of LACP is fairly simple; each port in an LACP group is monitored and is removed from the port group in the event of a malfunction. There are, however, many misconceptions about how LACP works with respect to performance:

- LACP does not require the configuration on the switch to match the endpoint. For example, ONTAP can be configured with IP-based load balancing, while a switch can use MAC-based load balancing.
- Each endpoint using an LACP connection can independently choose the packet transmission port, but it cannot choose the port used for receipt. This means that traffic from ONTAP to a particular destination is

tied to a particular port, and the return traffic could arrive on a different interface. This does not cause problems, however.

- LACP does not evenly distribute traffic all the time. In a large environment with many NFS clients, the result is typically even use of all ports in an LACP aggregation. However, any one NFS file system in the environment is limited to the bandwidth of only one port, not the entire aggregation.
- Although robin-robin LACP policies are available on ONTAP, these policies do not address the connection from a switch to a host. For example, a configuration with a four-port LACP trunk on a host and a four-port LACP trunk on ONTAP is still only able to read a file system using a single port. Although ONTAP can transmit data through all four ports, no switch technologies are currently available that send from the switch to the host through all four ports. Only one is used.

The most common approach in larger environments consisting of many database hosts is to build an LACP aggregate of an appropriate number of 10Gb (or faster) interfaces by using IP load balancing. This approach enables ONTAP to deliver even use of all ports, as long as enough clients exist. Load balancing breaks down when there are fewer clients in the configuration because LACP trunking does not dynamically redistribute load.

When a connection is established, traffic in a particular direction is placed on only one port. For example, a database performing a full table scan against an NFS file system connected through a four-port LACP trunk reads data though only one network interface card (NIC). If only three database servers are in such an environment, it is possible that all three are reading from the same port, while the other three ports are idle.

Bind LIFs to physical ports

Binding a LIF to a physical port results in more granular control over network configuration because a given IP address on a ONTAP system is associated with only one network port at a time. Resiliency is then accomplished through the configuration of failover groups and failover policies.

Failover policies and failover groups

The behavior of LIFs during network disruption is controlled by failover policies and failover groups. Configuration options have changed with the different versions of ONTAP. Consult the ONTAP network management documentation for failover groups and policies for specific details for the version of ONTAP being deployed.

ONTAP 8.3 and higher allow management of LIF failover based on broadcast domains. Therefore, an administrator can define all of the ports that have access to a given subnet and allow ONTAP to select an appropriate failover LIF. This approach can be used by some customers, but it has limitations in a high-speed storage network environment because of the lack of predictability. For example, an environment can include both 1Gb ports for routine file system access and 10Gb ports for datafile I/O. If both types of ports exist in the same broadcast domain, LIF failover can result in moving datafile I/O from a 10Gb port to a 1Gb port.

In summary, consider the following practices:

1. Configure a failover group as user-defined.
2. Populate the failover group with ports on the storage failover (SFO) partner controller so that the LIFs follow the aggregates during a storage failover. This avoids creating indirect traffic.
3. Use failover ports with matching performance characteristics to the original LIF. For example, a LIF on a single physical 10Gb port should include a failover group with a single 10Gb port. A four-port LACP LIF should fail over to another four-port LACP LIF. These ports would be a subset of the ports defined in the broadcast domain.
4. Set the failover policy to SFO-partner only. Doing so makes sure that the LIF follows the aggregate during failover.

Auto-revert

Set the `auto-revert` parameter as desired. Most customers prefer to set this parameter to `true` to have the LIF revert to its home port. However, in some cases, customers have set this to `false `so that an unexpected failover can be investigated before returning a LIF to its home port.

LIF-to-volume ratio

A common misconception is that there must be a 1:1 relationship between volumes and NFS LIFs. Although this configuration is required for moving a volume anywhere in a cluster while never creating additional interconnect traffic, it is categorically not a requirement. Intercluster traffic must be considered, but the mere presence of intercluster traffic does not create problems. Many of the published benchmarks created for ONTAP include predominantly indirect I/O.

For example, a database project containing a relatively small number of performance-critical databases that only required a total of 40 volumes might warrant a 1:1 volume to LIF strategy, an arrangement that would require 40 IP addresses. Any volume could then be moved anywhere in the cluster along with the associated LIF, and traffic would always be direct, minimizing every source of latency even at microsecond levels.

As a counter example, a large, hosted environment might be more easily managed with a 1:1 relationship between customers and LIFs. Over time, a volume might need to be migrated to a different node, which would cause some indirect traffic. However, the performance effect should be undetectable unless the network ports on the interconnect switch are saturating. If there is concern, a new LIF can be established on additional nodes and the host can be updated at the next maintenance window to remove indirect traffic from the configuration.

TCP/IP and ethernet configuration for Oracle databases

Many Oracle on ONTAP customers use ethernet, the network protocol of NFS, iSCSI, NVMe/TCP, and especially the cloud.

Host OS settings

Most application vendor documentation include specific TCP and ethernet settings intended to ensure the application is working optimally. These same settings are usually sufficient to also deliver optimal IP-based storage performance.

Ethernet flow control

This technology allows a client to request that a sender temporarily stop data transmission. This is usually done because the receiver is unable to process incoming data quickly enough. At one time, requesting that a sender cease transmission was less disruptive than having a receiver discard packets because buffers were full. This is no longer the case with the TCP stacks used in OSs today. In fact, flow control causes more problems than it solves.

Performance problems caused by Ethernet flow control have been increasing in recent years. This is because Ethernet flow control operates at the physical layer. If a network configuration permits any host OS to send an Ethernet flow control request to a storage system, the result is a pause in I/O for all connected clients. Because an increasing number of clients are served by a single storage controller, the likelihood of one or more of these clients sending flow control requests increases. The problem has been seen frequently at customer sites with extensive OS virtualization.

A NIC on a NetApp system should not receive flow-control requests. The method used to achieve this result varies based on the network switch manufacturer. In most cases, flow control on an Ethernet switch can be set to `receive desired` or `receive on`, which means that a flow control request is not forwarded to the

storage controller. In other cases, the network connection on the storage controller might not allow flow-control disabling. In these cases, the clients must be configured to never send flow control requests, either by changing to the NIC configuration on the host server itself or the switch ports to which the host server is connected.

NetApp recommends making sure that NetApp storage controllers do not receive Ethernet flow-control packets. This can generally be done by setting the switch ports to which the controller is attached, but some switch hardware has limitations that might require client-side changes instead.

MTU Sizes

The use of jumbo frames has been shown to offer some performance improvement in 1Gb networks by reducing CPU and network overhead, but the benefit is not usually significant.

NetApp recommends implementing jumbo frames when possible, both to realize any potential performance benefits and to future-proof the solution.

Using jumbo frames in a 10Gb network is almost mandatory. This is because most 10Gb implementations reach a packets-per-second limit without jumbo frames before they reach the 10Gb mark. Using jumbo frames improves efficiency in TCP/IP processing because it allows the OS, server, NICs, and the storage system to process fewer but larger packets. The performance improvement varies from NIC to NIC, but it is significant.

For jumbo-frame implementations, there is the common but incorrect belief that all connected devices must support jumbo frames and that the MTU size must match end-to-end. Instead, the two network end points negotiate the highest mutually acceptable frame size when establishing a connection. In a typical environment, a network switch is set to an MTU size of 9216, the NetApp controller is set to 9000, and the clients are set to a mix of 9000 and 1514. Clients that can support an MTU of 9000 can use jumbo frames, and clients that can only support 1514 can negotiate a lower value.

Problems with this arrangement are rare in a completely switched environment. However, take care in a routed environment that no intermediate router is forced to fragment jumbo frames.

NetApp recommends configuring the following:

- Jumbo frames are desirable but not required with 1Gb Ethernet (GbE).
- Jumbo frames are required for maximum performance with 10GbE and faster.

TCP parameters

Three settings are often misconfigured: TCP timestamps, selective acknowledgment (SACK), and TCP window scaling. Many out-of-date documents on the Internet recommend disabling one or more of these parameters to improve performance. There was some merit to this recommendation many years ago when CPU capabilities were much lower and there was a benefit to reducing the overhead on TCP processing whenever possible.

However, with modern OSs, disabling any of these TCP features usually results in no detectable benefit while also potentially damaging performance. Performance damage is especially likely in virtualized networking environments because these features are required for efficient handling of packet loss and changes in network quality.

NetApp recommends enabling TCP timestamps, SACK, and TCP window scaling on the host, and all three of these parameters should be on by default in any current OS.

FC configuration for Oracle databases

Configuring FC SAN for Oracle databases is primarily about following everyday SAN best practices.

This includes typical planning measures such as ensuring sufficient bandwidth exists on the SAN in between the host and storage system, checking that all SAN paths exist between all required devices, using the FC port settings required by your FC switch vendor, avoiding ISL contention, and using proper SAN fabric monitoring.

Zoning

An FC zone should never contain more than one initiator. Such an arrangement might appear to work initially, but crosstalk between initiators eventually interferes with performance and stability.

Multitarget zones are generally regarded as safe, although in rare circumstances the behavior of FC target ports from different vendors has caused problems. For example, avoid including the target ports from both a NetApp and an non-NetApp storage array in the same zone. In addition, placing a NetApp storage system and a tape device in the same zone is even more likely to cause problems.

Oracle database and direct-connect ONTAP connectivity

Storage administrators sometimes prefer to simplify their infrastructures by removing network switches from the configuration. This can be supported in some scenarios.

iSCSI and NVMe/TCP

A host using iSCSI or NVMe/TCP can be directly connected to a storage system and operate normally. The reason is pathing. Direct connections to two different storage controllers results in two independent paths for data flow. The loss of path, port, or controller does not prevent the other path from being used.

NFS

Direct-connected NFS storage can be used, but with a significant limitation - failover will not work without a significant scripting effort, which would be the responsibility of the customer.

The reason nondisruptive failover is complicated with direct-connected NFS storage is the routing that occurs on the local OS. For example, assume a host has an IP address of 192.168.1.1/24 and is directly connected to an ONTAP controller with an IP address of 192.168.1.50/24. During failover, that 192.168.1.50 address can fail over to the other controller, and it will be available to the host, but how does the host detect its presence? The original 192.168.1.1 address still exists on the host NIC that no longer connects to an operational system. Traffic destined for 192.168.1.50 would continue to be sent to an inoperable network port.

The second OS NIC could be configured as 19 2.168.1.2 and would be capable of communicating with the failed over 192.168.1.50 address, but the local routing tables would have a default of using one **and only one** address to communicate with the 192.168.1.0/24 subnet. A sysadmin could create a scripting framework that would detect a failed network connection and alter the local routing tables or bring interfaces up and down. The exact procedure would depend on the OS in use.

In practice, NetApp customers do have direct-connected NFS, but normally only for workloads where IO pauses during failovers are acceptable. When hard mounts are used, there should not be any IO errors during such pauses. The IO should hang until services are restored, either by a failback or manual intervention to move IP addresses between NICs on the host.

FC direct connect

It is not possible to directly connect a host to an ONTAP storage system using the FC protocol. The reason is the use of NPIV. The WWN that identifies an ONTAP FC port to the FC network uses a type of virtualization called NPIV. Any device connected to an ONTAP system must be able to recognize an NPIV WWN. There are no current HBA vendors who offer an HBA that can be installed in a host that would be able to support an NPIV target.

Storage configuration

FC SAN

LUN Alignment for Oracle database I/O

LUN alignment refers to optimizing I/O with respect to the underlying file system layout.

On a ONTAP system, storage is organized in 4KB units. A database or file system 8KB block should map to exactly two 4KB blocks. If an error in LUN configuration shifts the alignment by 1KB in either direction, each 8KB block would exist on three different 4KB storage blocks rather than two. This arrangement would cause increased latency and cause additional I/O to be performed within the storage system.

Alignment also affects LVM architectures. If a physical volume within a logical volume group is defined on the whole drive device (no partitions are created), the first 4KB block on the LUN aligns with the first 4KB block on the storage system. This is a correct alignment. Problems arise with partitions because they shift the starting location where the OS uses the LUN. As long as the offset is shifted in whole units of 4KB, the LUN is aligned.

In Linux environments, build logical volume groups on the whole drive device. When a partition is required, check alignment by running `fdisk -u` and verifying that the start of each partition is a multiple of eight. This means that the partition starts at a multiple of eight 512-byte sectors, which is 4KB.

Also see the discussion about compression block alignment in the section Efficiency. Any layout that is aligned with 8KB compression block boundaries is also aligned with 4KB boundaries.

Misalignment warnings

Database redo/transaction logging normally generates unaligned I/O that can cause misleading warnings about misaligned LUNs on ONTAP.

Logging performs a sequential write of the log file with writes of varying size. A log write operation that does not align to 4KB boundaries does not ordinarily cause performance problems because the next log write operation completes the block. The result is that ONTAP is able to process almost all writes as complete 4KB blocks, even though the data in some 4KB blocks was written in two separate operations.

Verify alignment by using by using utilities such as `sio` or `dd` that can generate I/O at a defined block size. The I/O alignment statistics on the storage system can be viewed with the `stats` command. See WAFL alignment verification for more information.

Alignment in Solaris environments is more complicated. Refer to ONTAP SAN Host Configuration for more information.

Caution
In Solaris x86 environments, take additional care about proper alignment because most configurations have several layers of partitions. Solaris x86 partition slices usually exist on top of a standard master boot record partition table.

Oracle database LUN sizing and LUN count

Selecting the optimal LUN size and the number of LUNs to be used is critical for optimal performance and manageability of Oracle databases.

A LUN is a virtualized object on ONTAP that exists across all of the drives in the hosting aggregate. As a result, the performance of the LUN is unaffected by its size because the LUN draws on the full performance potential of the aggregate no matter which size is chosen.

As a matter of convenience, customers might wish to use a LUN of a particular size. For example, if a database is built on an LVM or Oracle ASM diskgroup composed of two LUNs of 1TB each, then that diskgroup must be grown in increments of 1TB. It might be preferable to build the diskgroup from eight LUNs of 500GB each so that the diskgroup can be increased in smaller increments.

The practice of establishing a universal standard LUN size is discouraged because doing so can complicate manageability. For example, a standard LUN size of 100GB might work well when a database or datastore is in the range of 1TB to 2TB, but a database or datastore of 20TB in size would require 200 LUNs. This means that server reboot times are longer, there are more objects to manage in the various UIs, and products such as SnapCenter must perform discovery on many objects. Using fewer, larger LUNs avoids such problems.

- The LUN count is more important than the LUN size.
- LUN size is mostly controlled by LUN count requirements.
- Avoid creating more LUNs than required.

LUN count

Unlike the LUN size, the LUN count does affect performance. Application performance often depends on the ability to perform parallel I/O through the SCSI layer. As a result, two LUNs offer better performance than a single LUN. Using an LVM such as Veritas VxVM, Linux LVM2, or Oracle ASM is the simplest method to increase parallelism.

NetApp customers have generally experienced minimal benefit from increasing the number of LUNs beyond sixteen, although the testing of 100%-SSD environments with very heavy random I/O has demonstrated further improvement up to 64 LUNs.

NetApp recommends the following:

In general, four to sixteen LUNs are sufficient to support the I/O needs of any given database workload. Less than four LUNs might create performance limitations because of limitations in host SCSI implementations.

Oracle database LUN resizing and LVM-based resizing

When a SAN-based file system has reached its capacity limit, there are two options for increasing the space available:

- Increase the size of the LUNs
- Add a LUN to an existing volume group and grow the contained logical volume

Although LUN resizing is an option to increase capacity, it is generally better to use an LVM, including Oracle ASM. One of the principal reasons LVMs exist is to avoid the need for a LUN resize. With an LVM, multiple LUNs are bonded together into a virtual pool of storage. The logical volumes carved out of this pool are managed by the LVM and can be easily resized. An additional benefit is the avoidance of hotspots on a particular drive by distributing a given logical volume across all available LUNs. Transparent migration can usually be performed by using the volume manager to relocate the underlying extents of a logical volume to new LUNs.

LVM striping with Oracle databases

LVM striping refers to distributing data across multiple LUNs. The result is dramatically improved performance for many databases.

Before the era of flash drives, striping was used to help overcome the performance limitations of spinning drives. For example, if an OS needs to perform a 1MB read operation, reading that 1MB of data from a single drive would require a lot of drive head seeking and reading as the 1MB is slowly transferred. If that 1MB of data was striped across 8 LUNs, the OS could issue eight 128K read operations in parallel and reduce the time required to complete the 1MB transfer.

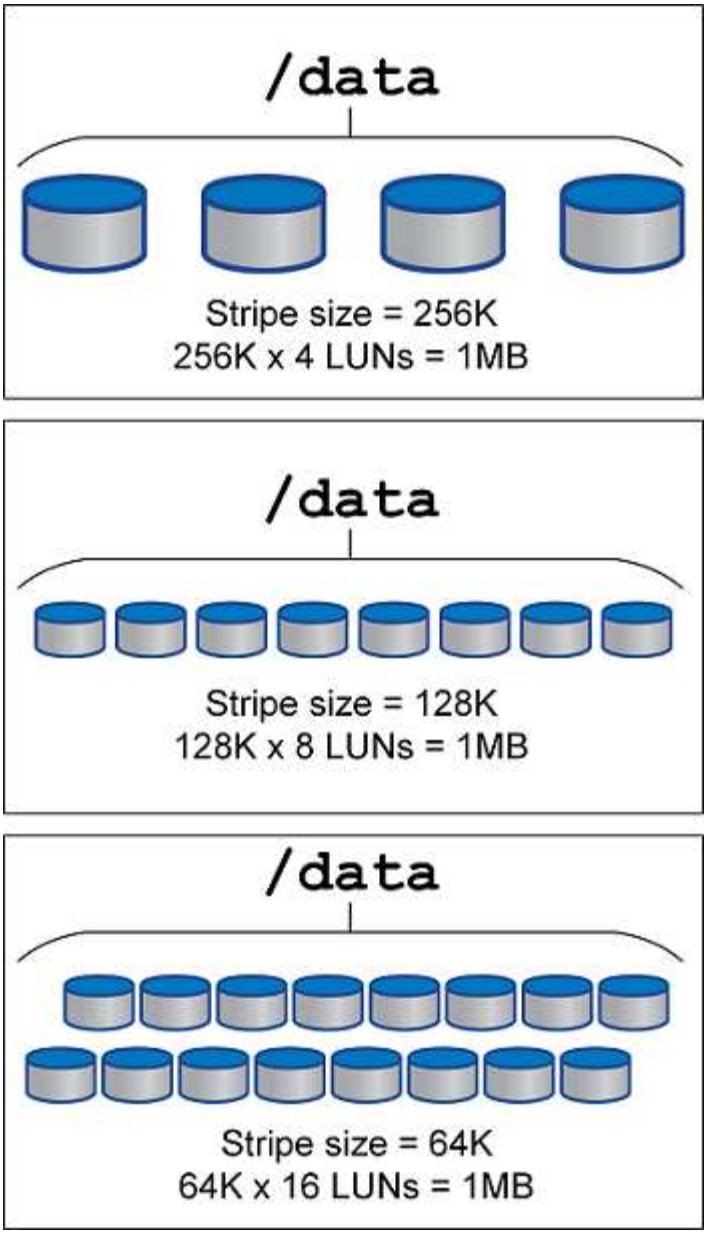
Striping with spinning drives was more difficult because the I/O pattern had to be known in advance. If the striping wasn’t correctly tuned for the true I/O patterns, striped configurations could damage performance. With Oracle databases, and especially with all-flash configurations, striping is much easier to configure and has been proven to dramatically improve performance.

Logical volume managers such as Oracle ASM stripe by default, but native OS LVM do not. Some of them bond multiple LUNs together as a concatenated device, which results in datafiles that exist on one and only one LUN device. This causes hot spots. Other LVM implementations default to distributed extents. This is similar to striping, but it’s coarser. The LUNs in the volume group are sliced into large pieces, called extents and typically measured in many megabytes, and the logical volumes are then distributed across those extents. The result is random I/O against a file should be well distributed across LUNs, but sequential I/O operations are not as efficient as they could be.

Performance-intensive application I/O is nearly always either (a) in units of the basic block size or (b) one megabyte.

The primary goal of a striped configuration is to ensure that single-file I/O can be performed as a single unit, and multiblock I/Os, which should be 1MB in size, can be parallelized evenly across all LUNs in the striped volume. This means that the stripe size must not be smaller than the database block size, and the stripe size multiplied by the number of LUNs should be 1MB.

The following figure shows three possible options for stripe size and width tuning. The number of LUNs is selected to meet performance requirements as described above, but in all cases the total data within a single stripe is 1MB.



NFS

NFS configuration for Oracle databases

NetApp has been providing enterprise-grade NFS storage for over 30 years, and its use is growing with the push toward cloud-based infrastructures because of it's simplicity.

The NFS protocol includes multiple versions with varying requirements. For a complete description of NFS configuration with ONTAP, please see TR-4067 NFS on ONTAP Best Practices. The following sections cover some of the more critical requirements and common user errors.

NFS versions

The operating system NFS client must be supported by NetApp.

- NFSv3 is supported with OSs that follow the NFSv3 standard.

- NFSv3 is supported with the Oracle dNFS client.
- NFSv4 is supported with all OSs that follow the NFSv4 standard.
- NFSv4.1 and NFSv4.2 require specific OS support. Consult the NetApp IMT for supported OSs.
- Oracle dNFS support for NFSv4.1 requires Oracle 12.2.0.2 or higher.

> The NetApp support matrix for NFSv3 and NFSv4 does not include specific operating systems. All OSs that obey the RFC are generally supported. When searching the online IMT for NFSv3 or NFSv4 support, do not select a specific OS because there will be no matches displayed. All OSs are implicitly supported by the general policy.

Linux NFSv3 TCP slot tables

TCP slot tables are the NFSv3 equivalent of host bus adapter (HBA) queue depth. These tables control the number of NFS operations that can be outstanding at any one time. The default value is usually 16, which is far too low for optimum performance. The opposite problem occurs on newer Linux kernels, which can automatically increase the TCP slot table limit to a level that saturates the NFS server with requests.

For optimum performance and to prevent performance problems, adjust the kernel parameters that control the TCP slot tables.

Run the `sysctl -a | grep tcp.*.slot_table` command, and observe the following parameters:

```
# sysctl -a | grep tcp.*.slot_table
sunrpc.tcp_max_slot_table_entries = 128
sunrpc.tcp_slot_table_entries = 128
```

All Linux systems should include `sunrpc.tcp_slot_table_entries`, but only some include `sunrpc.tcp_max_slot_table_entries`. They should both be set to 128.

Caution
Failure to set these parameters may have significant effects on performance. In some cases, performance is limited because the linux OS is not issuing sufficient I/O. In other cases, I/O latencies increases as the linux OS attempts to issue more I/O than can be serviced.

ADR and NFS

Some customers have reported performance problems resulting from an excessive amount of I/O on data in the `ADR` location. The problem does not generally occur until a lot of performance data has accumulated. The reason for the excessive I/O is unknown, but this problem appears to be a result of Oracle processes repeatedly scanning the target directory for changes.

Removal of the `noac` and/or `actimeo=0` mount options allows host OS caching to occur and reduces storage I/O levels.

> **NetApp recommends** to not place `ADR` data on a file system with `noac` or `actimeo=0` because performance problems are likely. Separate `ADR` data into a different mount point if necessary.

nfs-rootonly and mount-rootonly

ONTAP includes an NFS option called `nfs-rootonly` that controls whether the server accepts NFS traffic connections from high ports. As a security measure, only the root user is permitted to open TCP/IP connections using a source port below 1024 because such ports are normally reserved for OS use, not user processes. This restriction helps ensure that NFS traffic is from an actual operating system NFS client, and not a malicious process emulating an NFS client. The Oracle dNFS client is a userspace driver, but the process runs as root, so it is generally not required to change the value of `nfs-rootonly`. The connections is made from low ports.

The `mount-rootonly` option only applies to NFSv3. It controls whether the RPC MOUNT call be accepted from ports greater than 1024. When dNFS is used, the client is again running as root, so it able to open ports below 1024. This parameter has no effect.

Processes opening connections with dNFS over NFS versions 4.0 and higher do not run as root and therefore require ports over 1024. The `nfs-rootonly` parameter must be set to disabled for dNFS to complete the connection.

If `nfs-rootonly` is enabled, the result is a hang during the mount phase opening dNFS connections. The sqlplus output looks similar to this:

```
SQL>startup
ORACLE instance started.
Total System Global Area 4294963272 bytes
Fixed Size                  8904776 bytes
Variable Size             822083584 bytes
Database Buffers         3456106496 bytes
Redo Buffers                7868416 bytes
```

The parameter can be changed as follows:

```
Cluster01::> nfs server modify -nfs-rootonly disabled
```

> In rare situations, you might need to change both nfs-rootonly and mount-rootonly to disabled. If a server is managing an extremely large number of TCP connections, it is possible that no ports below 1024 is available, and the OS is forced to use higher ports. These two ONTAP parameters would need to be changed to allow the connection to complete.

NFS export polices: superuser and setuid

If Oracle binaries are located on an NFS share, the export policy must include superuser and setuid permissions.

Shared NFS exports used for generic file services such as user home directories usually squash the root user. This means a request from the root user on a host that has mounted a filesystem is remapped as a different user with lower privileges. This helps secure data by preventing a root user on a particular server from accessing data on the shared server. The setuid bit can also be a security risk on a shared environment. The setuid bit allows a process to be run as a different user than the user invoking the command. For example, a shell script that was owned by root with the setuid bit runs as root. If that shell script could be changed by other users, any non-root user could issue a command as root by updating the script.

The Oracle binaries include files owned by root and use the setuid bit. If Oracle binaries are installed on an NFS share, the export policy must include the appropriate superuser and setuid permissions. In the example below, the rule includes both `allow-suid` and permits `superuser` (root) access for NFS clients using system authentication.

```
Cluster01::> export-policy rule show -vserver vserver1 -policyname orabin
-fields allow-suid,superuser
vserver   policyname ruleindex superuser allow-suid
--------- ---------- --------- --------- ----------
vserver1  orabin     1         sys       true
```

NFSv4/4.1 configuration

For most applications, there is very little difference between NFSv3 and NFSv4. Application I/O is usually very simple I/O and does not benefit significantly from some of the advanced features available in NFSv4. Higher versions of NFS should not be viewed as an “upgrade” from a database storage perspective, but instead as versions of NFS that include additional features. For example, if the end-to-end security of kerberos privacy mode (krb5p) is required, then NFSv4 is required.

> **NetApp recommends** using NFSv4.1 if NFSv4 capabilities are required. There are some functional enhancements to the NFSv4 protocol in NFSv4.1 that improve resiliency in certain edge cases.

Switching to NFSv4 is more complicated than simply changing the mount options from vers=3 to vers=4.1. A more complete explanation of NFSv4 configuration with ONTAP, including guidance on configuring the OS, see TR-4067 NFS on ONTAP best practices. The following sections of this TR explain some of the basic requirements for using NFSv4.

NFSv4 domain

A complete explanation of NFSv4/4.1 configuration is beyond the scope of this document, but one commonly encountered problem is a mismatch in domain mapping. From a sysadmin point of view, the NFS file systems appear to behave normally, but applications report errors about permissions and/or setuid on the certain files. In some cases, administrators have incorrectly concluded that the permissions of the application binaries have been damaged and have run chown or chmod commands when the actual problem was the domain name.

The NFSv4 domain name is set on the ONTAP SVM:

```
Cluster01::> nfs server show -fields v4-id-domain
vserver   v4-id-domain
--------- ------------
vserver1  my.lab
```

The NFSv4 domain name on the host is set in `/etc/idmap.cfg`

```
[root@host1 etc]# head /etc/idmapd.conf
[General]
#Verbosity = 0
# The following should be set to the local NFSv4 domain name
# The default is the host's DNS domain name.
Domain = my.lab
```

The domain names must match. If they do not, mapping errors similar to the following appear in `/var/log/messages`:

```
Apr 12 11:43:08 host1 nfsidmap[16298]: nss_getpwnam: name 'root@my.lab'
does not map into domain 'default.com'
```

Application binaries, such as Oracle database binaries, include files owned by root with the setuid bit, which means a mismatch in the NFSv4 domain names causes failures with Oracle startup and a warning about the ownership or permissions of a file called `oradism`, which is located in the `$ORACLE_HOME/bin` directory. It should appear as follows:

```
[root@host1 etc]# ls -l /orabin/product/19.3.0.0/dbhome_1/bin/oradism
-rwsr-x--- 1 root oinstall 147848 Apr 17  2019
/orabin/product/19.3.0.0/dbhome_1/bin/oradism
```

If this file appears with ownership of nobody, there may be an NFSv4 domain mapping problem.

```
[root@host1 bin]# ls -l oradism
-rwsr-x--- 1 nobody oinstall 147848 Apr 17  2019 oradism
```

To fix this, check the `/etc/idmap.cfg` file against the v4-id-domain setting on ONTAP and ensure they are consistent. If they are not, make the required changes, run `nfsidmap -c`, and wait a moment for the changes to propagate. The file ownership should then be properly recognized as root. If a user had attempted to run `chown root` on this file before the NFS domains configure was corrected, it might be necessary to run `chown root` again.

Oracle directNFS

Oracle databases can use NFS in two ways.

First, it can use a filesystem mounted using the native NFS client that is part of the operating system. This is sometimes called kernel NFS, or kNFS. The NFS filesystem is mounted and used by the Oracle database exactly the same as any other application would use an NFS filesystem.

The second method is Oracle Direct NFS (dNFS). This is an implementation of the NFS standard within the Oracle database software. It does not change the way Oracle databases are configured or managed by the DBA. As long as the storage system itself has the correct settings, the use of dNFS should be transparent to the DBA team and end users.

A database with the dNFS feature enabled still has the usual NFS filesystems mounted. Once the database is open, the Oracle database opens a set of TCP/IP sessions and performs NFS operations directly.

Direct NFS

The primary value of Oracle's Direct NFS is to bypass the host NFS client and perform NFS file operations directly on an NFS server. Enabling it only requires changing the Oracle Disk Manager (ODM) library. Instructions for this process are provided in the Oracle documentation.

Using dNFS results in a significant improvement in I/O performance and decreases the load on the host and the storage system because I/O is performed in the most efficient way possible.

In addition, Oracle dNFS includes an **option** for network interface multipathing and fault-tolerance. For example, two 10Gb interfaces can be bound together to offer 20Gb of bandwidth. A failure of one interface results in I/O being retried on the other interface. The overall operation is very similar to FC multipathing. Multipathing was common years ago when 1Gb ethernet was the most common standard. A 10Gb NIC is sufficient for most Oracle workloads, but if more is required 10Gb NICs can be bonded.

When dNFS is used, it is critical that all patches described in Oracle Doc 1495104.1 are installed. If a patch cannot be installed, the environment must be evaluated to make sure that the bugs described in that document do not cause problems. In some cases, an inability to install the required patches prevents the use of dNFS.

Do not use dNFS with any type of round-robin name resolution, including DNS, DDNS, NIS or any other method. This includes the DNS load balancing feature available in ONTAP. When an Oracle database using dNFS resolves a host name to an IP address it must not change on subsequent lookups. This can result in Oracle database crashes and possible data corruption.

Direct NFS and host file system access

Using dNFS can occasionally cause problems for applications or user activities that rely on the visible file systems mounted on the host because the dNFS client accesses the file system out of band from the host OS. The dNFS client can create, delete, and modify files without the knowledge of the OS.

When the mount options for single-instance databases are used, they enable caching of file and directory attributes, which also means that the contents of a directory are cached. Therefore, dNFS can create a file, and there is a short lag before the OS rereads the directory contents and the file becomes visible to the user. This is not generally a problem, but, on rare occasions, utilities such as SAP BR*Tools might have issues. If this happens, address the problem by changing the mount options to use the recommendations for Oracle RAC. This change results in the disabling of all host caching.

Only change mount options when (a) dNFS is used and (b) a problem results from a lag in file visibility. If dNFS is not in use, using Oracle RAC mount options on a single-instance database results in degraded performance.

> See the note about `nosharecache` in Linux NFS mount options for a Linux-specific dNFS issue that can produce unusual results.

Oracle databases and NFS leases and locks

NFSv3 is stateless. That effectively means that the NFS server (ONTAP) doesn't keep track of which file systems are mounted, by whom, or which locks are truly in place.

ONTAP does have some features that will record mount attempts so you have an idea which clients may be accessing data, and there may be advisory locks present, but that information isn't guaranteed to be 100% complete. It can't be complete, because tracking NFS client state is not part of the NFSv3 standard.

NFSv4 statefulness

In contrast, NFSv4 is stateful. The NFSv4 server tracks which clients are using which file systems, which files exist, which files and/or regions of files are locked, etc. This means there needs to be regular communication between an NFSv4 server to keep the state data current.

The most important states being managed by the NFS server are NFSv4 Locks and NFSv4 Leases, and they are very much intertwined. You need to understand how each works by itself, and how they relate to one another.

NFSv4 locks

With NFSv3, locks are advisory. An NFS client can still modify or delete a "locked" file. An NFSv3 lock doesn't expire by itself, it must be removed. This creates problems. For example, if you have a clustered application that creates NFSv3 locks, and one of the nodes fails, what do you do? You can code the application on the surviving nodes to remove the locks, but how do you know that's safe? Maybe the "failed" node is operational, but isn't communicating with the rest of the cluster?

With NFSv4, locks have a limited duration. As long as the client holding the locks continues to check in with the NFSv4 server, no other client is permitted to acquire those locks. If a client fails to check in with the NFSv4, the locks eventually get revoked by the server and other clients will be able to request and obtain locks.

NFsv4 leases

NFSv4 locks are associated with an NFSv4 lease. When an NFSv4 client establishes a connection with an NFSv4 server, it gets a lease. If the client obtains a lock (there are many types of locks) then the lock is associated with the lease.

This lease has a defined timeout. By default, ONTAP will set the timeout value to 30 seconds:

```
Cluster01::*> nfs server show -vserver vserver1 -fields v4-lease-seconds

vserver   v4-lease-seconds
--------- ----------------
vserver1  30
```

This means that an NFSv4 client needs to check in with the NFSv4 server every 30 seconds to renew its leases.

The lease is automatically renewed by any activity, so if the client is doing work there's no need to perform addition operations. If an application becomes quiet and is not doing real work, it's going to need to perform a sort of keep-alive operation (called a SEQUENCE) instead. It's essentially just saying "I'm still here, please refresh my leases."

```
*Question:* What happens if you lose network connectivity for 31 seconds?
```

NFSv3 is stateless. It's not expecting communication from the clients. NFSv4 is stateful, and once that lease period elapses, the lease expires, and locks are revoked and the locked files are made available to other clients.

With NFSv3, you could move network cables around, reboot network switches, make configuration changes,

and be fairly sure that nothing bad would happen. Applications would normally just wait patiently for the network connection to work again.

With NFSv4, you have 30 seconds (unless you've increased the value of that parameter within ONTAP) to complete your work. If you exceed that, your leases time out. Normally this results in application crashes.

As an example, if you have an Oracle database, and you experience a loss of network connectivity (sometimes called a "network partition") that exceeds the lease timeout, you will crash the database.

Here's an example of what happens in the Oracle alert log if this happens:

```
2022-10-11T15:52:55.206231-04:00
Errors in file /orabin/diag/rdbms/ntap/NTAP/trace/NTAP_ckpt_25444.trc:
ORA-00202: control file: '/redo0/NTAP/ctrl/control01.ctl'
ORA-27072: File I/O error
Linux-x86_64 Error: 5: Input/output error
Additional information: 4
Additional information: 1
Additional information: 4294967295
2022-10-11T15:52:59.842508-04:00
Errors in file /orabin/diag/rdbms/ntap/NTAP/trace/NTAP_ckpt_25444.trc:
ORA-00206: error in writing (block 3, # blocks 1) of control file
ORA-00202: control file: '/redo1/NTAP/ctrl/control02.ctl'
ORA-27061: waiting for async I/Os failed
```

If you look at the syslogs, you should see several of these errors:

```
Oct 11 15:52:55 host1 kernel: NFS: nfs4_reclaim_open_state: Lock reclaim
failed!
Oct 11 15:52:55 host1 kernel: NFS: nfs4_reclaim_open_state: Lock reclaim
failed!
Oct 11 15:52:55 host1 kernel: NFS: nfs4_reclaim_open_state: Lock reclaim
failed!
```

The log messages are usually the first sign of a problem, other than the application freeze. Typically, you see nothing at all during the network outage because processes and the OS itself are blocked attempting to access the NFS file system.

The errors appear after the network is operational again. In the example above, once connectivity was reestablished, the OS attempted to reacquire the locks, but it was too late. The lease had expired and the locks were removed. That results in an error that propagates up to the Oracle layer and causes the message in the alert log. You might see variations on these patterns depending on the version and configuration of the database.

In summary, NFSv3 tolerates network interruption, but NFSv4 is more sensitive and imposes a defined lease period.

What if a 30 second timeout isn't acceptable? What if you manage a dynamically changing network where

switches are rebooted or cables are relocated and the result is the occasional network interruption? You could choose to extend the lease period, but whether you want to do that requires an explanation of NFSv4 grace periods.

NFSv4 grace periods

If an NFSv3 server is rebooted, it's ready to serve IO almost instantly. It was not maintaining any sort of state about clients. The result is that an ONTAP takeover operation often appears to be close to instantaneous. The moment a controller is ready to start serving data it will send an ARP to the network that signals the change in topology. Clients normally detect this almost instantly and data resumes flowing.

NFSv4, however, will produce a brief pause. It's just part of how NFSv4 works.

NFSv4 servers need to track the leases, locks, and who's using what data. If an NFS server panics and reboots, or loses power for a moment, or is restarted during maintenance activity, the result is the lease/lock and other client information is lost. The server needs to figure out which client is using what data before resuming operations. This is where the grace period comes in.

If you suddenly power cycle your NFSv4 server. When it comes back up, clients that attempt to resume IO will get a response that essentially says, "I have lost lease/lock information. Would you like to re-register your locks?" That's the start of the grace period. It defaults to 45 seconds on ONTAP:

```
Cluster01::> nfs server show -vserver vserver1 -fields v4-grace-seconds

vserver   v4-grace-seconds
--------- ----------------
vserver1  45
```

The result is that, after a restart, a controller will pause IO while all the clients reclaim their leases and locks. Once the grace period ends, the server will resume IO operations.

Lease timeouts vs grace periods

The grace period and the lease period are connected. As mentioned above, the default lease timeout is 30 seconds, which means NFSv4 clients must check in with the server at least every 30 seconds or they lose their leases and, in turn, their locks. The grace period exists to allow an NFS server to rebuild lease/lock data, and it defaults to 45 seconds. ONTAP requires the grace period to be 15 seconds longer than the lease period. This ensures that an NFS client environment that is designed to renew leases at least every 30 seconds will have the ability to check in with the server after a restart. A grace period of 45 seconds ensures that all those clients that expect to renew their leases at least every 30 seconds definitely have the opportunity to do so.

If a 30 second timeout isn't acceptable, you could choose to extend the lease period. If you want to increase the lease timeout to 60 seconds in order to withstand a 60 second network outage, you're going to have to increase the grace period to at least 75 seconds. ONTAP requires it to be 15 seconds higher than the lease period. That means you're going to experience longer IO pauses during controller failover.

This shouldn't normally be a problem. Typical users only update ONTAP controllers once or twice per year, and unplanned failover due to hardware failures are extremely rare. In addition, if you had a network where a 60-second network outage was a concerning possibility, and you needed to the lease timeout to 60 seconds, then you probably wouldn't object to rare storage system failover resulting in a 75 second pause either. You've already acknowledged you have a network that's pausing for 60+ seconds rather frequently.

NFS caching with Oracle databases

The presence of any of the following mount options causes host caching to be disabled:

```
cio, actimeo=0, noac, forcedirectio
```

These settings can have a severe negative effect on the speed of software installation, patching, and backup/restore operations. In some cases, especially with clustered applications, these options are required as an inevitable result of the need to deliver cache-coherency across all nodes in the cluster. In other cases, customers mistakenly use these parameters and the result is unnecessary performance damage.

Many customers temporarily remove these mount options during installation or patching of the application binaries. This removal can be performed safely if the user verifies that no other processes are actively using the target directory during the installation or patching process.

NFS transfer sizes with Oracle databases

By default, ONTAP limits NFS I/O sizes to 64K.

Random I/O with an most applications and databases uses a much smaller block size which is well below the 64K maximum. Large-block I/O is usually parallelized, so the 64K maximum is also not a limitation to obtaining maximum bandwidth.

There are some workloads where the 64K maximum does create a limitation. In particular, single-threaded operations such as backup or recovery operation or a database full table scan run faster and more efficiently if the database can perform fewer but larger I/Os. The optimum I/O handling size for ONTAP is 256K.

The maximum transfer size for a given ONTAP SVM can be changed as follows:

```
Cluster01::> set advanced
Warning: These advanced commands are potentially dangerous; use them only
when directed to do so by NetApp personnel.
Do you want to continue? {y|n}: y
Cluster01::*> nfs server modify -vserver vserver1 -tcp-max-xfer-size
262144
Cluster01::*>
```

Caution
Never decrease the maximum allowable transfer size on ONTAP below the value of rsize/wsize of currently mounted NFS file systems. This can create hangs or even data corruption with some operating systems. For example, if NFS clients are currently set at an rsize/wsize of 65536, then the ONTAP maximum transfer size could be adjusted between 65536 and 1048576 with no effect because the clients themselves are limited. Reducing the maximum transfer size below 65536 can damage availability or data.

Oracle databases and NVFAIL

NVFAIL is a feature within ONTAP that ensures the integrity during catastrophic failover scenarios.

Databases are vulnerable to corruption during storage failover events because they maintain large internal caches. If a catastrophic event requires forcing an ONTAP failover or forcing MetroCluster switchover, irrespective of the health of the overall configuration, the result is previously acknowledged changes may be effectively discarded. The contents of the storage array jump backward in time, and the state of the database cache no longer reflects the state of the data on disk. This inconsistency results in data corruption.

Caching can occur at the application or server layer. For example, an Oracle Real Application Cluster (RAC) configuration with servers active on both a primary and a remote site caches data within the Oracle SGA. A forced switchover operation that resulted in lost data would put the database at risk of corruption because the blocks stored in the SGA might not match the blocks on disk.

A less obvious use of caching is at the OS file system layer. Blocks from a mounted NFS file system might be cached in the OS. Alternatively, a clustered file system based on LUNs located on the primary site could be mounted on servers at the remote site, and once again data could be cached. A failure of NVRAM or a forced takeover or forced switchover in these situations could result in file system corruption.

ONTAP protects databases and operating systems from this scenario with NVFAIL and its associated settings.

ASM Reclamation Utility and ONTAP zero-block detection

ONTAP efficiently removes zeroed blocks written to a file or LUN when inline compression is enabled. Utilities such as the Oracle ASM Reclamation Utility (ASRU) work by writing zeros to unused ASM extents.

This allows DBAs to reclaim space on the storage array after data is deleted. ONTAP intercepts the zeros and deallocates the space from the LUN. The reclamation process is extremely fast because no data is being written within the storage system.

From a database perspective, the ASM diskgroup contains zeros, and reading those regions of the LUNs would result in a stream of zeros, but ONTAP does not store the zeros on drives. Instead, simple metadata changes are made that internally mark the zeroed regions of the LUN as empty of any data.

For similar reasons, performance testing involving zeroed data is not valid since blocks of zeros are not actually processed as writes within the storage array.

When using ASRU, ensure that all Oracle-recommended patches are installed.

Oracle database virtualization

Virtualization of databases with VMware, Oracle OLVM, or KVM is an increasingly common choice for NetApp customers who chose virtualization for even their most mission-critical databases.

Supportability

Many misconceptions exist about the Oracle support policies for virtualization, particularly for VMware products. It is not uncommon to hear that Oracle outright does not support virtualization. This notion is incorrect and leads to missed opportunities to benefit from virtualization. Oracle Doc ID 249212.1 discusses the actual requirements, and is rarely considered by customers to be a concern.

If a problem occurs on a virtualized server and that problem is previously unknown to Oracle Support, the customer might be asked to reproduce the problem on physical hardware. An Oracle customer running a

bleeding-edge version of a product might not want to use virtualization because of the potential for supportability problems, but this situation has not been a real-world for virtualization customers using generally available Oracle product versions.

Storage presentation

Customers considering virtualization of their databases should base their storage decisions on their business needs. Although this is a generally true statement for all IT decisions, it is especially important for database projects, because the size and scope of requirements vary considerably.

There are three basic options for storage presentation:

- Virtualized LUNs on hypervisor datastores
- iSCSI LUNs managed by the iSCSI initiator on the VM, not the hypervisor
- NFS file systems mounted by the VM (not from an NFS-based datastore)
- Direct device mappings. VMware RDMs are disfavored by customers, but physical devices are still often similarly directly mapped with KVM and OLVM virtualization.

Performance

The method of presenting storage to a virtualized guest does not generally affect performance. Host OSs, virtualized network drivers, and hypervisor datastore implementations are all highly optimized and can generally consume all available FC or IP network bandwidth between the hypervisor and the storage system as long as basic best practices are followed. In some cases, obtaining optimal performance might be slightly easier using one storage presentation approach as compared to another, but the end result should be comparable.

Manageability

The key factor in deciding how to present storage to a virtualized guest is mangeability. There is no right or wrong method. The best approach depends on IT operational needs, skills, and preferences.

Factors to consider include:

- **Transparency.** When a VM manages its file systems, it is easier for a database administrator or a system administrator to identify the source of the file systems for their data. The filesystems and LUNs are accessed no differently than with a physical server.
- **Consistency.** When a VM owns its file systems, the use or nonuse of a hypervisor layer affects manageability. The same procedures for provisioning, monitoring, data protection, and so on can be used across the entire estate, including both virtualized and nonvirtualized environments.

  On the other hand, in a otherwise 100% virtualized data center it may be preferable to also use datastore-based storage across the entire footprint on the same rationale mentioned above - consistency - the ability to use the same procedures for provisioning, protection, montoring, and data protection.

- **Stability and troubleshooting.** When a VM owns its file systems, delivering good, stable performance and troubleshooting problems are simpler because the entire storage stack is present on the VM. The hypervisor’s only role is to transport FC or IP frames. When a datastore is included in a configuration, it complicates the configuration by introducing another set of timeouts, parameters, log files, and potential bugs.
- **Portability.** When a VM owns its file systems, the process of moving an Oracle environment becomes much simpler. File systems can easily be moved between virtualized and nonvirtualized guests.

- **Vendor lock-in.** After data is placed in a datastore, using a different hypervisor or taking the data out of the virtualized environment entirely becomes difficult.
- **Snapshot enablement.** Traditional backup procedures in a virtualized environment can become a problem because of the relatively limited bandwidth. For example, a four-port 10GbE trunk might be sufficient to support the day-to-day performance needs of many virtualized databases, but such a trunk would be insufficient to perform backups using RMAN or other backup products that require streaming a full-sized copy of the data. The result is that an increasingly consolidated virtualized environment needs to perform backups via storage snapshots. This avoids the need to overbuild the hypervisor configuration purely to support the bandwidth and CPU requirements in the backup window.

  Using guest-owned file systems sometimes makes it easier to leverage snapshot-based backups and restores because the storage objects in need of protection can be targeted more easily. However, there are an increasingly large number of virtualization data protection products that integrate well with datastores and snapshots. The backup strategy should be fully considred before making a decision on how to present storage to a virtualized host.

Paravirtualized drivers

For optimum performance, the use of paravirtualized network drivers is critical. When a datastore is used, a paravirtualized SCSI driver is required. A paravirtualized device driver allows a guest to integrate more deeply into the hypervisor, as opposed to an emulated driver in which the hypervisor spends more CPU time mimicking the behavior of physical hardware.

Overcommitting RAM

Overcommitting RAM means configuring more virtualized RAM on various hosts than exists on the physical hardware. Doing so can cause unexpected performance problems. When virtualizing a database, the underlying blocks of the Oracle SGA must not be swapped out to storage by the hypervisor. Doing so causes highly unstable performance results.

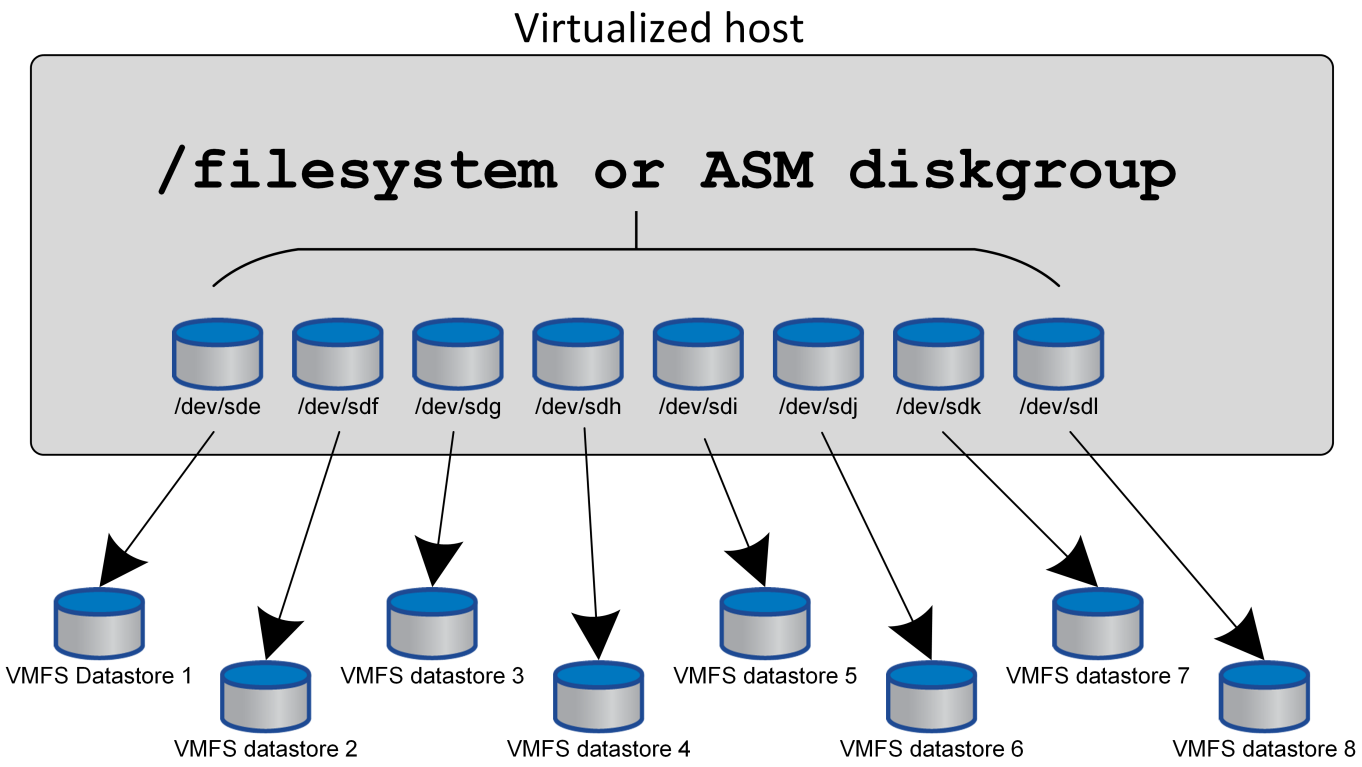
Datastore striping

When using databases with datastores, there is one critical factor to consider with respect to performance - striping.

Datastore technologies such as VMFS are able to span multiple LUNs, but they are not striped devices. The LUNs are concatenated. The end result can be LUN hot spots. For example, a typical Oracle database might have an 8-LUN ASM diskgroup. All 8 virtualized LUNs could be provisioned on an 8-LUN VMFS datastore, but there is no guarantee on which LUNs the data will reside. The resulting configuration could be all 8 virtualized LUN occupying a single LUN within the VMFS datastore. This becomes a performance bottleneck.

Striping is usually required. With some hypervisors, including KVM, it is possible to build a datastore using LVM striping as described here. With VMware, the architecture looks a little different. Each virtualized LUN needs to be placed on a different VMFS datastore.

For example:



The primary driver for this approach is not ONTAP, it's because of inherent limitation of the number of operations a single VM or hypervisor LUN can service in parallel. A single ONTAP LUN can generally support far more IOPS than a host can request. The single-LUN performance limit is almost universally a result of the host OS. The result is that most databases need between 4 and 8 LUNs to meet their performance needs.

VMware architectures need to plan their architectures carefully to ensure that datastore and/or LUN path maximumes are not encountered with this approach. Addtionally, there is no requirement for a unique set of VMFS datastores for every database. The primary need is to ensure each host has a clean set of 4-8 IO paths from the virtualized LUNs to the backend LUNs on the storage system itself. In rare occasions, even more datatores may be beneficial for truly extreme performance demands, but 4-8 LUNs is generally sufficient for 95% of all databases. A single ONTAP volume containing 8 LUNs can support up to 250,000 random Oracle block IOPS with a typical OS/ONTAP/network configuration.

Tiering

Oracle database FabricPool tiering overview

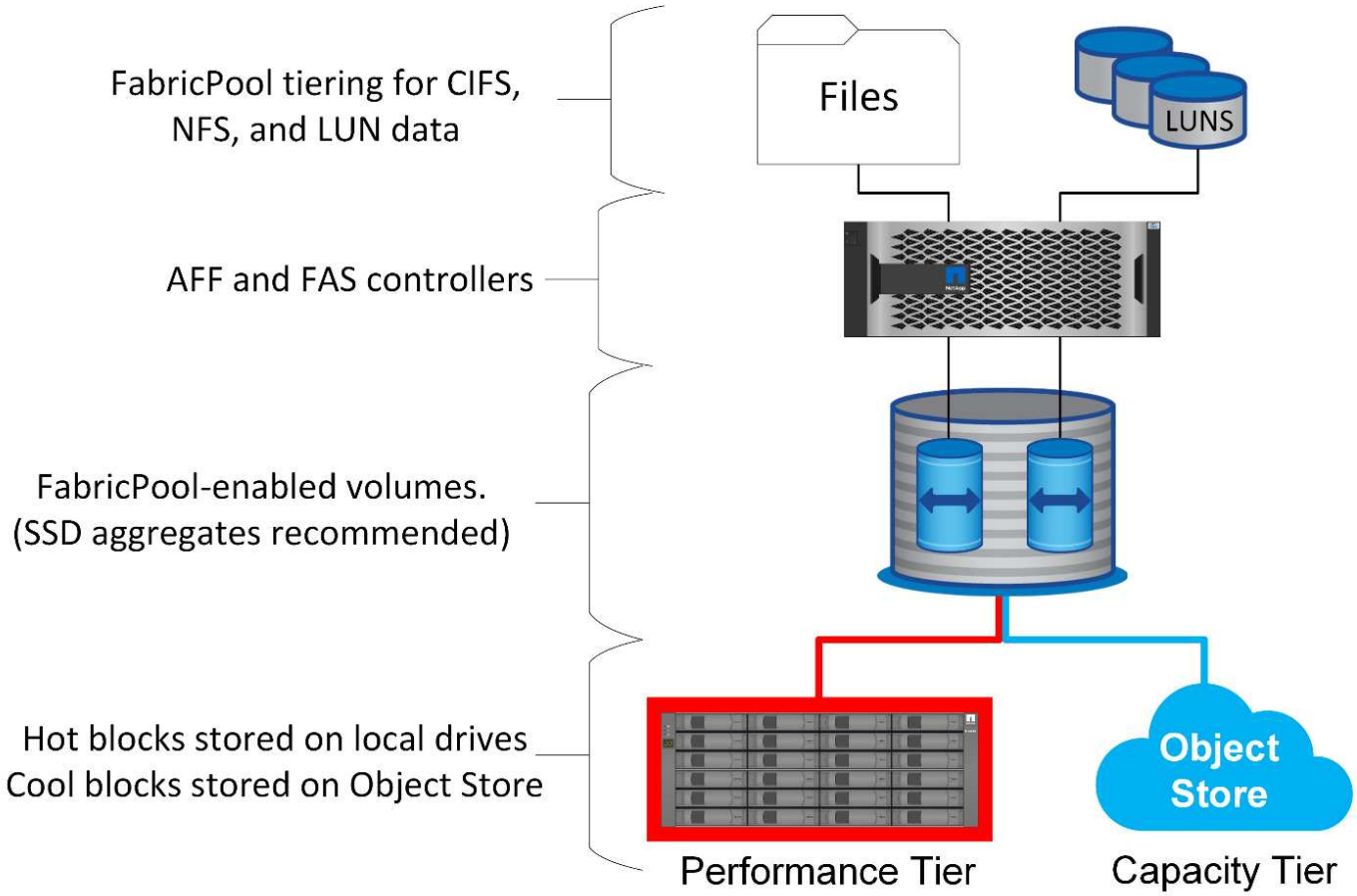
Understanding how FabricPool tiering affects Oracle and other databases requires an understanding of low-level FabricPool architecture.

Architecture

FabricPool is a tiering technology that classifies blocks as hot or cool and places them in the most appropriate tier of storage. The performance tier is most often located on SSD storage and hosts the hot data blocks. The capacity tier is located on an object store and hosts the cool data blocks. Object storage support includes NetApp StorageGRID, ONTAP S3, Microsoft Azure Blob storage, Alibaba Cloud Object Storage service, IBM Cloud Object Storage, Google Cloud storage, and Amazon AWS S3.

Multiple tiering policies are available that control how blocks are classified as hot or cool, and policies can be

set on a per-volume basis and changed as required. Only the data blocks are moved between the performance and capacity tiers. The metadata that defines the LUN and file system structure always remains on the performance tier. As a result, management is centralized on ONTAP. Files and LUNs appear no different from data stored on any other ONTAP configuration. The NetApp AFF or FAS controller applies the defined policies to move data to the appropriate tier.



Object store providers

Object storage protocols use simple HTTP or HTTPS requests for storing large numbers of data objects. Access to the object storage must be reliable, because data access from ONTAP depends on prompt servicing of requests. Options include the Amazon S3 Standard and Infrequent Access options, and Microsoft Azure Hot and Cool Blob Storage, IBM Cloud, and Google Cloud. Archival options such as Amazon Glacier and Amazon Archive are not supported because the time required to retrieve data can exceed the tolerances of host operating systems and applications.

NetApp StorageGRID is also supported and is an optimal enterprise-class solution. It is a high-performance, scalable, and highly secure object storage system that can provide geographic redundancy for FabricPool data as well as other object store applications that are increasingly likely to be part of enterprise application environments.

StorageGRID can also reduce costs by avoiding the egress charges imposed by many public cloud providers for reading data back from their services.

Data and metadata

Note that the term "data" here applies to the actual data blocks, not the metadata. Only data blocks are tiered, while metadata remains in the performance tier. In addition, the status of a block as hot or cool is only affected

by reading the actual data block. Simply reading the name, timestamp, or ownership metadata of a file does not affect the location of the underlying data blocks.

Backups

Although FabricPool can significantly reduce storage footprints, it is not by itself a backup solution. NetApp WAFL metadata always stays on the performance tier. If a catastrophic disaster destroys the performance tier, a new environment cannot be created using the data on the capacity tier because it contains no WAFL metadata.

FabricPool can, however, become part of a backup strategy. For example, FabricPool can be configured with NetApp SnapMirror replication technology. Each half of the mirror can have its own connection to an object storage target. The result is two independent copies of the data. The primary copy consists of the blocks on the performance tier and associated blocks in the capacity tier, and the replica is a second set of performance and capacity blocks.

Tiering policies

Oracle database FabricPool tiering policies

Four policies are available in ONTAP which control how Oracle data on the performance tier become a candidate to be relocated to the capacity tier.

Snapshot-only

The `snapshot-only` `tiering-policy` applies only to blocks that are not shared with the active file system. It essentially results in tiering of database backups. Blocks become candidates for tiering after a snapshot is created and the block is then overwritten, resulting in a block that exists only within the snapshot. The delay before a `snapshot-only` block is considered cool is controlled by the `tiering-minimum-cooling-days` setting for the volume. The range as of ONTAP 9.8 is from 2 to 183 days.

Many datasets have low change rates, resulting in minimal savings from this policy. For example, a typical database observed on ONTAP has a change rate of less than 5% per week. Database archive logs can occupy extensive space, but they usually continue to exist in the active file system and thus would not be candidates for tiering under this policy.

Auto

The `auto` tiering policy extends tiering to both snapshot-specific blocks as well as blocks within the active file system. The delay before a block is considered cool is controlled by the `tiering-minimum-cooling-days` setting for the volume. The range as of ONTAP 9.8 is from 2 to 183 days.

This approach enables tiering options that are not available with the `snapshot-only` policy. For example, a data protection policy might require 90 days of certain log files to be retained. Setting a cooling period of 3 days results in any log files older than 3 days to be tiered out from the performance layer. This action frees up substantial space on the performance tier while still allowing you to view and manage the full 90 days of data..

None

The `none` tiering policy prevents any additional blocks from being tiered from the storage layer, but any data still in the capacity tier remains in the capacity tier until it is read. If the block is then read, it is pulled back and placed on the performance tier.

The primary reason to use the `none` tiering policy is to prevent blocks from being tiered, but it could become

useful to change the policies over time. For example, let's say that a specific dataset is extensively tiered to the capacity layer, but an unexpected need for full performance capabilities arises. The policy can be changed to prevent any additional tiering and to confirm that any blocks read back as IO increases remain in the performance tier.

All

The `all` tiering policy replaces the `backup` policy as of ONTAP 9.6. The `backup` policy applied only to data protection volumes, meaning a SnapMirror or NetApp SnapVault destination. The `all` policy functions the same, but is not restricted to data protection volumes.

With this policy, blocks are immediately considered cool and eligible to be tiered to the capacity layer immediately.

This policy is especially appropriate for long-term backups. It can also be used as a form of Hierarchical Storage Management (HSM). In the past, HSM was commonly used to tier the data blocks of a file to tape while keeping the file itself visible on the file system. A FabricPool volume with the `all` policy allows you to store files in a visible and manageable yet consuming nearly no space on the local storage tier.

Oracle databases and FabricPool retrieval policies

The tiering policies control which Oracle database blocks are tiered from the performance tier to the capacity tier. Retrieval policies control what happens when a block that has been tiered is read.

Default

All FabricPool volumes are initially set at `default`, which means the behavior is controlled by the `cloud-retrieval-policy. `The exact behavior depends on the tiering policy used.

- `auto`– only retrieve randomly read data
- `snapshot-only`– retrieve all sequentially or randomly read data
- `none`– retrieve all sequentially or randomly read data
- `all`– do not retrieve data from the capacity tier

On-read

Setting `cloud-retrieval-policy` to on-read overrides the default behavior so a read of any tiered data results in that data being returned to the performance tier.

For example, a volume might have been lightly used for a long time under the `auto` tiering policy and most of the blocks are now tiered out.

If an unexpected change in business needs required some of the data to be repeatedly scanned in order to prepare a certain report, it may be desirable to change the `cloud-retrieval-policy` to `on-read` to ensure that all data that is read is returned to the performance tier, including both sequentially and randomly read data. This would improve performance of sequential I/O against the volume.

Promote

The behavior of the promote policy depends on the tiering policy. If the tiering policy is `auto`, then setting the `cloud-retrieval-policy` `to` `promote` brings back all blocks from the capacity tier on the next tiering

scan.

If the tiering policy is `snapshot-only`, then the only blocks that are returned are the blocks that are associated with the active file system. Normally this would not have any effect because the only blocks tiered under the `snapshot-only` policy would be blocks associated exclusively with snapshots. There would be no tiered blocks in the active file system.

If, however, data on a volume was restored by a volume SnapRestore or file-clone operation from a snapshot, some of the blocks that were tiered out because they were only associated with snapshots may now be required by the active file system. It may be desirable to temporarily change the `cloud-retrieval-policy` policy to `promote` to quickly retrieve all locally required blocks.

Never

Do not retrieve blocks from the capacity tier.

Tiering strategies

Oracle database full file FabricPool tiering

Although FabricPool tiering operates at the block level, in some cases it can be used to provide file-level tiering.

Many applications datasets are organized by date, and such data is generally increasingly less likely to be accessed as it ages. For example, a bank might have a repository of PDF files that contain five years of customer statements, but only the most recent few months are active. FabricPool can be used to relocate older datafiles to the capacity tier. A cooling period of 14 days would ensure the more recent 14 days of PDF files remain on the performance tier. Furthermore, files that are read at least every 14 days would remain hot and therefore remain on the performance tier.

Policies

To implement a file-based tiering approach, you must have files that are written and not subsequently modified. The `tiering-minimum-cooling-days` policy should be set high enough so that files that you might need remain on the performance tier. For example, a dataset for which the most recent 60 days of data is required with optimal performance warrants setting the `tiering-minimum-cooling-days` period to 60. Similar results can also be achieved based on the file access patterns. For example, if the most recent 90 days of data is required and the application is accessing that 90-day span of data, then the data would remain on the performance tier. By setting the `tiering-minimum-cooling-days` period to 2, you get prompt tiering after the data becomes less active.

The `auto` policy is required to drive tiering of these blocks because only the `auto` policy affects blocks that are in the active file system.

> Any type of access to data resets the heat map data. Virus scanning, indexing, and even backup activity that reads the source files prevents tiering because the required `tiering-minimum-cooling-days` threshold is never reached.

Oracle partial file FabricPool tiering

Because FabricPool works at the block level, files that are subject to change can be partially tiered to object storage while also remaining partially on performance tier.

This is common with databases. Databases that are known to contain inactive blocks are also candidates for FabricPool tiering. For example, a supply chain management database might contain historical information that must be available if needed but is not accessed during normal operations. FabricPool can be used to selectively relocate the inactive blocks.

For example, datafiles running on a FabricPool volume with a `tiering-minimum-cooling-days` period of 90 days retains any blocks accessed in the preceding 90 days on the performance tier. However, anything that is not accessed for 90 days is relocated to the capacity tier. In other cases, normal application activity preserves the correct blocks on the correct tier. For example, if a database is normally used to process the previous 60 days of data on a regular basis, a much lower `tiering-minimum-cooling-days` period can be set because the natural activity of the application makes sure that blocks are not relocated prematurely.

The `auto` policy should be used with care with databases. Many databases have periodic activities such as end-of-quarter process or reindexing operations. If the period of these operations is greater than the `tiering-minimum-cooling-days` performance problems can occur. For example, if end-of-quarter processing requires 1TB of data that was otherwise untouched, that data might now be present on the capacity tier. Reads from the capacity tier is often extremely fast and may not cause performance problems, but the exact results will depend on the object store configuration.

Policies

The `tiering-minimum-cooling-days` policy should be set high enough to retain files that might be required on the performance tier. For example, a database in which the most recent 60 days of data might be required with optimal performance would warrant setting the `tiering-minimum-cooling-days` period to 60 days. Similar results could also be achieved based on the access patterns of files. For example, if the most recent 90 days of data is required and the application is accessing that 90- day span of data, then the data would remain on the performance tier. Setting the `tiering-minimum-cooling-days` period to 2 days would tier the data promptly after the data becomes less active.

The `auto` policy is required to drive tiering of these blocks because only the `auto` policy affects blocks that are in the active file system.

> Any type of access to data resets the heat map data. Therefore, database full table scans and even backup activity that reads the source files prevents tiering because the required `tiering-minimum-cooling-days` threshold is never reached.

Oracle database archive log tiering

Perhaps the most important use for FabricPool is improving the efficiency of known cold data, such as database transaction logs.

Most relational databases operate in transaction log archival mode to deliver point-in-time recovery. Changes to the databases are committed by recording the changes in the transaction logs, and the transaction log is retained without being overwritten. The result can be a requirement to retain an enormous volume of archived transaction logs. Similar examples exist with many other application workflows that generate data that must be retained, but is highly unlikely to ever be accessed.

FabricPool solves these problems by delivering a single solution with integrated tiering. Files are stored and remain accessible in their usual location, but take up virtually no space on the primary array.

Policies

Use a `tiering-minimum-cooling-days` policy of a few days results in retention of blocks in the recently

created files (which are the files most likely to be required in the near term) on the performance tier. The data blocks from older files are then moved to the capacity tier.

The `auto` enforces prompt tiering when the cooling threshold has been reached irrespective of whether the logs have been deleted or continue to exist in the primary file system. Storing all the potentially required logs in a single location in the active file system also simplifies management. There is no reason to search through snapshots to locate a file that needs to be restored.

Some applications, such as Microsoft SQL Server, truncate transaction log files during backup operations so that the logs are no longer in the active file system. Capacity might be saved by using the `snapshot-only` tiering policy, but the `auto` policy is not useful for log data because there should rarely cooled log data in the active file system.

Oracle with FabricPool snapshot tiering

The initial release of FabricPool targeted the backup use case. The only type of blocks that could be tiered were blocks that were no longer associated with data in the active file system. Therefore, only the snapshot data blocks could be moved to the capacity tier. This remains one of the safest tiering options when you need to ensure performance is never affected.

Policies - local snapshots

Two options exist for tiering inactive snapshot blocks to the capacity tier. First, the `snapshot-only` policy only targets the snapshot blocks. Although the `auto` policy includes the `snapshot-only` blocks, it also tiers blocks from the active file system. This might not be desirable.

The `tiering-minimum-cooling-days` value should be set to a time period that makes data that might be required during a restoration available on the performance tier. For example, most restore scenarios of a critical production database include a restore point at some time in the previous few days. Setting a `tiering-minimum-cooling-days` value of 3 would make sure that any restoration of the file results in a file that immediately delivers maximum performance. All blocks in the active files are still present on fast storage without needing to recover them from the capacity tier.

Policies - replicated snapshots

A snapshot that is replicated with SnapMirror or SnapVault that is only used for recovery should generally use the FabricPool `all` policy. With this policy, metadata is replicated, but all data blocks are immediately sent to the capacity tier, which yields maximum performance. Most recovery processes involve sequential I/O, which is inherently efficient. The recovery time from the object store destination should be evaluated, but, in a well-designed architecture, this recovery process does not need to be significantly slower than recovery from local data.

If the replicated data is also intended to be used for cloning, the `auto` policy is more appropriate, with a `tiering-minimum-cooling-days` value that encompasses data that is expected to be regularly used in a cloning environment. For example, a database's active working set might include data read or written in the previous three days, but it could also include another 6 months of historical data. If so, then the `auto` policy at the SnapMirror destination makes the working set available on the performance tier.

Oracle database backup tiering

Traditional application backups include products such as Oracle Recovery Manager, which create file-based backups outside the location of the original database.

```
`tiering-minimum-cooling-days` policy of a few days preserves the most
recent backups, and therefore the backups most likely to be required for
an urgent recovery situation, on the performance tier. The data blocks of
the older files are then moved to the capacity tier.
```

The `auto` policy is the most appropriate policy for backup data. This ensures prompt tiering when the cooling threshold has been reached irrespective of whether the files have been deleted or continue to exist in the primary file system. Storing all the potentially required files in a single location in the active file system also simplifies management. There is no reason to search through snapshots to locate a file that needs to be restored.

The `snapshot-only` policy could be made to work, but that policy only applies to blocks that are no longer in the active file system. Therefore, files on an NFS or SMB share must be deleted first before the data can be tiered.

This policy would be even less efficient with a LUN configuration because deletion of a file from a LUN only removes the file references from the file system metadata. The actual blocks on the LUNs remain in place until they are overwritten. This situation can create a lengthy delay between the time a file is deleted and the time that the blocks are overwritten and become candidates for tiering. There is some benefit to moving the `snapshot-only` blocks to the capacity tier, but, overall, FabricPool management of backup data works best with the `auto` policy.

> This approach helps users manage the space required for backups more efficiently, but FabricPool itself is not a backup technology. Tiering backup files to object store simplifies management because the files are still visible on the original storage system, but the data blocks in the object store destination depend on the original storage system. If the source volume is lost, the object store data is no longer useable.

Oracle database and object store access interruptions

Tiering a dataset with FabricPool results in a dependency between the primary storage array and the object store tier. There are many object storage options that offer varying levels of availability. It is important to understand the impact of a possible loss of connectivity between the primary storage array and the object storage tier.

If an I/O issued to ONTAP requires data from the capacity tier and ONTAP cannot reach the capacity tier to retrieve blocks, then the I/O eventually times out. The effect of this timeout depends on the protocol used. In an NFS environment, ONTAP responds with either an EJUKEBOX or EDELAY response, depending on the protocol. Some older operating systems might interpret this as an error, but current operating systems and current patch levels of the Oracle Direct NFS client treat this as a retriable error and continue waiting for the I/O to complete.

A shorter timeout applies to SAN environments. If a block in the object store environment is required and remains unreachable for two minutes, a read error is returned to the host. The ONTAP volume and LUNs remain online, but the host OS might flag the file system as being in an error state.

Object storage connectivity problems `snapshot-only` policy is less of a concern because only backup data is tiered. Communication problems would slow data recovery but would not otherwise affect data being actively used. The `auto` and `all` policies allow tiering of cold data from the active LUN, which means that an error during object store data retrieval could affect database availability. A SAN deployment with these policies

should only be used with enterprise-class object storage and network connections designed for high availability. NetApp StorageGRID is the superior option.

Oracle data protection

Oracle data protection with ONTAP

NetApp knows the most mission-critical data is found in databases.

An enterprise cannot operate without access to its data, and sometimes, the data defines the business. This data must be protected; however, data protection is more than just ensuring a usable backup—it is about performing the backups quickly and reliably in addition to storing them safely.

The other side of data protection is data recovery. When data is inaccessible, the enterprise is affected and might be inoperative until data is restored. This process must be fast and reliable. Finally, most databases must be protected against disasters, which means maintaining a replica of the database. The replica must be sufficiently up to date. It must also be quick and simple to make the replica a fully operational database.

> This documentation replaces previously published technical report *TR-4591: Oracle data protection: Backup, recovery, and replication.*

Planning

The right enterprise data protection architecture depends on the business requirements surrounding data retention, recoverability, and tolerance for disruption during various events.

For example, consider the number of applications, databases, and important datasets in scope. Building a backup strategy for a single dataset that ensures compliance with typical SLAs is fairly straightforward because there are not many objects to manage. As the number of datasets increases, monitoring becomes more complicated and administrators might be forced to spend an increasing amount of time addressing backup failures. As an environment reaches cloud and service provider scales, a wholly different approach is needed.

Dataset size also affects strategy. For example, many options exist for backup and recovery with a 100GB database because the data set is so small. Simply copying the data from backup media with traditional tools typically delivers a sufficient RTO for recovery. A 100TB database normally needs a completely different strategy unless the RTO allows for a multiday outage, in which case a traditional copy-based backup and recovery procedure might be acceptable.

Finally, there are factors outside the backup and recovery process itself. For example, are there databases supporting critical production activities, making recovery a rare event that is only performed by skilled DBAs? Alternatively, are the databases part of a large development environment in which recovery is a frequent occurrence and managed by a generalist IT team?

Oracle database RTO, RPO, and SLA planning

ONTAP allows you to easily tailor an Oracle database data protection strategy to your business requirements.

These requirements include factors such as the speed of recovery, the maximum permissible data loss, and backup retention needs. The data protection plan must also take into consideration various regulatory requirements for data retention and restoration. Finally, different data recovery scenarios must be considered, ranging from the typical and foreseeable recovery resulting from user or application errors up to disaster recovery scenarios that include the complete loss of a site.

Small changes in data protection and recovery policies can have a significant effect on the overall architecture of storage, backup, and recovery. It is critical to define and document standards before starting design work to avoid complicating a data protection architecture. Unnecessary features or levels of protection lead to unnecessary costs and management overhead, and an initially overlooked requirement can lead a project in the wrong direction or require last-minute design changes.

Recovery time objective

The recovery time objective (RTO) defines the maximum time allowed for the recovery of a service. For example, a human resources database might have an RTO of 24 hours because, although it would be very inconvenient to lose access to this data during the workday, the business can still operate. In contrast, a database supporting the general ledger of a bank would have an RTO measured in minutes or even seconds. An RTO of zero is not possible, because there must be a way to differentiate between an actual service outage and a routine event such as a lost network packet. However, a near-zero RTO is a typical requirement.

Recovery point objective

The recovery point objective (RPO) defines the maximum tolerable data loss. In many cases, the RPO is solely determined by the frequency of snapshots or snapmirror updates.

In some cases, the RPO can be made more aggressive be selectively protecting certain data more frequently. In a database context, the RPO is usually a question of how much log data can be lost in a specific situation. In a typical recovery scenario in which a database is damaged due to a product bug or user error, the RPO should be zero, meaning there should be no data loss. The recovery procedure involves restoring an earlier copy of the database files and then replaying the log files to bring the database state up to the desired point in time. The log files required for this operation should already be in place in the original location.

In unusual scenarios, log data might be lost. For example, an accidental or malicious `rm -rf *` of database files could result in the deletion of all data. The only option would be to restore from backup, including log files, and some data would inevitably be lost. The only option to improve the RPO in a traditional backup environment would be to perform repeated backups of the log data. This has limitations, however, because of the constant data movement and the difficulty maintaining a backup system as a constantly running service. One of the benefits of advanced storage systems is the ability to protect data from accidental or malicious damage to files and thus deliver a better RPO without data movement.

Disaster recovery

Disaster recovery includes the IT architecture, policies, and procedures required to recover a service in the event of a physical disaster. This can include floods, fire, or person acting with malicious or negligent intent.

Disaster recovery is more than just a set of recovery procedures. It is the complete process of identifying the various risks, defining the data recovery and service continuity requirements, and delivering the right architecture with associated procedures.

When establishing data protection requirements, it is critical to differentiate between typical RPO and RTO requirements and the RPO and RTO requirements needed for disaster recovery. Some applications environments require an RPO of zero and a near-zero RTO for data loss situations ranging from a relatively normal user error right up to a fire that destroys a data center. However, there are cost and administrative consequences for these high levels of protection.

In general, nondisaster data recovery requirements should be strict for two reasons. First, application bugs and user errors that damage data are foreseeable to the point they are almost inevitable. Second, it is not difficult to design a backup strategy that can deliver an RPO of zero and a low RTO as long as the storage system is not destroyed. There is no reason not to address a significant risk that is easily remedied, which is why the RPO and RTO targets for local recovery should be aggressive.

Disaster recovery RTO and RPO requirements vary more widely based on the likelihood of a disaster and the consequences of the associated data loss or disruption to a business. RPO and RTO requirements should be based on the actual business needs and not on general principles. They must account for multiple logical and physical disaster scenarios.

Logical disasters

Logical disasters include data corruption caused by users, application or OS bugs, and software malfunctions. Logical disasters can also include malicious attacks by outside parties with viruses or worms or by exploiting application vulnerabilities. In these cases, the physical infrastructure is undamaged but the underlying data is no longer valid.

An increasingly common type of logical disaster is known as ransomware, in which an attack vector is used to encrypt data. Encryption does not damage the data, but it makes it unavailable until payment is made to a third party. An increasing number of enterprises are being specifically targeted with ransomware hacks. For this threat, NetApp offers tamperproof snapshots where not even the storage administrator can change protected data before the configured expiry date.

Physical disasters

Physical disasters include the failure of components of an infrastructure that exceeds its redundancy capabilities and result in a loss of data or an extended loss of service. For example, RAID protection provides disk-drive redundancy, and the use of HBAs provides FC port and FC cable redundancy. Hardware failures of such components is foreseeable and does not impact availability.

In an enterprise environment, it is generally possible to protect the infrastructure of an entire site with redundant components to the point where the only foreseeable physical disaster scenario is complete loss of the site. Disaster recovery planning then depends on site-to-site replication.

Synchronous and asynchronous data protection

In an ideal world, all data would be synchronously replicated across geographically dispersed sites. Such replication is not always feasible or even possible for several reasons:

- Synchronous replication unavoidably increases write latency because all changes must be replicated to both locations before the application/database can proceed with processing. The resulting performance effect is sometimes unacceptable, ruling out the use of synchronous mirroring.
- The increased adoption of 100% SSD storage means that additional write latency is more likely to be noticed because performance expectations include hundreds of thousands of IOPS and submillisecond latency. Gaining the full benefits of using 100% SSDs can require revisiting the disaster recovery strategy.
- Datasets continue to grow in terms of bytes, creating challenges with ensuring sufficient bandwidth to sustain synchronous replication.
- Datasets also grow in terms of complexity, creating challenges with the management of large-scale synchronous replication.
- Cloud-based strategies frequently involve greater replication distances and latency, further precluding the use of synchronous mirroring.

NetApp offers solutions that include both synchronous replication for the most exacting data recovery demands and asynchronous solutions that allow for better performance and flexibility. In addition, NetApp technology integrates seamlessly with many third-party replication solutions, such as Oracle DataGuard

Retention Time

The final aspect of a data protection strategy is the data retention time, which can vary dramatically.

- A typical requirement is 14 days of nightly backups on the primary site and 90 days of backups stored on a secondary site.
- Many customers create standalone quarterly archives stored on different media.
- A constantly updated database might have no need for historical data, and backups need only be retained for a few days.
- Regulatory requirements might require recoverability to the point of any arbitrary transaction in a 365-day window.

Oracle database availability with ONTAP

ONTAP is designed to deliver maximum Oracle database availability. A complete description of ONTAP high availability features is beyond the scope of this document. However, as with data protection, a basic understanding of this functionality is important when designing a database infrastructure.

HA pairs

The basic unit of high availability is the HA pair. Each pair contains redundant links to support replication of data to NVRAM. NVRAM is not a write cache. The RAM inside the controller serves as the write cache. The purpose of NVRAM is to temporarily journal data as a safeguard against unexpected system failure. In this respect, it is similar to a database redo log.

Both NVRAM and a database redo log are used to store data quickly, allowing changes to data to be committed as quickly as possible. The update to the persistent data on drives (or datafiles) does not take place until later during a process called a checkpoint on both ONTAP and most databases platforms. Neither NVRAM data nor database redo logs are read during normal operations.

If a controller fails abruptly, there are likely to be pending changes stored in NVRAM that have not yet been written to the drives. The partner controller detects the failure, take control of the drives, and applies the required changes that have been stored in NVRAM.

Takeover and giveback

Takeover and giveback refers to the process of transferring responsibility for storage resources between nodes in an HA pair. There are two aspects to takeover and giveback:

- Management of the network connectivity that allows access to the drives
- Management of the drives themselves

Network interfaces supporting CIFS and NFS traffic are configured with both a home and failover location. A takeover includes moving the network interfaces to their temporary home on a physical interface located on the same subnet(s) as the original location. A giveback includes moving the network interfaces back to their original locations. The exact behavior can be tuned as required.

Network interfaces supporting SAN block protocols such as iSCSI and FC are not relocated during takeover and giveback. Instead, LUNs should be provisioned with paths that includes a complete HA pair which results in a primary path and a secondary path.

Additional paths to additional controllers can also be configured to support relocating data between nodes in a larger cluster, but this is not part of the HA process.

The second aspect of takeover and giveback is the transfer of disk ownership. The exact process depends on multiple factors including the reason for the takeover/giveback and the command line options issued. The goal is to perform the operation as efficiently as possible. Although the overall process might appear to require several minutes, the actual moment in which ownership of the drive is transitioned from node to node can generally be measured in seconds.

Takeover time

Host I/O experiences a short pause in I/O during takeover and giveback operations, but there should not be application disruption in a correctly configured environment. The actual transition process in which I/O is delayed is generally measured in seconds, but the host might require additional time to recognize the change in data paths and resubmit I/O operations.

The nature of the disruption depends on the protocol:

- A network interface supporting NFS and CIFS traffic issues an Address Resolution Protocol (ARP) request to the network after the transition to a new physical location. This causes the network switches to update their media access control (MAC) address tables and resume processing I/O. Disruption in the case of planned takeover and giveback is usually measured in seconds and in many cases is not detectable. Some networks might be slower to fully recognize the change in network path, and some OSs might queue up a lot of I/O in a very short time that must be retried. This can extend the time required to resume I/O.
- A network interface supporting SAN protocols does not transition to a new location. A host OS must change the path or paths in use. The pause in I/O observed by the host depends on multiple factors. From a storage system point of view, the period where I/O cannot be served is just a few seconds. However, different host OSs might require additional time to allow an I/O to time out before retry. Newer OSs are better able to recognize a path change much more quickly, but older OSs typically require up to 30 seconds to recognize a change.

The expected takeover times during which the storage system cannot serve data to an application environment are shown in the table below. There should not be any errors in any application environment, the takeover should instead appear as a short pause in IO processing.

	NFS	AFF	ASA
Planned takeover	15 sec	6-10 sec	2-3 sec
Unplanned takeover	30 sec	6-10 sec	2-3 sec

Checksums and Oracle database integrity

ONTAP and its supported protocols include multiple features that protect Oracle database integrity, including both data at rest and data being transmitted over the network network.

Logical data protection within ONTAP consists of three key requirements:

- Data must be protected against data corruption.
- Data must be protected against drive failure.
- Changes to data must be protected against loss.

These three needs are discussed in the following sections.

Network corruption: checksums

The most basic level of data protection is the checksum, which is a special error-detecting code stored alongside the data. Corruption of data during network transmission is detected with the use of a checksum and, in some instances, multiple checksums.

For example, an FC frame includes a form of checksum called a cyclic redundancy check (CRC) to make sure that the payload is not corrupted in transit. The transmitter sends both the data and the CRC of the data. The receiver of an FC frame recalculates the CRC of the received data to make sure that it matches the transmitted CRC. If the newly computed CRC does not match the CRC attached to the frame, the data is corrupt and the FC frame is discarded or rejected. An iSCSI I/O operation includes checksums at the TCP/IP and Ethernet layers, and, for extra protection, it can also include optional CRC protection at the SCSI layer. Any bit corruption on the wire is detected by the TCP layer or IP layer, which results in retransmission of the packet. As with FC, errors in the SCSI CRC result in a discard or rejection of the operation.

Drive corruption: checksums

Checksums are also used to verify the integrity of data stored on drives. Data blocks written to drives are stored with a checksum function that yields an unpredictable number that is tied to the original data. When data is read from the drive, the checksum is recomputed and compared to the stored checksum. If it does not match, then the data has become corrupt and must be recovered by the RAID layer.

Data corruption: lost writes

One of the most difficult types of corruption to detect is a lost or a misplaced write. When a write is acknowledged, it must be written to the media in the correct location. In-place data corruption is relatively easy to detect by using a simple checksum stored with the data. However, if the write is simply lost, then the prior version of data might still exist and the checksum would be correct. If the write is placed at the wrong physical location, the associated checksum would once again be valid for the stored data, even though the write has destroyed other data.

The solution to this challenge is as follows:

- A write operation must include metadata that indicates the location where the write is expected to be found.
- A write operation must include some sort of version identifier.

When ONTAP writes a block, it includes data on where the block belongs. If a subsequent read identifies a block, but the metadata indicates that it belongs at location 123 when it was found at location 456, then the write has been misplaced.

Detecting a wholly lost write is more difficult. The explanation is very complicated, but essentially ONTAP is storing metadata in a way that a write operation results in updates to two different locations on the drives. If a write is lost, a subsequent read of the data and associated metadata shows two different version identities. This indicates that the write was not completed by the drive.

Lost and misplaced write corruption is exceedingly rare, but, as drives continue to grow and datasets push into exabyte scale, the risk increases. Lost write detection should be included in any storage system supporting database workloads.

Drive failures: RAID, RAID DP, and RAID-TEC

If a block of data on a drive is discovered to be corrupt, or the entire drive fails and is wholly unavailable, the

data must be reconstituted. This is done in ONTAP by using parity drives. Data is striped across multiple data drives, and then parity data is generated. This is stored separately from the original data.

ONTAP originally used RAID 4, which uses a single parity drive for each group of data drives. The result was that any one drive in the group could fail without resulting in data loss. If the parity drive failed, no data was damaged and a new parity drive could be constructed. If a single data drive failed, the remaining drives could be used with the parity drive to regenerate the missing data.

When drives were small, the statistical chance of two drives failing simultaneously was negligible. As drive capacities have grown, so has the time required to reconstruct data after a drive failure. This has increased the window in which a second drive failure would result in data loss. In addition, the rebuild process creates a lot of additional I/O on the surviving drives. As drives age, the risk of the additional load leading to a second drive failure also increases. Finally, even if the risk of data loss did not increase with the continued use of RAID 4, the consequences of data loss would become more severe. The more data that would be lost in the event of a RAID-group failure, the longer it would take to recover the data, extending business disruption.

These issues led NetApp to develop the NetApp RAID DP technology, a variant of RAID 6. This solution includes two parity drives, meaning that any two drives in a RAID group can fail without creating data loss. Drives have continued to grow in size, which eventually led NetApp to develop the NetApp RAID-TEC technology, which introduces a third parity drive.

Some historical database best practices recommend the use of RAID-10, also known as striped mirroring. This offers less data protection than even RAID DP because there are multiple two-disk failure scenarios, whereas in RAID DP there are none.

There are also some historical database best practices that indicate RAID-10 is preferred to RAID-4/5/6 options due to performance concerns. These recommendations sometimes refer to a RAID penalty. Although these recommendations are generally correct, they are inapplicable to the implementations of RAID within ONTAP. The performance concern is related to parity regeneration. With traditional RAID implementations, processing the routine random writes performed by a database requires multiple disk reads to regenerate the parity data and complete the write. The penalty is defined as the additional read IOPS required to perform write operations.

ONTAP does not incur a RAID penalty because writes are staged in memory where parity is generated and then written to disk as a single RAID stripe. No reads are required to complete the write operation.

In summary, when compared to RAID 10, RAID DP and RAID-TEC deliver much more usable capacity, better protection against drive failure, and no performance sacrifice.

Hardware failure protection: NVRAM

Any storage array servicing a database workload must service write operations as quickly as possible. Furthermore, a write operation must be protected from loss from an unexpected event such as a power failure. This means any write operation must be safely stored in at least two locations.

AFF and FAS systems rely on NVRAM to meet these requirements. The write process works as follows:

1. The inbound write data is stored in RAM.
2. The changes that must be made to data on disk are journaled into NVRAM on both the local and partner node. NVRAM is not a write cache; rather it is a journal similar to a database redo log. Under normal conditions, it is not read. It is only used for recovery, such as after a power failure during I/O processing.
3. The write is then acknowledged to the host.

The write process at this stage is complete from the application point of view, and the data is protected against

loss because it is stored in two different locations. Eventually, the changes are written to disk, but this process is out-of-band from the application point of view because it occurs after the write is acknowledged and therefore does not affect latency. This process is once again similar to database logging. A change to the database is recorded in the redo logs as quickly as possible, and the change is then acknowledged as committed. The updates to the datafiles occur much later and do not directly affect the speed of processing.

In the event of a controller failure, the partner controller takes ownership of the required disks and replays the logged data in NVRAM to recover any I/O operations that were in-flight when the failure occurred.

Hardware failure protection: NVFAIL

As discussed earlier, a write is not acknowledged until it has been logged into local NVRAM and NVRAM on at least one other controller. This approach makes sure that a hardware failure or power outage does not result in the loss of in-flight I/O. If the local NVRAM fails or the connectivity to HA partner fails, then this in-flight data would no longer be mirrored.

If the local NVRAM reports an error, the node shuts down. This shutdown results in failover to a HA partner controller. No data is lost because the controller experiencing the failure has not acknowledged the write operation.

ONTAP does not permit a failover when the data is out of sync unless the failover is forced. Forcing a change in conditions in this manner acknowledges that data might be left behind in the original controller and that data loss is acceptable.

Databases are especially vulnerable to corruption if a failover is forced because databases maintain large internal caches of data on disk. If a forced failover occurs, previously acknowledged changes are effectively discarded. The contents of the storage array effectively jump backward in time, and the state of the database cache no longer reflects the state of the data on disk.

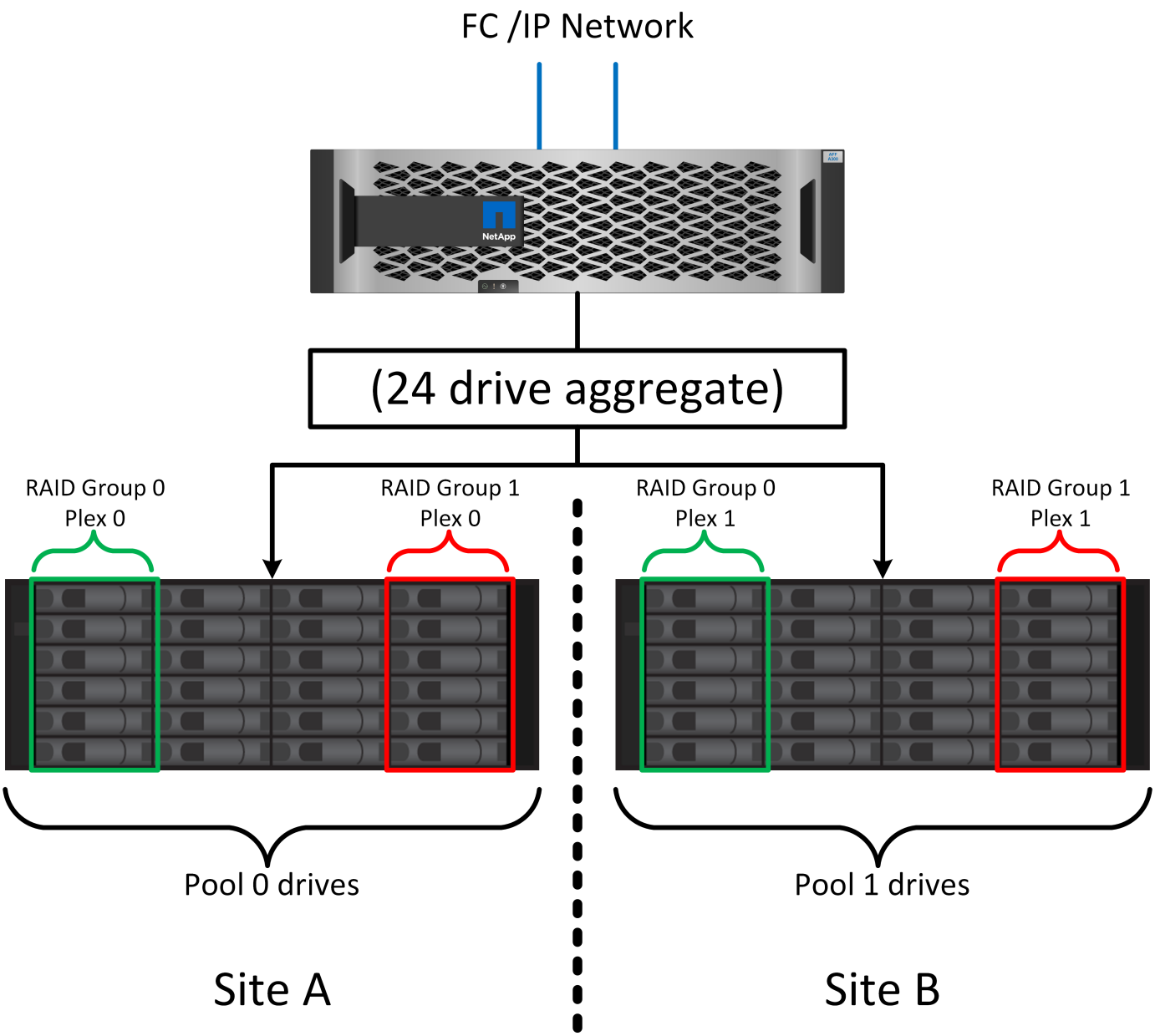
To protect data from this situation, ONTAP allows volumes to be configured for special protection against NVRAM failure. When triggered, this protection mechanism results in a volume entering a state called NVFAIL. This state results in I/O errors that cause a an application shutdown so that they do not use stale data. Data should not be lost because any acknowledged write should be present on the storage array.

The usual next steps are for an administrator to fully shut down the hosts before manually placing the LUNs and volumes back online again. Although these steps can involve some work, this approach is the safest way to make sure of data integrity. Not all data requires this protection, which is why NVFAIL behavior can be configured on a volume-by-volume basis.

Site and shelf failure protection: SyncMirror and plexes

SyncMirror is a mirroring technology that enhances, but does not replace, RAID DP or RAID-TEC. It mirrors the contents of two independent RAID groups. The logical configuration is as follows:

- Drives are configured into two pools based on location. One pool is composed of all drives on site A, and the second pool is composed of all drives on site B.
- A common pool of storage, known as an aggregate, is then created based on mirrored sets of RAID groups. An equal number of drives is drawn from each site. For example, a 20-drive SyncMirror aggregate would be composed of 10 drives from site A and 10 drives from site B.
- Each set of drives on a given site is automatically configured as one or more fully redundant RAID-DP or RAID-TEC groups, independent of the use of mirroring. This provides continuous data protection, even after the loss of a site.



The figure above illustrates a sample SyncMirror configuration. A 24-drive aggregate was created on the controller with 12 drives from a shelf allocated on Site A and 12 drives from a shelf allocated on Site B. The drives were grouped into two mirrored RAID groups. RAID Group 0 includes a 6-drive plex on Site A mirrored to a 6-drive plex on Site B. Likewise, RAID Group 1 includes a 6-drive plex on Site A mirrored to a 6-drive plex on Site B.

SyncMirror is normally used to provide remote mirroring with MetroCluster systems, with one copy of the data at each site. On occasion, it has been used to provide an extra level of redundancy in a single system. In particular, it provides shelf-level redundancy. A drive shelf already contains dual power supplies and controllers and is overall little more than sheet metal, but in some cases the extra protection might be warranted. For example, one NetApp customer has deployed SyncMirror for a mobile real-time analytics platform used during automotive testing. The system was separated into two physical racks supplied by independent power feeds from independent UPS systems.

==Checksums

The topic of checksums is of particular interest to DBAs who are accustomed to using Oracle RMAN streaming

backups migrates to snapshot-based backups. One feature of RMAN is that it performs integrity checks during backup operations. Although this feature has some value, its primary benefit is for a database that is not used on a modern storage array. When physical drives are used for an Oracle database, it is nearly certain that corruption eventually occurs as the drives age, a problem that is addressed by array-based checksums in true storage arrays.

With a real storage array, data integrity is protected by using checksums at multiple levels. If data is corrupted in an IP-based network, the Transmission Control Protocol (TCP) layer rejects the packet data and requests retransmission. The FC protocol includes checksums, as does encapsulated SCSI data. After it is on the array, ONTAP has RAID and checksum protection. Corruption can occur, but, as in most enterprise arrays, it is detected and corrected. Typically, an entire drive fails, prompting a RAID rebuild, and database integrity is unaffected. Less often, ONTAP detects a checksum error, meaning that data on the drive is damaged. The drive is then failed out and a RAID rebuild begins. Once again, data integrity is unaffected.

The Oracle datafile and redo log architecture is also designed to deliver the highest possible level of data integrity, even under extreme circumstances. At the most basic level, Oracle blocks include checksum and basic logical checks with almost every I/O. If Oracle has not crashed or taken a tablespace offline, then the data is intact. The degree of data integrity checking is adjustable, and Oracle can also be configured to confirm writes. As a result, almost all crash and failure scenarios can be recovered, and in the extremely rare event of an unrecoverable situation, corruption is promptly detected.

Most NetApp customers using Oracle databases discontinue the use of RMAN and other backup products after migrating to snapshot-based backups. There are still options in which RMAN can be used to perform block-level recovery with SnapCenter. However, on a day-to-day basis, RMAN, NetBackup, and other products are only used occasionally to create monthly or quarterly archival copies.

Some customers choose to run `dbv` periodically to perform integrity checks on their existing databases. NetApp discourages this practice because it creates unnecessary I/O load. As discussed above, if the database was not previously experiencing problems, the chance of `dbv` detecting a problem is close to zero, and this utility creates a very high sequential I/O load on the network and storage system. Unless there is reason to believe corruption exists, such as exposure to a known Oracle bug, there is no reason to run `dbv`.

Backup and recovery basics

Oracle databases and snapshot-based backups

The foundation of Oracle database data protection on ONTAP is NetApp Snapshot technology.

The key values are as follows:

- **Simplicity.** A snapshot is a read-only copy of the contents of a container of data at a specific point in time.
- **Efficiency.** Snapshots require no space at the moment of creation. Space is only consumed when data is changed.
- **Manageability.** A backup strategy based on snapshots is easy to configure and manage because snapshots are a native part of the storage OS. If the storage system is powered on, it is ready to create backups.
- **Scalability.** Up to 1024 backups of a single container of files and LUNs can be preserved. For complex datasets, multiple containers of data can be protected by a single, consistent set of snapshots.
- Performance is unaffected, whether a volume contains 1024 snapshots or none.

Although many storage vendors offer snapshot technology, the Snapshot technology within ONTAP is unique

and offers significant benefits to enterprise application and database environments:

- Snapshot copies are part of the underlying Write-Anywhere File Layout (WAFL). They are not an add-on or external technology. This simplifies management because the storage system is the backup system.
- Snapshot copies do not affect performance, except for some edge cases such as when so much data is stored in snapshots that the underlying storage system fills up.
- The term "consistency group" is often used to refer to a grouping of storage objects that are managed as a consistent collection of data. A snapshot of a particular ONTAP volume constitutes consistency group backup.

ONTAP snapshots also scale better than competing technology. Customers can store 5, 50, or 500 snapshots without affecting performance. The maximum number of snapshots currently allowed in a volume is 1024. If additional snapshot retention is required, there are options to cascade the snapshots to additional volumes.

As a result, protecting a dataset hosted on ONTAP is simple and highly scalable. Backups do not require movement of data, therefore a backup strategy can be tailored to the needs of the business rather than the limitations of network transfer rates, large number of tape drives, or disk staging areas.

Is a snapshot a backup?

One commonly asked question about the use of snapshots as a data protection strategy is the fact that the "real" data and the snapshot data are located on the same drives. Loss of those drives would result in the loss of both the primary data and the backup.

This is a valid concern. Local snapshots are used for day-to-day backup and recovery needs, and in that respect the snapshot is a backup. Close to 99% of all recovery scenarios in NetApp environments rely on snapshots to meet even the most aggressive RTO requirements.

Local snapshots should, however, never be the only backup strategy, which is why NetApp offers technology such as SnapMirror and SnapVault replication to quickly and efficiently replicate snapshots to an independent set of drives. In a properly architected solution with snapshots plus snapshot replication, the use of tape can be minimized to perhaps a quarterly archive or eliminated entirely.

Snapshot-based backups

There are many options for using ONTAP Snapshot copies to protect your data, and snapshots are the basis for many other ONTAP features, including replication, disaster recovery, and cloning. A complete description of snapshot technology is beyond the scope of this document, but the following sections provide a general overview.

There are two primary approaches to creating a snapshot of a dataset:

- Crash-consistent backups
- Application-consistent backups

A crash-consistent backup of a dataset refers to the capture of the entire dataset structure at a single point in time. If the dataset is stored in a single NetApp FlexVol volume, then the process is simple; a Snapshot can be created at any time. If a dataset spans volumes, a consistency group (CG) snapshot must be created. Several options exist for creating CG snapshots, including NetApp SnapCenter software, native ONTAP consistency group features, and user-maintained scripts.

Crash-consistent backups are primarily used when point-of-the-backup recovery is sufficient. When more granular recover is required, application-consistent backups are usually required.

The word "consistent" in "application-consistent" is often a misnomer. For example, placing an Oracle database in backup mode is referred to as an application-consistent backup, but the data is not made consistent or quiesced in any way. The data continue to change throughout the backup. In contrast, most MySQL and Microsoft SQL Server backups do indeed quiesce the data before executing the backup. VMware may or may not make certain files consistent.

Consistency groups

The term "consistency group" refers to the ability of a storage array to manage multiple storage resources as a single image. For example, a database might consist of 10 LUNs. The array must be able to back up, restore, and replicate those 10 LUNs in a consistent manner. Restoration is not possible if the images of the LUNs were not consistent at the point of backup. Replicating those 10 LUNs requires that all the replicas are perfectly synchronized with each other.

The term "consistency group" is not often used when discussing ONTAP because consistency has always been a basic function of the volume and aggregate architecture within ONTAP. Many other storage arrays manage LUNs or file systems as individual units. They could then be optionally configured as a "consistency group" for purposes of data protection, but this is an extra step in the configuration.

ONTAP has always been able to capture consistent local and replicated images of data. Although the various volumes on an ONTAP system are not usually formally described as a consistency group, that is what they are. A snapshot of that volume is a consistency group image, restoration for that snapshot is a consistency group restoration, and both SnapMirror and SnapVault offer consistency group replication.

Consistency group snapshots

Consistency group snapshots (cg-snapshots) are an extension of the basic ONTAP Snapshot technology. A standard snapshot operation creates a consistent image of all data within a single volume, but sometimes it is necessary to create a consistent set of snapshots across multiple volumes and even across multiple storage systems. The result is a set of snapshots that can be used in the same way as a snapshot of just one individual volume. They can be used for local data recovery, replicated for disaster recovery purposes, or cloned as a single consistent unit.

The largest known use of cg-snapshots is for a database environment of approximately 1PB in size spanning 12 controllers. The cg-snapshots created on this system have been used for backup, recovery and cloning.

Most of the time, when a data set spans volumes and write order must be preserved, a cg-snapshot is automatically used by the chosen management software. There is no need to understand the technical details of cg-snapshots in such cases. However, there are situations in which complicated data protection requirements require detailed control over the data protection and replication process. Automation workflows or the use of custom scripts to call the cg-snapshot APIs are some of options. Understanding the best option and the role of cg-snapshot requires a more detailed explanation of the technology.

Creation of a set of cg-snapshots is a two-step process:

1. Establish write fencing on all target volumes.
2. Create snapshots of those volumes while in the fenced state.

Write fencing is established serially. This means that as the fencing process is set up across multiple volumes, write I/O is frozen on the first volume in the sequence as it continues to be committed to volumes that appear later. This might initially appear to violate the requirement for write order to be preserved, but that only applies to I/O that is issued asynchronously on the host and does not depend on any other writes.

For example, a database might issue a lot of asynchronous datafile updates and allow the OS to reorder the I/O and complete them according to its own scheduler configuration. The order of this type of I/O cannot be

guaranteed because the application and operating system have already released the requirement to preserve write order.

As a counter example, most database logging activity is synchronous. The database does not proceed with further log writes until the I/O is acknowledged, and the order of those writes must be preserved. If a log I/O arrives on a fenced volume, it is not acknowledged and the application blocks on further writes. Likewise, file system metadata I/O is usually synchronous. For example, a file deletion operation must not be lost. If an operating system with an xfs file system deleted a file and the I/O that updated the xfs file system metadata to remove the reference to that file landed on a fenced volume, then the file system activity would pause. This guarantees the integrity of the file system during cg-snapshot operations.

After write fencing is set up across the target volumes, they are ready for snapshot creation. The snapshots need not be created at precisely the same time because the state of the volumes is frozen from a dependent write point of view. To guard against a flaw in the application creating the cg-snapshots, the initial write fencing includes a configurable timeout in which ONTAP automatically releases the fencing and resumes write processing after a defined number of seconds. If all the snapshots are created before the timeout period lapses, then the resulting set of snapshots are a valid consistency group.

Dependent write order

From a technical point of view, the key to a consistency group is preserving write order and, specifically, dependent write order. For example, a database writing to 10 LUNs writes simultaneously to all of them. Many writes are issued asynchronously, meaning that the order in which they are completed is unimportant and the actual order they are completed varies based on operating system and network behavior.

Some write operations must be present on disk before the database can proceed with additional writes. These critical write operations are called dependent writes. Subsequent write I/O depends on the presence of these writes on disk. Any snapshot, recovery, or replication of these 10 LUNs must make sure that dependent write order is guaranteed. File system updates are another example of write-order dependent writes. The order in which file system changes are made must be preserved or the entire file system could become corrupt.

Strategies

There are two primary approaches to snapshot-based backups:

- Crash-consistent backups
- Snapshot-protected hot backups

A crash-consistent backup of a database refers to the capture of the entire database structure, including datafiles, redo logs, and control files, at a single point in time. If the database is stored in a single NetApp FlexVol volume, then the process is simple; a Snapshot can be created at any time. If a database spans volumes, a consistency group (CG) snapshot must be created. Several options exist for creating CG snapshots, including NetApp SnapCenter software, native ONTAP consistency group features, and user-maintained scripts.

Crash-consistent Snapshot backups are primarily used when point-of-the-backup recovery is sufficient. Archive logs can be applied under some circumstances, but when more granular point-in-time recovery is required, a online backup is preferable.

The basic procedure for a snapshot-based online backup is as follows:

1. Place the database in `backup` mode.
2. Create a snapshot of all volumes hosting datafiles.

3. Exit `backup` mode.
4. Run the command `alter system archive log current` to force log archiving.
5. Create snapshots of all volumes hosting the archive logs.

This procedure yields a set of snapshots containing datafiles in backup mode and the critical archive logs generated while in backup mode. These are the two requirements for recovering a database. Files such as control files should also be protected for convenience, but the only absolute requirement is protection for datafiles and archive logs.

Although different customers might have very different strategies, almost all of these strategies are ultimately based on the the same principles outlined below.

Snapshot-based recovery

When designing volume layouts for Oracle databases, the first decision is whether to use volume-based NetApp SnapRestore (VBSR) technology.

Volume-based SnapRestore allows a volume to be almost instantly reverted to an earlier point in time. Because all of the data on the volume is reverted, VBSR might not be appropriate for all use cases. For example, if an entire database, including datafiles, redo logs, and archive logs, is stored on a single volume and this volume is restored with VBSR, then data is lost because the newer archive log and redo data are discarded.

VBSR is not required for restore. Many databases can be restored by using file-based single-file SnapRestore (SFSR) or by simply copying files from the snapshot back into the active file system.

VBSR is preferred when a database is very large or when it must be recovered as quickly as possible, and the use of VBSR requires isolation of the datafiles. In an NFS environment, the datafiles of a given database must be stored in dedicated volumes that are uncontaminated by any other type of file. In a SAN environment, datafiles must be stored in dedicated LUNs on dedicated FlexVol volumes. If a volume manager is used (including Oracle Automatic Storage Management [ASM]), the diskgroup must also be dedicated to datafiles.

Isolating datafiles in this manner allows them to be reverted to an earlier state without damaging other file systems.

Snapshot reserve

For each volume with Oracle data in a SAN environment, the `percent-snapshot-space` should be set to zero because reserving space for a snapshot in a LUN environment is not useful. If the fractional reserve is set to 100, a snapshot of a volume with LUNs requires enough free space in the volume, excluding the snapshot reserve, to absorb 100% turnover of all of the data. If the fractional reserve is set to a lower value, then a correspondingly smaller amount of free space is required, but it always excludes the snapshot reserve. This means that the snapshot reserve space in a LUN environment is wasted.

In an NFS environment, there are two options:

- Set the `percent-snapshot-space` based on expected snapshot space consumption.
- Set the `percent-snapshot-space` to zero and manage active and snapshot space consumption collectively.

With the first option, `percent-snapshot-space` is set to a nonzero value, typically around 20%. This space is then hidden from the user. This value does not, however, create a limit on utilization. If a database with a 20% reservation experiences 30% turnover, the snapshot space can grow beyond the bounds of the 20%

reserve and occupy unreserved space.

The main benefit of setting a reserve to a value such as 20% is to verify that some space is always available for snapshots. For example, a 1TB volume with a 20% reserve would only permit a database administrator (DBA) to store 800GB of data. This configuration guarantees at least 200GB of space for snapshot consumption.

When `percent-snapshot-space` is set to zero, all space in the volume is available to the end user, which delivers better visibility. A DBA must understand that, if he or she sees a 1TB volume that leverages snapshots, this 1TB of space is shared between active data and Snapshot turnover.

There is no clear preference between option one and option two among end users.

ONTAP and third-party snapshots

Oracle Doc ID 604683.1 explains the requirements for third-party snapshot support and the multiple options available for backup and restore operations.

The third-party vendor must guarantee that the company's snapshots conform to the following requirements:

- Snapshots must integrate with Oracle's recommended restore and recovery operations.
- Snapshots must be database crash consistent at the point of the snapshot.
- Write ordering is preserved for each file within a snapshot.

ONTAP and NetApp Oracle management products comply with these requirements.

Rapid Oracle database recovery with SnapRestore

Rapid data restoration in ONTAP from a snapshot is delivered by NetApp SnapRestore technology.

When a critical dataset is unavailable, critical business operations are down. Tapes can break, and even restores from disk-based backups can be slow to transfer across the network. SnapRestore avoids these problems by delivering near instantaneous restoration of datasets. Even petabyte-scale databases can be completely restored with just a few minutes of effort.

There are two forms of SnapRestore - file/LUN-based and volume-based.

- Individual files or LUNs can be restored in seconds, whether it is a 2TB LUN or a 4KB file.
- The container of files or LUNs can be restored in seconds, whether it is 10GB or 100TB of data.

A "container of files or LUNs" would typically refer to a FlexVol volume. For example, you may have 10 LUNs that make up a LVM diskgroup in a single volume, or a volume might store the NFS home directories of 1000 users. Rather than executing a restore operation for each individual file or LUN, you can restore the entire volume as a single operation. This process also work with scale-out containers that include multiple volumes, such as a FlexGroup or an ONTAP Consistency Group.

The reason SnapRestore works so quickly and efficiently is due to the nature of a snapshot, which is essentially a parallel read-only view of the contents of a volume at a specific point in time. The active blocks are the real blocks that can be changed, while the snapshot is a read-only view into the state of the blocks that constitute the files and LUNs at the time the snapshot was created.

ONTAP only permits read-only access to snapshot data, but the data can be reactivated with SnapRestore. The snapshot is reenabled as a read-write view of the data, returning the data to its prior state. SnapRestore

can operate at the volume or the file level. The technology is essentially the same with a few minor differences in behavior.

Volume SnapRestore

Volume-based SnapRestore returns the entire volume of data to an earlier state. This operation does not require data movement, meaning that the restore process is essentially instantaneous, although the API or CLI operation might take a few seconds to be processed. Restoring 1GB of data is no more complicated or time-consuming than restoring 1PB of data. This capability is the primary reason many enterprise customers migrate to ONTAP storage systems. It delivers an RTO measured in seconds for even the largest datasets.

One drawback to volume-based SnapRestore is caused by the fact that changes within a volume are cumulative over time. Therefore, each snapshot and the active file data are dependent on the changes leading up to that point. Reverting a volume to an earlier state means discarding all the subsequent changes that had been made to the data. What is less obvious, however, is that this includes subsequently created snapshots. This is not always desirable.

For example, a data retention SLA might specify 30 days of nightly backups. Restoring a dataset to a snapshot created five days ago with volume SnapRestore would discard all the snapshots created on the previous five days, violating the SLA.

There are a number of options available to address this limitation:

1. Data can be copied from a prior snapshot, as opposed to performing a SnapRestore of the entire volume. This method works best with smaller datasets.
2. A snapshot can be cloned rather than restored. The limitation to this approach is that the source snapshot is a dependency of the clone. Therefore, it cannot be deleted unless the clone is also deleted or is split into an independent volume.
3. Use of file-based SnapRestore.

File SnapRestore

File-based SnapRestore is a more granular snapshot-based restoration process. Rather than reverting the state of an entire volume, the state of an individual file or LUN is reverted. No snapshots need to be deleted, nor does this operation create any dependency on a prior snapshot. The file or LUN becomes immediately available in the active volume.

No data movement is required during a SnapRestore restore of a file or LUN. However, some internal metadata updates are required to reflect the fact that the underlying blocks in a file or LUN now exist in both a snapshot and the active volume. There should be no effect on performance, but this process blocks the creation of snapshots until it is complete. The processing rate is approximately 5GBps (18TB/hour) based on the total size of the files restored.

Oracle database online backups

Two sets of data are required to protect and recover an Oracle database in backup mode. Note that this is not the only Oracle backup option, but it is the most common.

- A snapshot of the datafiles in backup mode
- The archive logs created while the datafiles were in backup mode

If complete recovery including all committed transactions is required, a third item is required:

- A set of current redo logs

There are a number of ways to drive recovery of an online backup. Many customers restore snapshots by using the ONTAP CLI and then using Oracle RMAN or sqlplus to complete the recovery. This is especially common with large production environments in which the probability and frequency of database restores is extremely low and any restore procedure is handled by a skilled DBA. For complete automation, solutions such as NetApp SnapCenter include an Oracle plug-in with both command- line and graphical interfaces.

Some large-scale customers have taken a simpler approach by configuring basic scripting on the hosts to place the databases in backup mode at a specific time in preparation for a scheduled snapshot. For example, schedule the command `alter database begin backup` at 23:58, `alter database end backup` at 00:02, and then schedule snapshots directly on the storage system at midnight. The result is a simple, highly scalable backup strategy that requires no external software or licenses.

Data layout

The simplest layout is to isolate datafiles into one or more dedicated volumes. They must be uncontaminated by any other file type. This is to make sure that the datafile volumes can be rapidly restored through a SnapRestore operation without destroying an important redo log, controlfile, or archive log.

SAN has similar requirements for datafile isolation within dedicated volumes. With an operating system such as Microsoft Windows, a single volume might contain multiple datafile LUNs, each with an NTFS file system. With other operating systems, there is generally a logical volume manager. For example, with Oracle ASM, the simplest option would be to confine the LUNs of an ASM disk group to a single volume that can be backed up and restored as a unit. If additional volumes are required for performance or capacity management reasons, creating an additional disk group on the new volume results in simpler management.

If these guidelines are followed, snapshots can be scheduled directly on the storage system with no requirement for performing a consistency group snapshot. The reason is that Oracle backups do not require datafiles to be backed up at the same time. The online backup procedure was designed to allow datafiles to continue to be updated as they are slowly streamed to tape over the course of hours.

A complication arises in situations such as the use of an ASM disk group that is distributed across volumes. In these cases, a cg-snapshot must be performed to make sure that the ASM metadata is consistent across all constituent volumes.

Caution: Verify that the ASM `spfile` and `passwd` files are not in the disk group hosting the datafiles. This interferes with the ability to selectively restore datafiles and only datafiles.

Local recovery procedure—NFS

This procedure can be driven manually or through an application such as SnapCenter. The basic procedure is as follows:

1. Shut down the database.
2. Recover the datafile volume(s) to the snapshot immediately prior to the desired restore point.
3. Replay archive logs to the desired point.
4. Replay current redo logs if complete recovery is desired.

This procedure assumes that the desired archive logs are still present in the active file system. If they are not, the archive logs must be restored or rman/sqlplus can be directed to the data in the snapshot directory.

In addition, for smaller databases, datafiles can be recovered by an end user directly from the `.snapshot` directory without assistance from automation tools or storage administrators to execute a `snaprestore`

command.

Local recovery procedure—SAN

This procedure can be driven manually or through an application such as SnapCenter. The basic procedure is as follows:

1. Shut down the database.
2. Quiesce the disk group(s) hosting the datafiles. The procedure varies depending on the logical volume manager chosen. With ASM, the process requires dismounting the disk group. With Linux, the file systems must be dismounted, and the logical volumes and volume groups must be deactivated. The objective is to stop all updates on the target volume group to be restored.
3. Restore the datafile disk groups to the snapshot immediately prior to the desired restore point.
4. Reactivate the newly restored disk groups.
5. Replay archive logs to the desired point.
6. Replay all redo logs if complete recovery is desired.

This procedure assumes that the desired archive logs are still present in the active file system. If they are not, the archive logs must be restored by taking the archive log LUNs offline and performing a restore. This is also an example in which dividing up archive logs into dedicated volumes is useful. If the archive logs share a volume group with redo logs, then the redo logs must be copied elsewhere before restoration of the overall set of LUNs. This step prevents the loss of those final recorded transactions.

Oracle Database Storage Snapshot Optimized backups

Snapshot-based backup and recovery became even simpler back when Oracle 12c was released because there is no need to place a database in hot backup mode. The result is an ability to schedule snapshot-based backups directly on a storage system and still preserve the ability to perform complete or point-in-time recovery.

Although the hot backup recovery procedure is more familiar to DBAs, it has, for a long time, been possible to use snapshots that were not created while the database was in hot backup mode. Extra manual steps were required with Oracle 10g and 11g during recovery to make the database consistent. With Oracle 12c, `sqlplus` and `rman` contain the extra logic to replay archive logs on datafile backups that were not in hot backup mode.

As discussed previously, recovering a snapshot-based hot backup requires two sets of data:

- A snapshot of the datafiles created while in backup mode
- The archive logs generated while the datafiles were in hot backup mode

During recovery, the database reads metadata from the datafiles to select the required archive logs for recovery.

Storage snapshot-optimized recovery requires slightly different datasets to accomplish the same results:

- A snapshot of the datafiles, plus a method to identify the time the snapshot was created
- Archive logs from the time of the most recent datafile checkpoint through the exact time of the snapshot

During recovery, the database reads metadata from the datafiles to identify the earliest archive log required. Full or point-in-time recovery can be performed. When performing a point-in-time recovery, it is critical to know the time of the snapshot of the datafiles. The specified recovery point must be after the creation time of the

snapshots. NetApp recommends adding at least a few minutes to the snapshot time to account for clock variation.

For complete details, see Oracle's documentation on the topic, "Recovery Using Storage Snapshot Optimization" available in various releases of the Oracle 12c documentation. Also, see Oracle Document ID Doc ID 604683.1 regarding Oracle third-party snapshot support.

Data layout

The simplest layout is to isolate the datafiles into one or more dedicated volumes. They must be uncontaminated by any other file type. This is to make sure that the datafile volumes can be rapidly restored with a SnapRestore operation without destroying an important redo log, controlfile, or archive log.

SAN has similar requirements for datafile isolation within dedicated volumes. With an operating system such as Microsoft Windows, a single volume might contain multiple datafile LUNs, each with an NTFS file system. With other operating systems, there is generally a logical volume manager as well. For example, with Oracle ASM, the simplest option would be to confine disk groups to a single volume that can be backed up and restored as a unit. If additional volumes are required for performance or capacity management reasons, creating an additional disk group on the new volume results in easier management.

If these guidelines are followed, snapshots can be scheduled directly on ONTAP with no requirement for performing a consistency group snapshot. The reason is that snapshot-optimized backups do not require that datafiles be backed up at the same time.

A complication arises in situations such as an ASM disk group that is distributed across volumes. In these cases, a cg-snapshot must be performed to make sure that the ASM metadata is consistent across all constituent volumes.

[Note]Verify that the ASM spfile and passwd files are not in the disk group hosting the datafiles. This interferes with the ability to selectively restore datafiles and only datafiles.

Local recovery procedure—NFS

This procedure can be driven manually or through an application such as SnapCenter. The basic procedure is as follows:

1. Shut down the database.
2. Recover the datafile volume(s) to the snapshot immediately prior to the desired restore point.
3. Replay archive logs to the desired point.

This procedure assumes that the desired archive logs are still present in the active file system. If they are not, the archive logs must be restored, or `rman` or `sqlplus` can be directed to the data in the `.snapshot` directory.

In addition, for smaller databases, datafiles can be recovered by an end user directly from the `.snapshot` directory without assistance from automation tools or a storage administrator to execute a SnapRestore command.

Local recovery procedure—SAN

This procedure can be driven manually or through an application such as SnapCenter. The basic procedure is as follows:

1. Shut down the database.

2. Quiesce the disk group(s) hosting the datafiles. The procedure varies depending on the logical volume manager chosen. With ASM, the process requires dismounting the disk group. With Linux, the file systems must be dismounted, and the logical volumes and volume groups are deactivated. The objective is to stop all updates on the target volume group to be restored.
3. Restore the datafile disk groups to the snapshot immediately prior to the desired restore point.
4. Reactivate the newly restored disk groups.
5. Replay archive logs to the desired point.

This procedure assumes that the desired archive logs are still present in the active file system. If they are not, the archive logs must be restored by taking the archive log LUNs offline and performing a restore. This is also an example in which dividing up archive logs into dedicated volumes is useful. If the archive logs share a volume group with redo logs, the redo logs must be copied elsewhere before restoration of the overall set of LUNs to avoid losing the final recorded transactions.

Full recovery example

Assume the datafiles have been corrupted or destroyed and full recovery is required. The procedure to do so is as follows:

```
[oracle@host1 ~]$ sqlplus / as sysdba
Connected to an idle instance.
SQL> startup mount;
ORACLE instance started.
Total System Global Area 1610612736 bytes
Fixed Size                  2924928 bytes
Variable Size            1040191104 bytes
Database Buffers          553648128 bytes
Redo Buffers               13848576 bytes
Database mounted.
SQL> recover automatic;
Media recovery complete.
SQL> alter database open;
Database altered.
SQL>
```

Point-in-time recovery example

The entire recovery procedure is a single command: `recover automatic`.

If point-in-time recovery is required, the timestamp of the snapshot(s) must be known and can be identified as follows:

```
Cluster01::> snapshot show -vserver vserver1 -volume NTAP_oradata -fields
create-time
vserver   volume        snapshot   create-time
--------  ------------  ---------  ------------------------
vserver1  NTAP_oradata  my-backup  Thu Mar 09 10:10:06 2017
```

The snapshot creation time is listed as March 9th and 10:10:06. To be safe, one minute is added to the snapshot time:

```
[oracle@host1 ~]$ sqlplus / as sysdba
Connected to an idle instance.
SQL> startup mount;
ORACLE instance started.
Total System Global Area 1610612736 bytes
Fixed Size                  2924928 bytes
Variable Size            1040191104 bytes
Database Buffers          553648128 bytes
Redo Buffers               13848576 bytes
Database mounted.
SQL> recover database until time '09-MAR-2017 10:44:15' snapshot time '09-
MAR-2017 10:11:00';
```

The recovery is now initiated. It specified a snapshot time of 10:11:00, one minute after the recorded time to account for possible clock variance, and a target recovery time of 10:44. Next, sqlplus requests the archive logs required to reach the desired recovery time of 10:44.

```
ORA-00279: change 551760 generated at 03/09/2017 05:06:07 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_31_930813377.dbf
ORA-00280: change 551760 for thread 1 is in sequence #31
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 552566 generated at 03/09/2017 05:08:09 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_32_930813377.dbf
ORA-00280: change 552566 for thread 1 is in sequence #32
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 553045 generated at 03/09/2017 05:10:12 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_33_930813377.dbf
ORA-00280: change 553045 for thread 1 is in sequence #33
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 753229 generated at 03/09/2017 05:15:58 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_34_930813377.dbf
ORA-00280: change 753229 for thread 1 is in sequence #34
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
Log applied.
Media recovery complete.
SQL> alter database open resetlogs;
Database altered.
SQL>
```

Complete recovery of a database using snapshots using the `recover automatic` command does not require specific licensing, but point-in-time recovery using `snapshot time` requires the Oracle Advanced Compression license.

Oracle database management and automation tools

The primary value of ONTAP in an Oracle database environment comes from the core ONTAP technologies such as instant Snapshot copies, simple SnapMirror replication, and efficient creation of FlexClone volumes.

In some cases, simple configuration of these core features directly on ONTAP meets requirements, but more complicated needs require an orchestration layer.

SnapCenter

SnapCenter is the flagship NetApp data protection product. At a very low level, it is similar to the SnapManager products in terms of how it executes database backups, but it was built from the ground up to deliver a single-pane-of-glass for data protection management on NetApp storage systems.

SnapCenter includes the basic functions such as snapshot-based backups and restores, SnapMirror and SnapVault replication, and other features required to operate at scale for large enterprises. These advanced

features include an expanded role-based access control (RBAC) capability, RESTful APIs to integrate with third-party orchestration products, nondisruptive central management of SnapCenter plug-ins on database hosts, and a user interface designed for cloud-scale environments.

REST

ONTAP also contains a rich RESTful API set. This allows 3rd party vendors to create data protection and other management application with deep integration with ONTAP. Furthermore, the RESTful API is easy to consume by customers who wish to create their own automation workflows and utilities.

Oracle disaster recovery

Oracle database disaster recovery with ONTAP

Disaster recovery refers to restoring data services after a catastrophic event, such as a fire that destroys a storage system or even an entire site.

> This documentation replaces previously published technical reports *TR-4591: Oracle Data Protection* and *TR-4592: Oracle on MetroCluster.*

Disaster recovery can be accomplished by simple replication of data using SnapMirror, of course, with many customers updating mirrored replicas as often as hourly.

For most customers, DR requires more than just possessing a remote copy of data, it requires the ability to rapidly make use of that data. NetApp offers two technologies that address this need - MetroCluster and SnapMirror active sync

MetroCluster refers to ONTAP in a hardware configuration that includes low-level synchronously mirrored storage and numerous additional features. Integrated solutions such as MetroCluster simplify today's complicated, scale-out database, application, and virtualization infrastructures. It replaces multiple, external data protection products and strategies with one simple, central storage array. It also provides integrated backup, recovery, disaster recovery, and high availability (HA) within a single clustered storage system.

SnapMirror active sync is based on SnapMirror Synchronous. With MetroCluster, each ONTAP controller is responsible for replicating its drive data to a remote location. With SnapMirror active sync, you essentially have two different ONTAP systems maintaining independent copies of your LUN data, but cooperating to present a single instance of that LUN. From a host point of view, it's a single LUN entity.

Although SnapMirror active sync and MetroCluster work very differently internally, to a host the result is very similar. The primary difference is granularity. If you only need select workloads to be synchronous replicated, SnapMirror active sync is the better option. If you need to replicate entire environments or even data centers, MetroCluster is a better option. In addition, SnapMirror active sync is currently SAN-only while MetroCluster is multiprotocol, including SAN, NFS, and SMB.

MetroCluster

MetroCluster physical architecture and Oracle databases

Understanding how Oracle databases operate in a MetroCluster environment requires some explanation of physical design of a MetroCluster system.

This documentation replaces previously published technical report *TR-4592: Oracle on MetroCluster.*

MetroCluster is avialable in 3 different configurations

- HA pairs with IP connectivity
- HA pairs with FC connectivity
- Single controller with FC connectivity

[NOTE]The term 'connectivity' refers to the cluster connection used for cross-site replication. It does not refer to the host protocols. All host-side protocols are supported as usual in a MetroCluster configuration irrespective of the type of connection used for inter-cluster communication.

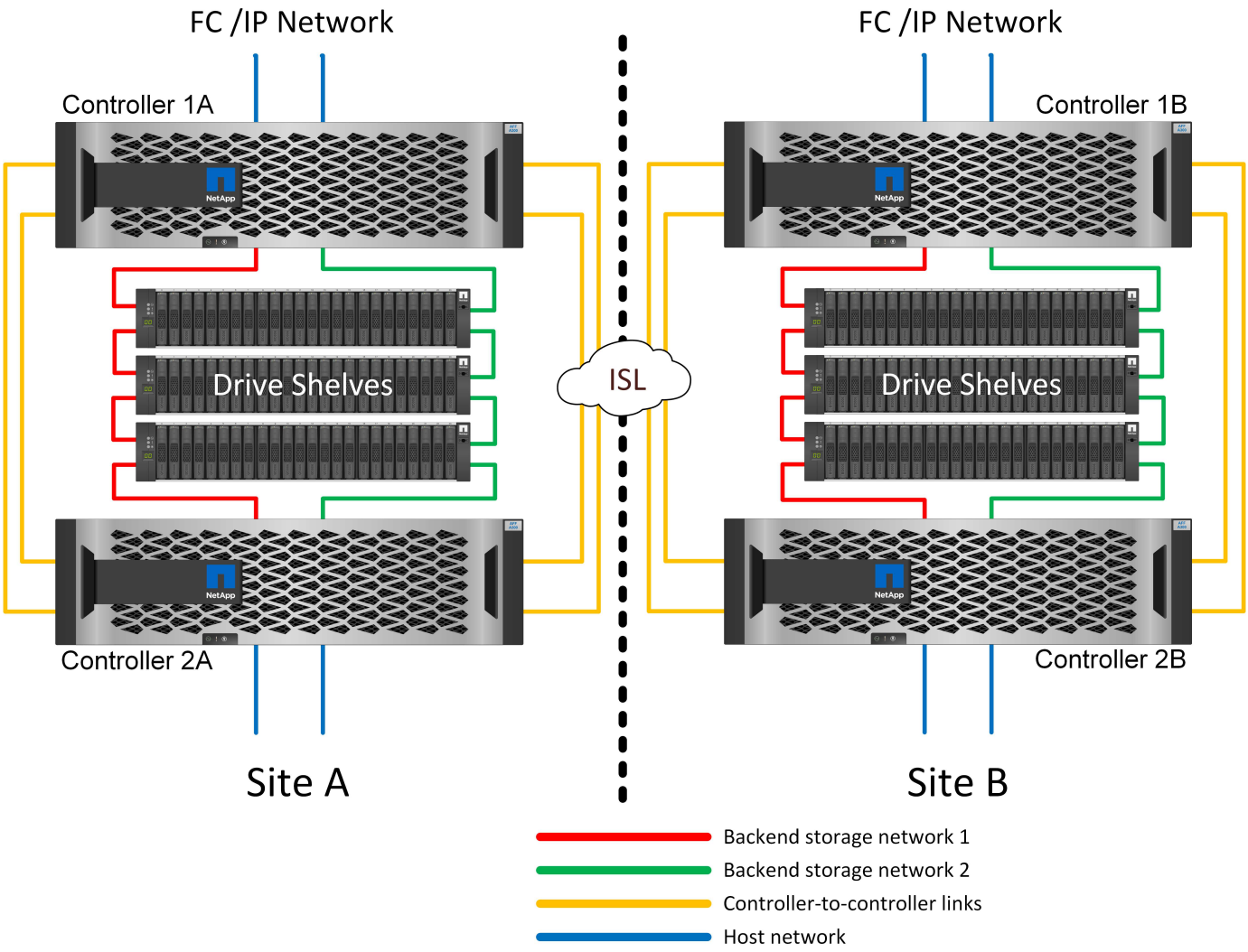
MetroCluster IP

The HA-pair MetroCluster IP configuration uses two or four nodes per site. This configuration option increases the complexity and costs relative to the two-node option, but it delivers an important benefit: intrasite redundancy. A simple controller failure does not require data access across the WAN. Data access remains local through the alternate local controller.

Most customers are choosing IP connectivity because the infrastructure requirements are simpler. In the past, high-speed cross-site connectivity was generally easier to provision using dark fibre and FC switches, but today high-speed, low latency IP circuits are more readily available.

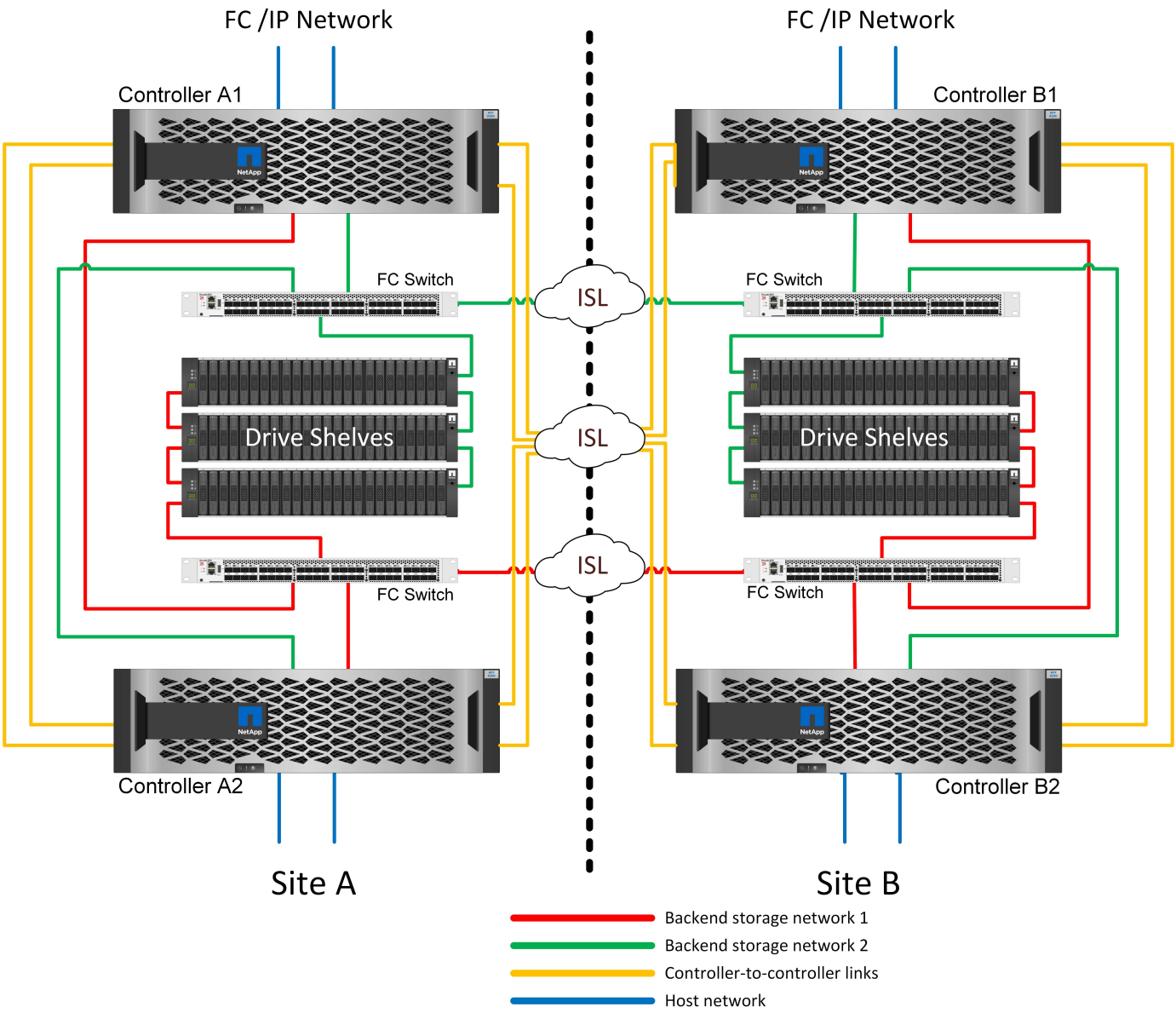
The architecture is also simpler because the only cross-site connections are for the controllers. In FC SAN attached MetroClusters, a controller writes directly to the drives on the opposite site and thus requires additional SAN connections, switches, and bridges. In contrast, a controller in an IP configuration writes to the opposite drives via the controller.

For additional information, refer to the official ONTAP documentation and MetroCluster IP Solution Architecture and Design.



HA-Pair FC SAN-attached MetroCluster

The HA-pair MetroCluster FC configuration uses two or four nodes per site. This configuration option increases the complexity and costs relative to the two-node option, but it delivers an important benefit: intrasite redundancy. A simple controller failure does not require data access across the WAN. Data access remains local through the alternate local controller.

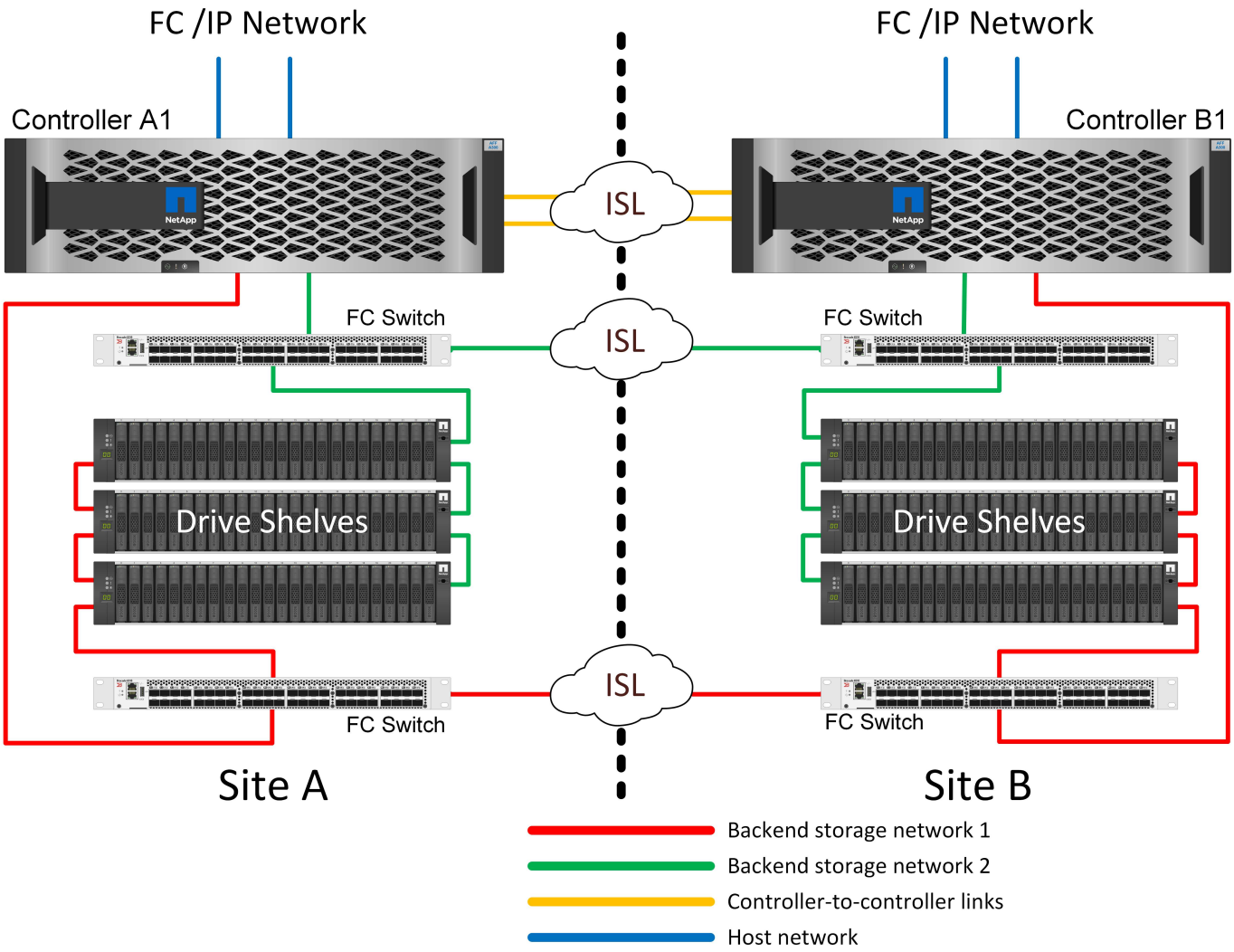


Some multisite infrastructures are not designed for active-active operations, but rather are used more as a primary site and disaster recovery site. In this situation, an HA-pair MetroCluster option is generally preferable for the following reasons:

- Although a two-node MetroCluster cluster is an HA system, unexpected failure of a controller or planned maintenance requires that data services must come online on the opposite site. If the network connectivity between sites cannot support the required bandwidth, performance is affected. The only option would be to also fail over the various host OSs and associated services to the alternate site. The HA-pair MetroCluster cluster eliminates this problem because loss of a controller results in simple failover within the same site.
- Some network topologies are not designed for cross-site access, but instead use different subnets or isolated FC SANs. In these cases, the two-node MetroCluster cluster no longer functions as an HA system because the alternate controller cannot serve data to the servers on the opposite site. The HA-pair MetroCluster option is required to deliver complete redundancy.
- If a two-site infrastructure is viewed as a single highly available infrastructure, the two-node MetroCluster configuration is suitable. However, if the system must function for an extended period of time after site failure, then an HA pair is preferred because it continues to provide HA within a single site.

Two-node FC SAN-attached MetroCluster

The two-node MetroCluster configuration uses only one node per site. This design is simpler than the HA-pair option because there are fewer components to configure and maintain. It also has reduced infrastructure demands in terms of cabling and FC switching. Finally, it reduces costs.



The obvious impact of this design is that controller failure on a single site means that data is available from the opposite site. This restriction is not necessarily a problem. Many enterprises have multisite data center operations with stretched, high-speed, low-latency networks that function essentially as a single infrastructure. In these cases, the two-node version of MetroCluster is the preferred configuration. Two-node systems are currently used at petabyte scale by several service providers.

MetroCluster resiliency features

There are no single points of failure in a MetroCluster solution:

- Each controller has two independent paths to the drive shelves on the local site.
- Each controller has two independent paths to the drive shelves on the remote site.
- Each controller has two independent paths to the controllers on the opposite site.
- In the HA-pair configuration, each controller has two paths to its local partner.

In summary, any one component in the configuration can be removed without compromising the ability of MetroCluster to serve data. The only difference in terms of resiliency between the two options is that the HA-pair version is still an overall HA storage system after a site failure.

MetroCluster logical architecture and Oracle databases

Understanding how Oracle databases operate in a MetroCluster environment alsop requires some explanation of the logical functionality of a MetroCluster system.

Site failure protection: NVRAM and MetroCluster

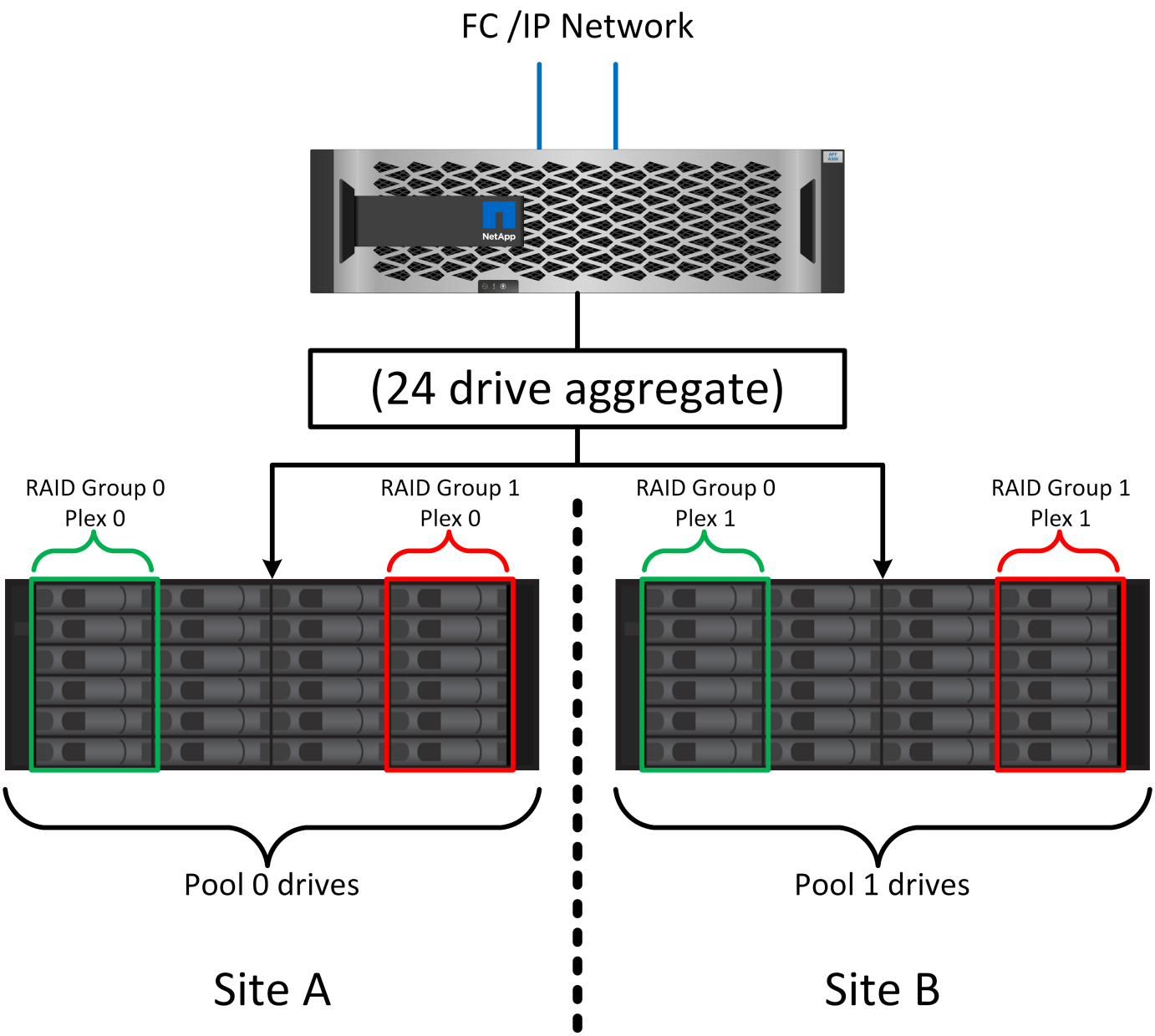
MetroCluster extends NVRAM data protection in the following ways:

- In a two-node configuration, NVRAM data is replicated using the Inter-Switch Links (ISLs) to the remote partner.
- In an HA-pair configuration, NVRAM data is replicated to both the local partner and a remote partner.
- A write is not acknowledged until it is replicated to all partners. This architecture protects in-flight I/O from site failure by replicating NVRAM data to a remote partner. This process is not involved with drive-level data replication. The controller that owns the aggregates is responsible for data replication by writing to both plexes in the aggregate, but there still must be protection against in-flight I/O loss in the event of site loss. Replicated NVRAM data is only used if a partner controller must take over for a failed controller.

Site and shelf failure protection: SyncMirror and plexes

SyncMirror is a mirroring technology that enhances, but does not replace, RAID DP or RAID-TEC. It mirrors the contents of two independent RAID groups. The logical configuration is as follows:

1. Drives are configured into two pools based on location. One pool is composed of all drives on site A, and the second pool is composed of all drives on site B.
2. A common pool of storage, known as an aggregate, is then created based on mirrored sets of RAID groups. An equal number of drives is drawn from each site. For example, a 20-drive SyncMirror aggregate would be composed of 10 drives from site A and 10 drives from site B.
3. Each set of drives on a given site is automatically configured as one or more fully redundant RAID DP or RAID-TEC groups, independent of the use of mirroring. This use of RAID underneath mirroring provides data protection even after the loss of a site.



The figure above illustrates a sample SyncMirror configuration. A 24-drive aggregate was created on the controller with 12 drives from a shelf allocated on site A and 12 drives from a shelf allocated on site B. The drives were grouped into two mirrored RAID groups. RAID group 0 includes a 6-drive plex on site A mirrored to a 6-drive plex on site B. Likewise, RAID group 1 includes a 6-drive plex on site A mirrored to a 6-drive plex on site B.

SyncMirror is normally used to provide remote mirroring with MetroCluster systems, with one copy of the data at each site. On occasion, it has been used to provide an extra level of redundancy in a single system. In particular, it provides shelf-level redundancy. A drive shelf already contains dual power supplies and controllers and is overall little more than sheet metal, but in some cases the extra protection might be warranted. For example, one NetApp customer has deployed SyncMirror for a mobile real-time analytics platform used during automotive testing. The system was separated into two physical racks supplied with independent power feeds and independent UPS systems.

Redundancy failure: NVFAIL

As discussed earlier, a write is not acknowledged until it has been logged into local NVRAM and NVRAM on at

least one other controller. This approach makes sure that a hardware failure or power outage does not result in the loss of in-flight I/O. If the local NVRAM fails or the connectivity to other nodes fails, then data would no longer be mirrored.

If the local NVRAM reports an error, the node shuts down. This shutdown results in failover to a partner controller when HA pairs are used. With MetroCluster, the behavior depends on the overall configuration chosen, but it can result in automatic failover to the remote note. In any case, no data is lost because the controller experiencing the failure has not acknowledged the write operation.

A site-to-site connectivity failure that blocks NVRAM replication to remote nodes is a more complicated situation. Writes are no longer replicated to the remote nodes, creating a possibility of data loss if a catastrophic error occurs on a controller. More importantly, attempting to fail over to a different node during these conditions results in data loss.

The controlling factor is whether NVRAM is synchronized. If NVRAM is synchronized, node-to-node failover is safe to proceed without risk of data loss. In a MetroCluster configuration, if NVRAM and the underlying aggregate plexes are in sync, it is safe to proceed with switchover without risk of data loss.

ONTAP does not permit a failover or switchover when the data is out of sync unless the failover or switchover is forced. Forcing a change in conditions in this manner acknowledges that data might be left behind in the original controller and that data loss is acceptable.

Databases and other applications are especially vulnerable to corruption if a failover or switchover is forced because they maintain larger internal caches of data on disk. If a forced failover or switchover occurs, previously acknowledged changes are effectively discarded. The contents of the storage array effectively jump backward in time, and the state of the cache no longer reflects the state of the data on disk.

To prevent this situation, ONTAP allows volumes to be configured for special protection against NVRAM failure. When triggered, this protection mechanism results in a volume entering a state called NVFAIL. This state results in I/O errors that cause an application crash. This crash causes the applications to shut down so that they do not use stale data. Data should not be lost because any committed transaction data should be present in the logs. The usual next steps are for an administrator to fully shut down the hosts before manually placing the LUNs and volumes back online again. Although these steps can involve some work, this approach is the safest way to make sure of data integrity. Not all data requires this protection, which is why NVFAIL behavior can be configured on a volume-by-volume basis.

HA pairs and MetroCluster

MetroCluster is available in two configurations: two-node and HA pair. The two-node configuration behaves the same as an HA pair with respect to NVRAM. In the event of sudden failure, the partner node can replay NVRAM data to make the drives consistent and make sure that no acknowledged writes have been lost.

The HA-pair configuration replicates NVRAM to the local partner node as well. A simple controller failure results in an NVRAM replay on the partner node, as is the case with a standalone HA-pair without MetroCluster. In the event of sudden complete site loss, the remote site also has the NVRAM required to make the drives consistent and start serving data.

One important aspect of MetroCluster is that the remote nodes have no access to partner data under normal operational conditions. Each site functions essentially as an independent system that can assume the personality of the opposite site. This process is known as a switchover and includes a planned switchover in which site operations are migrated nondisruptively to the opposite site. It also includes unplanned situations in which a site is lost and a manual or automatic switchover is required as part of disaster recovery.

Switchover and switchback

The terms switchover and switchback refer to the process of transitioning volumes between remote controllers in a MetroCluster configuration. This process only applies to the remote nodes. When MetroCluster is used in a four-volume configuration, local node failover is the same takeover and giveback process described previously.

Planned switchover and switchback

A planned switchover or switchback is similar to a takeover or giveback between nodes. The process has multiple steps and might appear to require several minutes, but what is actually happening is a multiphase graceful transition of storage and network resources. The moment when control transfers occurs much more quickly than the time required for the complete command to execute.

The primary difference between takeover/giveback and switchover/switchback is with the effect on FC SAN connectivity. With local takeover/giveback, a host experiences the loss of all FC paths to the local node and relies on its native MPIO to change over to available alternate paths. Ports are not relocated. With switchover and switchback, the virtual FC target ports on the controllers transition to the other site. They effectively cease to exist on the SAN for a moment and then reappear on an alternate controller.

SyncMirror timeouts

SyncMirror is a ONTAP mirroring technology that provides protection against shelf failures. When shelves are separated across a distance, the result is remote data protection.

SyncMirror does not deliver universal synchronous mirroring. The result is better availability. Some storage systems use constant all-or-nothing mirroring, sometimes called domino mode. This form of mirroring is limited in application because all write activity must cease if the connection to the remote site is lost. Otherwise, a write would exist at one site but not at the other. Typically, such environments are configured to take LUNs offline if site-to-site connectivity is lost for more than a short period (such as 30 seconds).

This behavior is desirable for a small subset of environments. However, most applications require a solution that delivers guaranteed synchronous replication under normal operating conditions, but with the ability to suspend replication. A complete loss of site-to-site connectivity is frequently considered a near-disaster situation. Typically, such environments are kept online and serving data until connectivity is repaired or a formal decision is made to shut down the environment to protect data. A requirement for automatic shutdown of the application purely because of remote replication failure is unusual.

SyncMirror supports synchronous mirroring requirements with the flexibility of a timeout. If connectivity to the remote controller and/or plex is lost, a 30- second timer begins counting down. When the counter reaches 0, write I/O processing resumes using the local data. The remote copy of the data is usable, but it is frozen in time until connectivity is restored. Resynchronization leverages aggregate-level snapshots to return the system to synchronous mode as quickly as possible.

Notably, in many cases, this sort of universal all-or-nothing domino mode replication is better implemented at the application layer. For example, Oracle DataGuard includes maximum protection mode, which guarantees long-instance replication under all circumstances. If the replication link fails for a period exceeding a configurable timeout, the databases shut down.

Automatic unattended switchover with Fabric Attached MetroCluster

Automatic unattended switchover (AUSO) is a Fabric Attached MetroCluster feature that delivers a form of cross-site HA. As discussed previously, MetroCluster is available in two types: a single controller on each site or an HA pair on each site. The principal advantage of the HA option is that planned or unplanned controller shutdown still allows all I/O to be local. The advantage of the single-node option is reduced costs, complexity, and infrastructure.

The primary value of AUSO is to improve the HA capabilities of Fabric Attached MetroCluster systems. Each site monitors the health of the opposite site, and, if no nodes remain to serve data, AUSO results in rapid switchover. This approach is especially useful in MetroCluster configurations with just a single node per site because it brings the configuration closer to an HA pair in terms of availability.

AUSO cannot offer comprehensive monitoring at the level of an HA pair. An HA pair can deliver extremely high availability because it includes two redundant physical cables for direct node-to-node communication. Furthermore, both nodes in an HA pair have access to the same set of disks on redundant loops, delivering another route for one node to monitor the health of another.

MetroCluster clusters exist across sites for which both node-to-node communication and disk access rely on the site-to-site network connectivity. The ability to monitor the heartbeat of the rest of the cluster is limited. AUSO has to discriminate between a situation where the other site is actually down rather than unavailable due to a network problem.

As a result, a controller in an HA pair can prompt a takeover if it detects a controller failure that occurred for a specific reason, such as a system panic. It can also prompt a takeover if there is a complete loss of connectivity, sometimes known as a lost heartbeat.

A MetroCluster system can only safely perform an automatic switchover when a specific fault is detected on the original site. Also, the controller taking ownership of the storage system must be able to guarantee that disk and NVRAM data is in sync. The controller cannot guarantee the safety of a switchover just because it lost contact with the source site, which could still be operational. For additional options for automating a switchover, see the information on the MetroCluster tiebreaker (MCTB) solution in the next section.

MetroCluster tiebreaker with fabric attached MetroCluster

The NetApp MetroCluster Tiebreaker software can run on a third site to monitor the health of the MetroCluster environment, send notifications, and optionally force a switchover in a disaster situation. A complete description of the tiebreaker can be found on the NetApp support site, but the primary purpose of the MetroCluster Tiebreaker is to detect site loss. It must also discriminate between site loss and a loss of connectivity. For example, switchover should not occur because the tiebreaker was unable to reach the primary site, which is why the tiebreaker also monitors the remote site's ability to contact the primary site.

Automatic switchover with AUSO is also compatible with the MCTB. AUSO reacts very quickly because it is designed to detect specific failure events and then invoke the switchover only when NVRAM and SyncMirror plexes are in sync.

In contrast, the tiebreaker is located remotely and therefore must wait for a timer to elapse before declaring a site dead. The tiebreaker eventually detects the sort of controller failure covered by AUSO, but in general AUSO has already started the switchover and possibly completed the switchover before the tiebreaker acts. The resulting second switchover command coming from the tiebreaker would be rejected.

*Caution: *The MCTB software does not verify that NVRAM was and/or plexes are in sync when forcing a switchover. Automatic switchover, if configured, should be disabled during maintenance activities that result in loss of sync for NVRAM or SyncMirror plexes.

Additionally, the MCTB might not address a rolling disaster that leads to the following sequence of events:

1. Connectivity between sites is interrupted for more than 30 seconds.
2. SyncMirror replication times out, and operations continue on the primary site, leaving the remote replica stale.
3. The primary site is lost.The result is the presence of unreplicated changes on the primary site. A switchover might then be undesirable for a number of reasons, including the following:

- Critical data might be present on the primary site, and that data might be eventually recoverable. A switchover that allowed the application to continue operating would effectively discard that critical data.
- An application on the surviving site that was using storage resources on the primary site at the time of site loss might have cached data. A switchover would introduce a stale version of the data that does not match the cache.
- An operating system on the surviving site that was using storage resources on the primary site at the time of site loss might have cached data. A switchover would introduce a stale version of the data that does not match the cache. The safest option is to configure the tiebreaker to send an alert if it detects site failure and then have a person make a decision on whether to force a switchover. Applications and/or operating systems might first need to be shut down to clear any cached data. In addition, the NVFAIL settings can be used to add further protection and help streamline the failover process.

ONTAP Mediator with MetroCluster IP

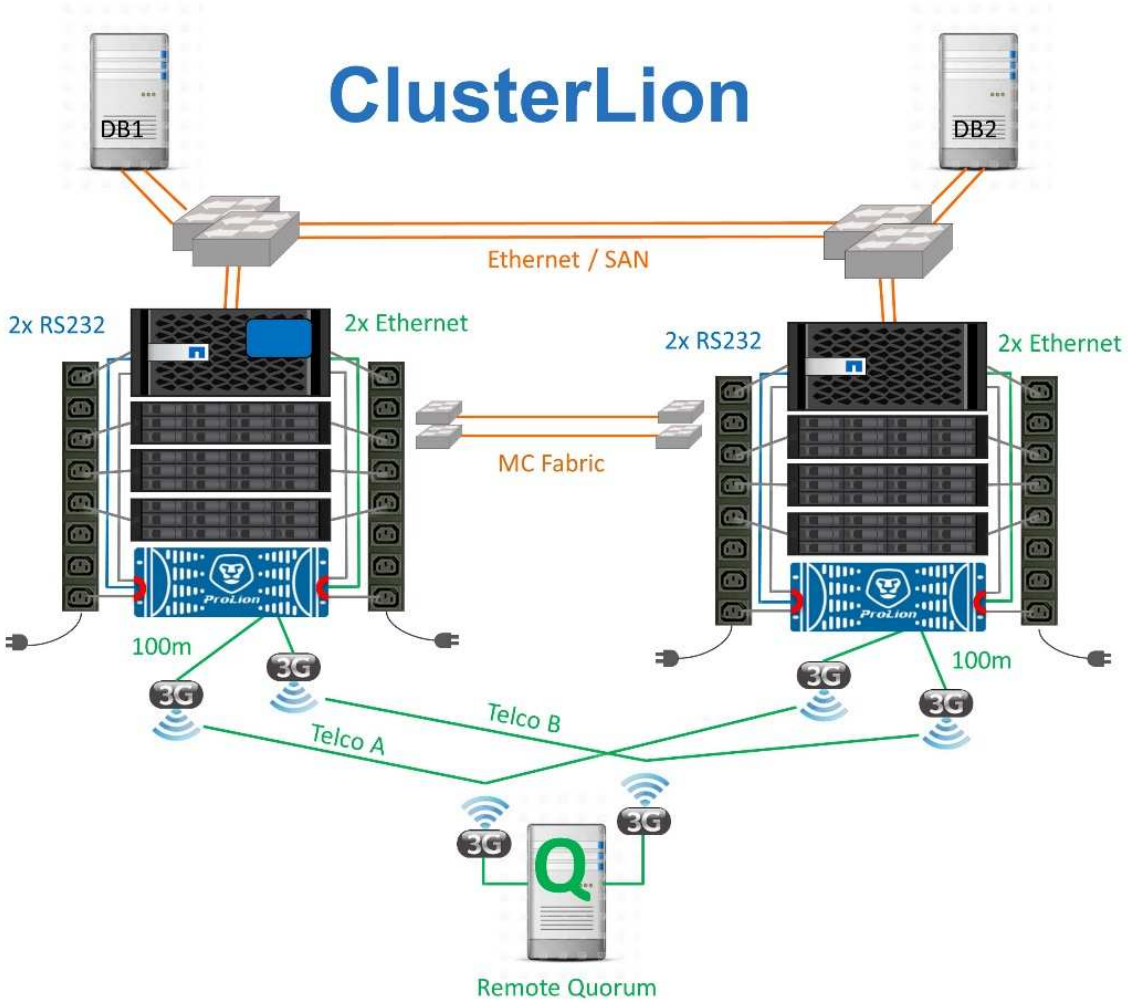
The ONTAP Mediator is used with MetroCluster IP and certain other ONTAP solutions. It functions as a traditional tiebreaker service, much like the MetroCluster Tiebreaker software discussed above, but also includes a critical feature – performing automated unattended switchover.

A fabric-attached MetroCluster has direct access to the storage devices on the opposite site. This allows one MetroCluster controller to monitor the health of the other controllers by reading heartbeat data from the drives. This allows one controller to recognize the failure of another controller and perform a switchover.

In contrast, the MetroCluster IP architecture routes all I/O exclusively through the controller-controller connection; there is no direct access to storage devices on the remote site. This limits the ability of a controller to detect failures and perform a switchover. The ONTAP Mediator is therefore required as a tiebreaker device to detect site loss and automatically perform a switchover.

Virtual third site with ClusterLion

ClusterLion is an advanced MetroCluster monitoring appliance that functions as a virtual third site. This approach allows MetroCluster to be safely deployed in a two-site configuration with fully automated switchover capability. Furthermore, ClusterLion can perform additional network level monitor and execute post-switchover operations. Complete documentation is available from ProLion.



- The ClusterLion appliances monitor the health of the controllers with directly connected Ethernet and serial cables.
- The two appliances are connected to each other with redundant 3G wireless connections.
- Power to the ONTAP controller is routed through internal relays. In the event of a site failure, ClusterLion, which contains an internal UPS system, cuts the power connections before invoking a switchover. This process makes sure that no split-brain condition occurs.
- ClusterLion performs a switchover within the 30-second SyncMirror timeout or not at all.
- ClusterLion does not perform a switchover unless the states of NVRAM and SyncMirror plexes are in sync.
- Because ClusterLion only performs a switchover if MetroCluster is fully in sync, NVFAIL is not required. This configuration permits site-spanning environments such as an extended Oracle RAC to remain online, even during an unplanned switchover.
- Support includes both Fabric-attached MetroCluster and MetroCluster IP

Oracle databases with SyncMirror

The foundation of Oracle data protection with a MetroCluster system is SyncMirror, a maximum-performance, scale-out synchronous mirroring technology.

Data protection with SyncMirror

At the simplest level, synchronous replication means any change must be made to both sides of mirrored storage before it is acknowledged. For example, if a database is writing a log, or a VMware guest is being patched, a write must never be lost. As a protocol level, the storage system must not acknowledge the write until it has been committed to nonvolatile media on both sites. Only then is it safe to proceed without the risk of data loss.

The use of a synchronous replication technology is the first step in designing and managing a synchronous replication solution. The most important consideration is understanding what could happen during various planned and unplanned failure scenarios. Not all synchronous replication solutions offer the same capabilities. If you need a solution that delivers a recovery point objective (RPO) of zero, meaning zero data loss, all failure scenarios must be considered. In particular, what is the expected result when replication is impossible due to loss of connectivity between sites?

SyncMirror data availability

MetroCluster replication is based on NetApp SyncMirror technology, which is designed to efficiently switch into and out of synchronous mode. This capability meets the requirements of customers who demand synchronous replication, but who also need high availability for their data services. For example, if connectivity to a remote site is severed, it is generally preferable to have the storage system continue operating in a non-replicated state.

Many synchronous replication solutions are only capable of operating in synchronous mode. This type of all-or-nothing replication is sometimes called domino mode. Such storage systems stop serving data rather than allowing the local and remote copies of data to become un-synchronized. If replication is forcibly broken, resynchronization can be extremely time consuming and can leave a customer exposed to complete data loss during the time that mirroring is reestablished.

Not only can SyncMirror seamlessly switch out of synchronous mode if the remote site is unreachable, it can also rapidly resync to an RPO = 0 state when connectivity is restored. The stale copy of data at the remote site can also be preserved in a usable state during resynchronization, which ensures that local and remote copies of data exist at all times.

Where domino mode is required, NetApp offers SnapMirror Synchronous (SM-S). Application-level options also exist, such as Oracle DataGuard or extended timeouts for host-side disk mirroring. Consult your NetApp or partner account team for additional information and options.

Oracle database failover with MetroCluster

```
Metrocluster is an ONTAP feature that can protect your Oracle databases
with RPO=0 synchronous mirroring across sites, and it scales up to support
hundreds of databases on a single MetroCluster system. It's also simple to
use. The use of MetroCluster does not necessarily add to or change any
best practices for operating a enterprise applications and databases.
```

The usual best practices still apply, and if your needs only require RPO=0 data protection then that need is met with MetroCluster. However, most customers use MetroCluster not only for RPO=0 data protection, but also to improve RTO during disaster scenarios as well as provide transparent failover as part of site maintenance activities.

Failover with a preconfigured OS

SyncMirror delivers a synchronous copy of the data at the disaster recovery site, but making that data available requires an operating system and the associated applications. Basic automation can dramatically improve the failover time of the overall environment. Clusterware products such as Oracle RAC, Veritas Cluster Server (VCS) or VMware HA are often used to create a cluster across the sites, and in many cases the failover process can be driven with simple scripts.

If the primary nodes are lost, the clusterware (or scripts) is configured to bring the applications online at the alternate site. One option is to create standby servers that are preconfigured for the NFS or SAN resources that make up the application. If the primary site fails, the clusterware or scripted alternative performs a sequence of actions similar to the following:

1. Forcing a MetroCluster switchover
2. Performing discovery of FC LUNs (SAN only)
3. Mounting file systems
4. Starting the application

The primary requirement of this approach is a running OS in place on the remote site. It must be preconfigured with application binaries, which also means that tasks such as patching must be performed on the primary and standby site. Alternatively, the application binaries can be mirrored to the remote site and mounted if a disaster is declared.

The actual activation procedure is simple. Commands such as LUN discovery require just a few commands per FC port. File system mounting is nothing more than a `mount` command, and both databases and ASM can be started and stopped at the CLI with a single command. If the volumes and file systems are not in use at the disaster recovery site prior to the switchover, there is no requirement to set `dr-force- nvfail` on volumes.

Failover with a virtualized OS

Failover of database environments can be extended to include the operating system itself. In theory, this failover can be done with boot LUNs, but most often it is done with a virtualized OS. The procedure is similar to the following steps:

1. Forcing a MetroCluster switchover
2. Mounting the datastores hosting the database server virtual machines
3. Starting the virtual machines
4. Starting databases manually or configuring the virtual machines to automatically start the databases

For example, an ESX cluster could span sites. In the event of disaster, the virtual machines can be brought online at the disaster recovery site after the switchover. As long as the datastores hosting the virtualized database servers are not in use at the time of the disaster, there is no requirement for setting `dr-force-nvfail` on associated volumes.

Oracle databases, MetroCluster, and NVFAIL

NVFAIL is a general data integrity feature in ONTAP that is designed to maximize data integrity protection with databases.

> This section expands on the explanation of basic ONTAP NVFAIL to cover MetroCluster-specific topics.

With MetroCluster, a write is not acknowledged until it has been logged into local NVRAM and NVRAM on at least one other controller. This approach makes sure that a hardware failure or power outage does not result in the loss of in-flight I/O. If the local NVRAM fails or the connectivity to other nodes fails, then data would no longer be mirrored.

If the local NVRAM reports an error, the node shuts down. This shutdown results in failover to a partner controller when HA pairs are used. With MetroCluster, the behavior depends on the overall configuration chosen, but it can result in automatic failover to the remote note. In any case, no data is lost because the controller experiencing the failure has not acknowledged the write operation.

A site-to-site connectivity failure that blocks NVRAM replication to remote nodes is a more complicated situation. Writes are no longer replicated to the remote nodes, creating a possibility of data loss if a catastrophic error occurs on a controller. More importantly, attempting to fail over to a different node during these conditions results in data loss.

The controlling factor is whether NVRAM is synchronized. If NVRAM is synchronized, node-to-node failover is safe to proceed without the risk of data loss. In a MetroCluster configuration, if NVRAM and the underlying aggregate plexes are in sync, it is safe to proceed with the switchover without the risk of data loss.

ONTAP does not permit a failover or switchover when the data is out of sync unless the failover or switchover is forced. Forcing a change in conditions in this manner acknowledges that data might be left behind in the original controller and that data loss is acceptable.

Databases are especially vulnerable to corruption if a failover or switchover is forced because databases maintain larger internal caches of data on disk. If a forced failover or switchover occurs, previously acknowledged changes are effectively discarded. The contents of the storage array effectively jump backward in time, and the state of the database cache no longer reflects the state of the data on disk.

To protect applications from this situation, ONTAP allows volumes to be configured for special protection against NVRAM failure. When triggered, this protection mechanism results in a volume entering a state called NVFAIL. This state results in I/O errors that cause an application shutdown so that they do not use stale data. Data should not be lost because any acknowledged writes are still present on the storage system, and with databases any committed transaction data should be present in the logs.

The usual next steps are for an administrator to fully shut down the hosts before manually placing the LUNs and volumes back online again. Although these steps can involve some work, this approach is the safest way to make sure of data integrity. Not all data requires this protection, which is why NVFAIL behavior can be configured on a volume-by-volume basis.

Manually forced NVFAIL

The safest option to force a switchover with an application cluster (including VMware, Oracle RAC, and others) that is distributed across sites is by specifying `-force-nvfail-all` at the command line. This option is available as an emergency measure to make sure that all cached data is flushed. If a host is using storage resources originally located on the disaster-stricken site, it receives either I/O errors or a stale file handle (`ESTALE`) error. Oracle databases crash and file systems either go offline entirely or switch to read-only mode.

After the switchover is complete, the `in-nvfailed-state` flag needs to be cleared, and the LUNs need to be placed online. After this activity is complete, the database can be restarted. These tasks can be automated to reduce the RTO.

dr-force-nvfail

As a general safety measure, set the `dr-force-nvfail` flag on all volumes that might be accessed from a remote site during normal operations, meaning they are activities used prior to failover. The result of this

setting is that select remote volumes become unavailable when they enter `in-nvfailed-state` during a switchover. After the switchover is complete, the `in-nvfailed-state` flag must be cleared, and the LUNs must be placed online. After these activities are complete, the applications can be restarted. These tasks can be automated to reduce the RTO.

The result is like using the `-force-nvfail-all` flag for manual switchovers. However, the number of volumes affected can be limited to just those volumes that must be protected from applications or operating systems with stale caches.

There are two critical requirements for an environment that does not use `dr-force-nvfail` on application volumes:

- A forced switchover must occur no more than 30 seconds after primary site loss.
- A switchover must not occur during maintenance tasks or any other conditions in which SyncMirror plexes or NVRAM replication are out of sync. The first requirement can be met by using tiebreaker software that is configured to perform a switchover within 30 seconds of a site failure. This requirement does not mean the switchover must be performed within 30 seconds of the detection of a site failure. It does mean that it is no longer safe to force a switchover if 30 seconds have elapsed since a site was confirmed to be operational.

The second requirement can be partially met by disabling all automated switchover capabilities when the MetroCluster configuration is known to be out of sync. A better option is to have a tiebreaker solution that can monitor the health of NVRAM replication and the SyncMirror plexes. If the cluster is not fully synchronized, the tiebreaker should not trigger a switchover.

The NetApp MCTB software cannot monitor the synchronization status, so it should be disabled when MetroCluster is not in sync for any reason. ClusterLion does include NVRAM-monitoring and plex-monitoring capabilities and can be configured to not trigger the switchover unless the MetroCluster system is confirmed to be fully synchronized.

Oracle single-instance on MetroCluster

As stated previously, the presence of a MetroCluster system does not necessarily add to or change any best practices for operating a database. The majority of databases currently running on customer MetroCluster systems are single instance and follow the recommendations in the Oracle on ONTAP documentation.

Failover with a preconfigured OS

SyncMirror delivers a synchronous copy of the data at the disaster recovery site, but making that data available requires an operating system and the associated applications. Basic automation can dramatically improve the failover time of the overall environment. Clusterware products such as Veritas Cluster Server (VCS) are often used to create a cluster across the sites, and in many cases the failover process can be driven with simple scripts.

If the primary nodes are lost, the clusterware (or scripts) is configured to bring the databases online at the alternate site. One option is to create standby servers that are preconfigured for the NFS or SAN resources that make up the database. If the primary site fails, the clusterware or scripted alternative performs a sequence of actions similar to the following:

1. Forcing a MetroCluster switchover
2. Performing discovery of FC LUNs (SAN only)
3. Mounting file systems and/or mounting ASM disk groups

4. Starting the database

The primary requirement of this approach is a running OS in place on the remote site. It must be preconfigured with Oracle binaries, which also means that tasks such as Oracle patching must be performed on the primary and standby site. Alternatively, the Oracle binaries can be mirrored to the remote site and mounted if a disaster is declared.

The actual activation procedure is simple. Commands such as LUN discovery require just a few commands per FC port. File system mounting is nothing more than a `mount` command, and both databases and ASM can be started and stopped at the CLI with a single command. If the volumes and file systems are not in use at the disaster recovery site prior to the switchover, there is no requirement to set `dr-force- nvfail` on volumes.

Failover with a virtualized OS

Failover of database environments can be extended to include the operating system itself. In theory, this failover can be done with boot LUNs, but most often it is done with a virtualized OS. The procedure is similar to the following steps:

1. Forcing a MetroCluster switchover
2. Mounting the datastores hosting the database server virtual machines
3. Starting the virtual machines
4. Starting databases manually or configuring the virtual machines to automatically start the databases For example, an ESX cluster could span sites. In the event of disaster, the virtual machines can be brought online at the disaster recovery site after the switchover. As long as the datastores hosting the virtualized database servers are not in use at the time of the disaster, there is no requirement for setting `dr-force-nvfail` on associated volumes.

Extended Oracle RAC on MetroCluster

Many customers optimize their RTO by stretching an Oracle RAC cluster across sites, yielding a fully active-active configuration. The overall design becomes more complicated because it must include quorum management of Oracle RAC. Additionally, data is accessed from both sites, which means a forced switchover might lead to the use of an out-of-date copy of the data.

Although a copy of the data is present on both sites, only the controller that currently owns an aggregate can serve data. Therefore, with extended RAC clusters, the nodes that are remote must perform I/O across a site-to-site connection. The result is added I/O latency, but this latency is not generally a problem. The RAC interconnect network must also be stretched across sites, which means a high-speed, low-latency network is required anyway. If the added latency does cause a problem, the cluster can be operated in an active-passive manner. I/O-intensive operations would then need to be directed to the RAC nodes that are local to the controller that owns the aggregates. The remote nodes then perform lighter I/O operations or are used purely as warm standby servers.

If active-active extended RAC is required, ASM mirroring should be considered in place of MetroCluster. ASM mirroring allows a specific replica of the data to be preferred. Therefore, a extended RAC cluster can be built in which all reads occur locally. Read I/O never crosses sites, which delivers the lowest possible latency. All write activity must still transit the intersite connection, but such traffic is unavoidable with any synchronous mirroring solution.

> If boot LUNs, including virtualized boot disks, are used with Oracle RAC, the `misscount` parameter might need to be changed. For more information about RAC timeout parameters, see Oracle RAC with ONTAP.

Two-site configuration

A two-site extended RAC configuration can deliver active-active database services that can survive many, but not all, disaster scenarios nondisruptively.

RAC voting files

The first consideration when deploying extended RAC on MetroCluster should be quorum management. Oracle RAC has two mechanisms to manage quorum: disk heartbeat and network heartbeat. The disk heartbeat monitors storage access using the voting files. With a single-site RAC configuration, a single voting resource is sufficient as long as the underlying storage system offers HA capabilities.

In earlier versions of Oracle, the voting files were placed on physical storage devices, but in current versions of Oracle the voting files are stored in ASM diskgroups.

> Oracle RAC is supported with NFS. During the grid installation process, a set of ASM processes is created to present the NFS location used for grid files as an ASM diskgroup. The process is nearly transparent to the end user and requires no ongoing ASM management after the installation is complete.

The first requirement in a two-site configuration is making sure that each site can always access more than half of the voting files in a way that guarantees a nondisruptive disaster recovery process. This task was simple before the voting files were stored in ASM diskgroups, but today administrators need to understand basic principles of ASM redundancy.

ASM diskgroups have three options for redundancy `external`, `normal`, and `high`. In other words, unmirrored, mirrored, and 3-way mirrored. A newer option called `Flex` is also available, but rarely used. The redundancy level and placement of the redundant devices controls what happens in failure scenarios. For example:

- Placing the voting files on a `diskgroup` with `external` redundancy resource guarantees eviction of one site if intersite connectivity is lost.
- Placing the voting files on a `diskgroup` with `normal` redundancy with only one ASM disk per site guarantees node eviction on both sites if intersite connectivity is lost because neither site would have a majority quorum.
- Placing the voting files on a `diskgroup` with `high` redundancy with two disks on one site and a single disk on the other site allows for active-active operations when both sites are operational and mutually reachable. However, if the single-disk site is isolated from the network, then that site is evicted.

RAC network heartbeat

The Oracle RAC network heartbeat monitors node reachability across the cluster interconnect. To remain in the cluster, a node must be able to contact more than half of the other nodes. In a two-site architecture, this requirement creates the following choices for the RAC node count:

- Placement of an equal number of nodes per site results in eviction at one site in the event network connectivity is lost.

- Placement of N nodes on one site and N+1 nodes on the opposite site guarantees that loss of intersite connectivity results in the site with the larger number of nodes remaining in network quorum and the site with fewer nodes evicting.

Prior to Oracle 12cR2, it was not feasible to control which side would experience an eviction during site loss. When each site has an equal number of nodes, eviction is controlled by the master node, which in general is the first RAC node to boot.

Oracle 12cR2 introduces node weighting capability. This capability gives an administrator more control over how Oracle resolves split-brain conditions. As a simple example, the following command sets the preference for a particular node in an RAC:

```
[root@host-a ~]# /grid/bin/crsctl set server css_critical yes
CRS-4416: Server attribute 'CSS_CRITICAL' successfully changed. Restart
Oracle High Availability Services for new value to take effect.
```

After restarting Oracle High-Availability Services, the configuration looks as follows:

```
[root@host-a lib]# /grid/bin/crsctl status server -f | egrep
'^NAME|CSS_CRITICAL='
NAME=host-a
CSS_CRITICAL=yes
NAME=host-b
CSS_CRITICAL=no
```

Node `host-a` is now designated as the critical server. If the two RAC nodes are isolated, `host-a` survives, and `host-b` is evicted.

> For complete details, see the Oracle white paper “Oracle Clusterware 12c Release 2 Technical Overview. ”

For versions of Oracle RAC prior to 12cR2, the master node can be identified by checking the CRS logs as follows:

```
[root@host-a ~]# /grid/bin/crsctl status server -f | egrep
'^NAME|CSS_CRITICAL='
NAME=host-a
CSS_CRITICAL=yes
NAME=host-b
CSS_CRITICAL=no
 [root@host-a ~]# grep -i 'master node' /grid/diag/crs/host-
a/crs/trace/crsd.trc
2017-05-04 04:46:12.261525 :   CRSSE:2130671360: {1:16377:2} Master Change
Event; New Master Node ID:1 This Node's ID:1
2017-05-04 05:01:24.979716 :   CRSSE:2031576832: {1:13237:2} Master Change
Event; New Master Node ID:2 This Node's ID:1
2017-05-04 05:11:22.995707 :   CRSSE:2031576832: {1:13237:221} Master
Change Event; New Master Node ID:1 This Node's ID:1
2017-05-04 05:28:25.797860 :   CRSSE:3336529664: {1:8557:2} Master Change
Event; New Master Node ID:2 This Node's ID:1
```

This log indicates that the master node is `2` and the node `host-a` has an ID of `1`. This fact means that `host-a` is not the master node. The identity of the master node can be confirmed with the command `olsnodes -n`.

```
[root@host-a ~]# /grid/bin/olsnodes -n
host-a  1
host-b  2
```

The node with an ID of `2` is `host-b`, which is the master node. In a configuration with equal numbers of nodes on each site, the site with `host-b` is the site that survives if the two sets lose network connectivity for any reason.

It is possible that the log entry that identifies the master node can age out of the system. In this situation, the timestamps of the Oracle Cluster Registry (OCR) backups can be used.

```
[root@host-a ~]#  /grid/bin/ocrconfig -showbackup
host-b     2017/05/05 05:39:53     /grid/cdata/host-cluster/backup00.ocr
0
host-b     2017/05/05 01:39:53     /grid/cdata/host-cluster/backup01.ocr
0
host-b     2017/05/04 21:39:52     /grid/cdata/host-cluster/backup02.ocr
0
host-a     2017/05/04 02:05:36     /grid/cdata/host-cluster/day.ocr     0
host-a     2017/04/22 02:05:17     /grid/cdata/host-cluster/week.ocr     0
```

This example shows that the master node is `host-b`. It also indicates a change in the master node from `host-a` to `host-b` somewhere between 2:05 and 21:39 on May 4. This method of identifying the master node is only safe to use if the CRS logs have also been checked because it is possible that the master node has

changed since the previous OCR backup. If this change has occurred, then it should be visible in the OCR logs.

Most customers choose a single voting diskgroup that services the entire environment and an equal number of RAC nodes on each site. The diskgroup should be placed on the site that that contains the database. The result is that loss of connectivity results in eviction on the remote site. The remote site would no longer have quorum, nor would it have access to the database files, but the local site continues running as usual. When connectivity is restored, the remote instance can be brought online again.

In the event of disaster, a switchover is required to bring the database files and voting diskgroup online on the surviving site. If the disaster allows AUSO to trigger the switchover, NVFAIL is not triggered because the cluster is known to be in sync, and the storage resources come online normally. AUSO is a very fast operation and should complete before the `disktimeout` period expires.

Because there are only two sites, it is not feasible to use any type of automated external tiebreaking software, which means forced switchover must be a manual operation.

Three-site configurations

An extended RAC cluster is much easier to architect with three sites. The two sites hosting each half of the MetroCluster system also support the database workloads, while the third site serves as a tiebreaker for both the database and the MetroCluster system. The Oracle tiebreaker configuration may be as simple as placing a member of the ASM diskgroup used for voting on a 3rd site, and may also include an operational instance on the 3rd site to ensure there is an odd number of nodes in the RAC cluster.

> Consult the Oracle documentation on “quorum failure group” for important information on using NFS in an extended RAC configuration. In summary, the NFS mount options may need to be modified to include the soft option to ensure that loss of connectivity to the 3rd site hosting quorum resources does not hang the primary Oracle servers or Oracle RAC processes.

SnapMirror active sync

Oracle databases with SnapMirror active sync

SnapMirror active sync enables selective RPO=0 synchronous mirroring for individual Oracle databases and application environments.

SnapMirror active sync is essentially an enhanced SnapMirror capability for SAN that allows hosts to access a LUN from both the system hosting the LUN as well as the system hosting its replica.

SnapMirror active sync and SnapMirror Sync share a replication engine, however, SnapMirror active sync includes additional features such as transparent application failover and failback for enterprise applications.

In practice, it works similar to a granular version of MetroCluster by enabling selective and granular RPO=0 synchronous replication for individual workloads. The low-level path behavior is very different from MetroCluster, but the end result from a host point of view is similar.

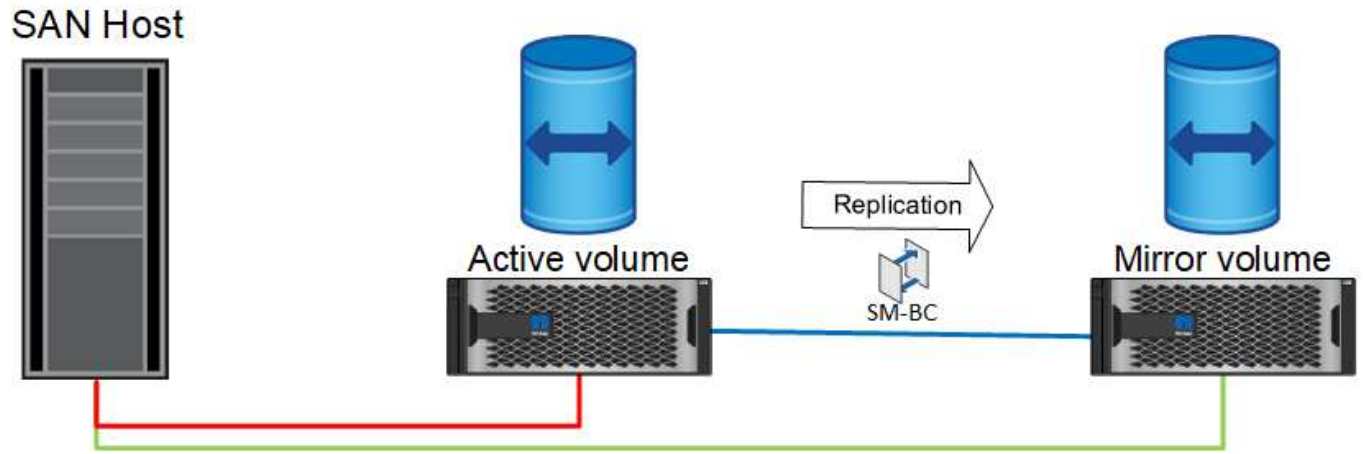
Path access

with SnapMirror active sync makes storage devices visible to host operating systems from both the primary and remote storage arrays. Paths are managed through asymmetric logical unit access (ALUA), which is an industry standard protocol for identifying optimized paths between a storage system and a host.

The device path that is the shortest to access I/O is considered Active/Optimized paths and the rest of the

paths are considered Active/Nonoptimized paths.

The SnapMirror active sync relationship is between a pair of SVMs located on different clusters. Both SVMs are capable of serving data, but ALUA will preferentially use the SVM that currently has ownership of the drives on which the LUNs reside. IO to the remote SVM will be proxied across the with SnapMirror active sync interconnect.



Synchronous replication

In normal operation, the remote copy is an RPO=0 synchronous replica at all times, with one exception. If data cannot be replicated, with SnapMirror active sync will release the requirement to replicate data and resume serving IO. This option is preferred by customers who consider loss of the replication link a near-disaster, or who do not want business operations to halt when data cannot be replicated.

Storage hardware

Unlike other storage disaster recovery solutions, SnapMirror active sync offers asymmetric platform flexibility. The hardware at each site does not need to be identical. This capability allows you to right-size the hardware used to support SnapMirror active sync. The remote storage system can be identical to the primary site if it needs to support a full production workload, but if a disaster results in reduced I/O, than a smaller system at the remote site might be more cost-effective.

ONTAP mediator

The ONTAP Mediator is a software application that is downloaded from NetApp support. The Mediator automates failover operations for both the primary and remote site storage cluster. It can be deployed on a small virtual machine (VM) hosted either on-premises or in the cloud. After it is configured, it acts as a third site to monitor failover scenarios for both the sites.

Oracle database failover with SnapMirror active sync

The primary reason for hosting an Oracle database on SnapMirror active sync is to deliver transparent failover during planned and unplanned storage events.

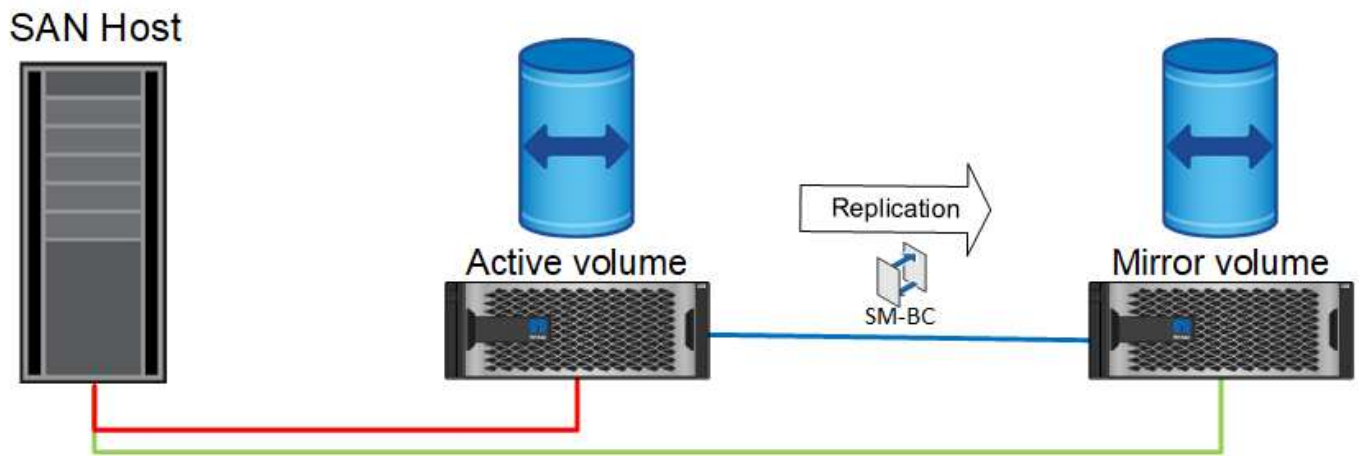
SnapMirror active sync supports two types of storage failover operations: planned and unplanned, which work in slightly different ways. A planned failover is initiated manually by the administrator for quick switchover to a remote site whereas unplanned failover is initiated automatically by the mediator on the third site. The primary purpose of a planned failover is to perform incremental patching and upgrades, perform disaster recovery testing, or adopt a formal policy of switching operations between sites throughout the year to prove full active sync capability.

The diagrams show what happens during normal, failover, and failback operations. For ease of illustration, they depict a replicated LUN. In an actual SnapMirror active sync configuration, the replication is based on volumes, where each volume contains one or more LUNs, but to make the picture simpler, the volume layer has been removed.

Normal operation

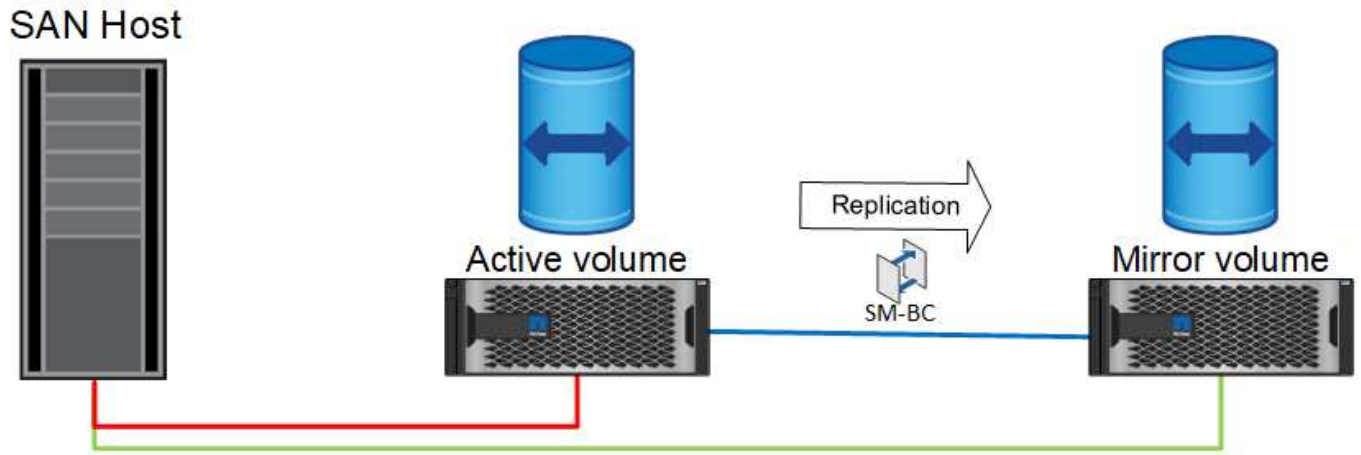
In normal operation a LUN can be accessed from either the local or remote replica. The red line indicates the optimized path as advertised by ALUA, and the result should be that IO is preferentially sent down this path.

The green line is an active path, but it would incur more latency because IO on that path would need to be passed across the SnapMirror active sync path. The additional latency would depend on the speed of the interconnect between sites that is used for SnapMirror active sync.



Failure

If the active mirror copy becomes unavailable, either because of planned or unplanned failover, it obviously will no longer be usable. However, the remote system possesses a synchronous replica and SAN paths to the remote site already exist. The remote system is able to service IO for that LUN.

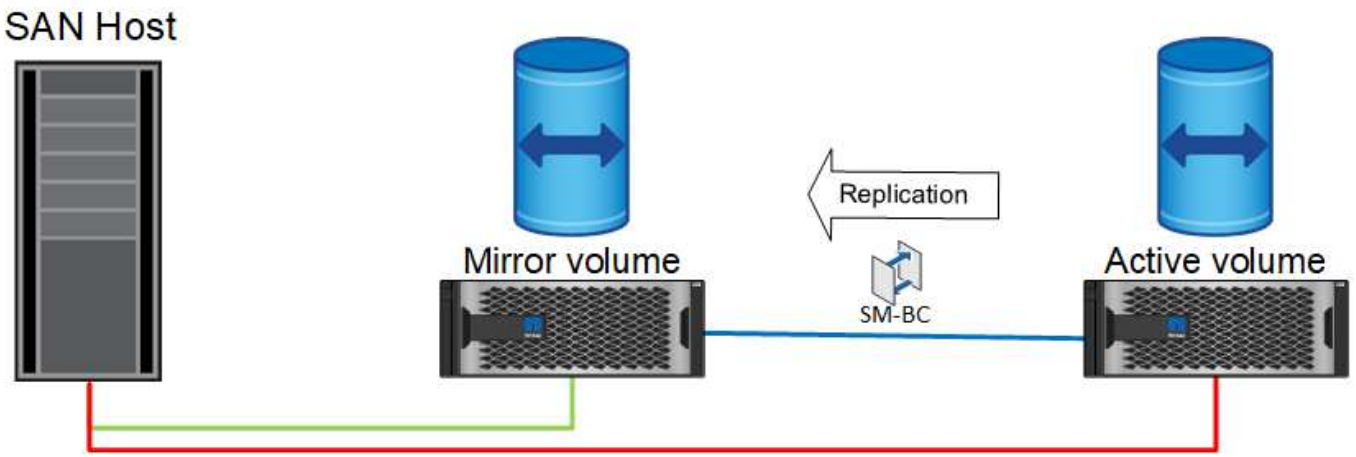


Failover

Failover results in the remote copy becoming the active copy. The paths are changed from Active to Active/Optimized and IO continues to be serviced without data loss.

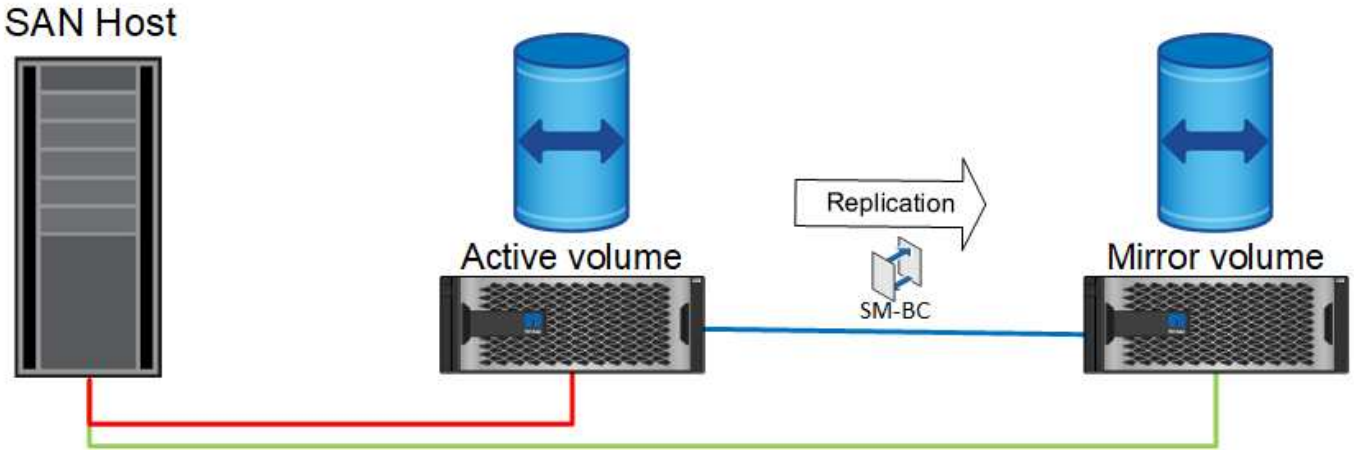
Repair

Once the source system is returned to service, SnapMirror active sync can resync replication but running the other direction. The configuration now is essentially the same as the starting point, except the active-mirror sites have been flipped.



Failback

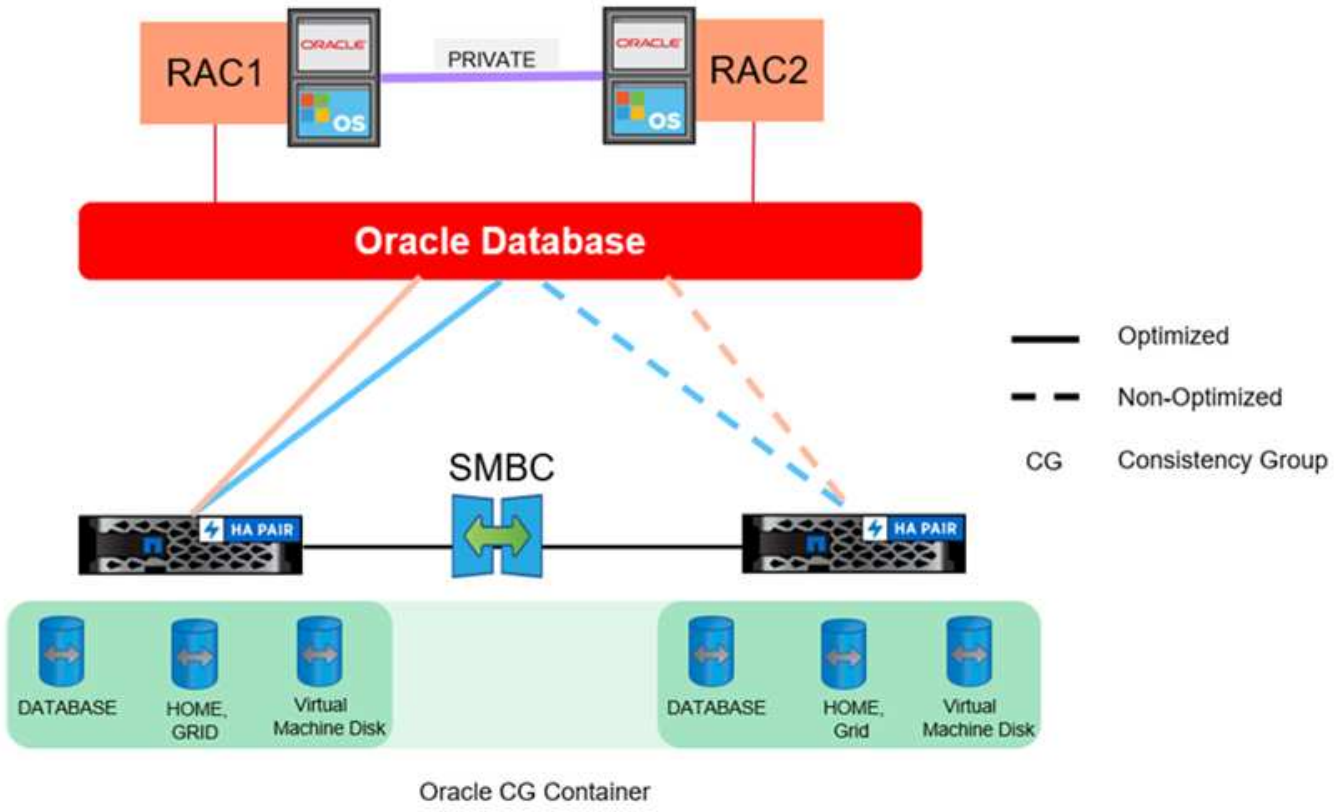
If desired, an administrator can perform a failback and move the active copy of the LUN(s) back to the original controllers.



Single-Instance Oracle databases with SnapMirror active sync

The diagram below shows a simple deployment model where you have storage devices being zoned or connected from both the primary and remote storage clusters for an Oracle database.

Oracle is configured on the primary only. This model addresses seamless storage failover in the event of storage side disasters providing no loss of data without any application downtime. This model would not, however, provide high availability of the database environment during a site failure. This type of architecture is useful for customers looking for a zero data loss solution with high availability of the storage services but accept that a total loss of the database cluster would require manual work.



This approach also saves money on Oracle licensing costs. Preconfiguration of Oracle database nodes on the remote site would require that all cores be licensed under most Oracle licensing agreements. If the delay caused by the time required to install an Oracle database server and mount the surviving copy of data is acceptable, this design can be very cost effective.

Oracle RAC with SnapMirror active sync

SnapMirror active sync delivers granular control over dataset replication for purposes such as load balancing or individual application failover. The overall architecture looks like an extended RAC cluster, but some databases are dedicated to specific sites and the overall load is distributed.

For example, you could build an Oracle RAC cluster hosting six individual databases. The storage for three of the databases would be primarily hosted on site A, and storage for the other three databases would be hosted on site B. This configuration ensures the best possible performance by minimizing cross-site traffic. In addition, applications would be configured to use the database instances that are local to the storage system with active

paths. This minimizes RAC interconnect traffic. Finally, this overall design ensures that all compute resources are used evenly. As workloads change, databases can be selectively failed back and forth across sites to ensure even loading.

Other than granularity, the basic principles and options for Oracle RAC using SnapMirror active syncare the same as Oracle RAC on MetroCluster

Oracle databases and SnapMirror active sync failure scenarios

There are multiple SnapMirror active sync (SM-AS) failure scenarios each having different results.

Scenario	Result
Replication link failure	Mediator recognizes this split-brain scenario and resumes I/O on the node that holds the master copy. When the connectivity between sites is back online, the alternate site performs automatic resync.
Primary site storage failure	Automated unplanned failover is initiated by Mediator. No I/O disruption.
Remote site storage failure	There is no I/O disruption. There is a momentary pause due to the network causing sync replication to abort and the master establishing that it is the rightful owner to continue to serve I/O (consensus). Therefore, there is an I/O pause of a few seconds and then the I/O will resume. There is an automatic resync when the site is online.
Loss of Mediator or link between Mediator and the storage arrays	I/O continues and remains in sync with the remote cluster but automated unplanned/planned failover and failback is not possible in the absence of Mediator.
Loss of one of the storage controllers in the HA cluster	The partner node in the HA cluster attempts a takeover (NDO). If takeover fails, Mediator notices that both the node in the storage is down and performs an automatic unplanned failover to the remote cluster.
Loss of disks	IO continues for up to three consecutive disk failures. This is part of RAID-TEC.
Loss of the entire site in a typical deployment	Servers on the failed site will obviously no longer be available. Applications that support clustering can be configured to run at both sites and continue operations on the alternate site, although most such applications require a 3rd site tiebreaker similar to how SM-AS requires the mediator. Without application level clusters, applications will need to be started at the surviving site. This would affect availbility, but RPO=0 is preserved. No data would be lost.

Oracle database migration

Migration of Oracle databases to ONTAP storage systems

Leveraging the capabilities of a new storage platform has one unavoidable requirement; data must be placed on the new storage system. ONTAP makes the migration process simple, including both ONTAP to ONTAP migrations and upgrades, foreign LUN imports, and procedures for using the host operating system or Oracle database software directly.

> This documentation replaces previously published technical report *TR-4534: Migration of Oracle Databases to NetApp Storage Systems*

In the case of a new database project, this is not a concern because the database and application environments are constructed in place. Migration, however, poses special challenges regarding business disruption, the time required for the completion of migration, needed skill sets, and risk minimization.

Scripts

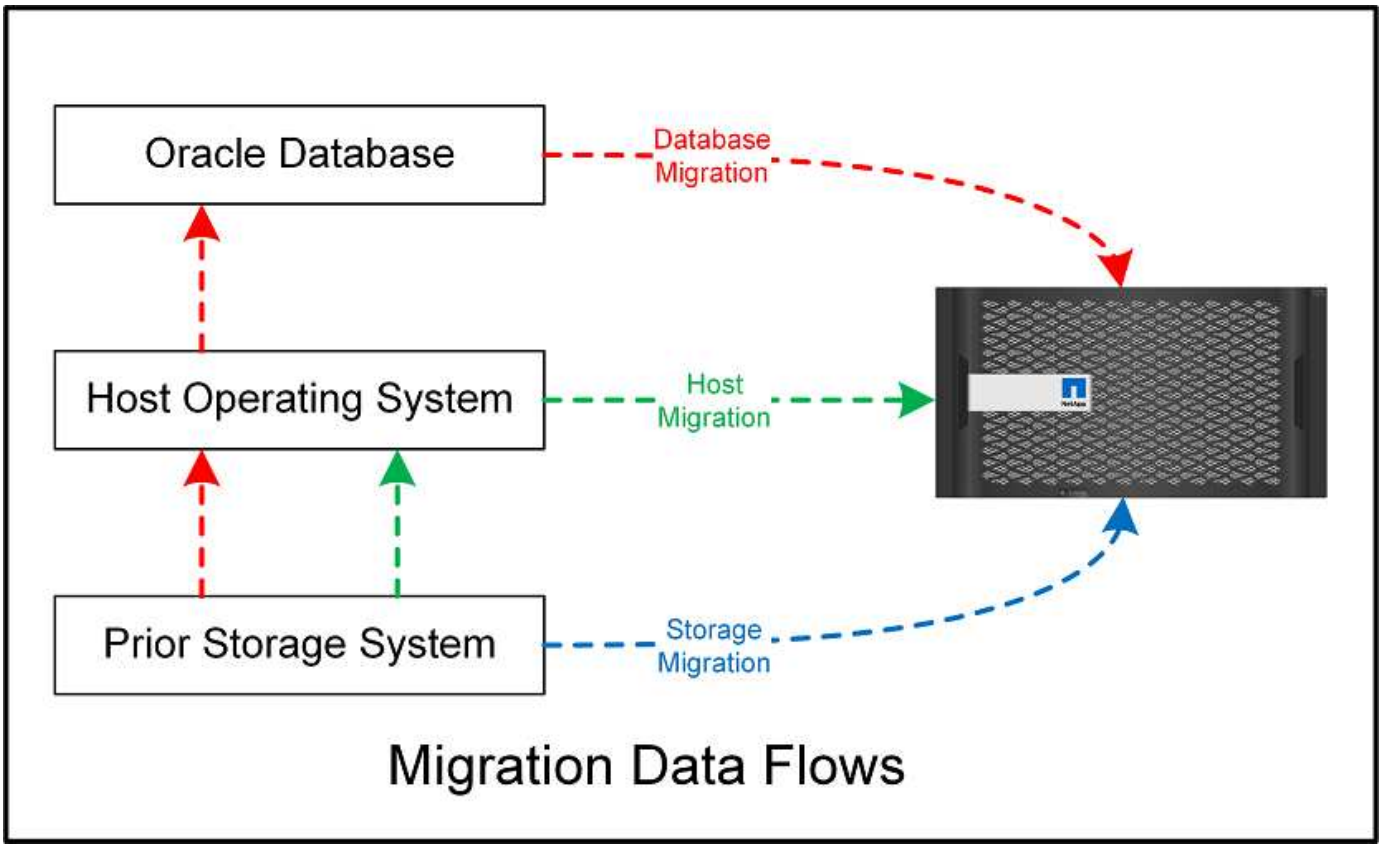
Sample scripts are provided in this documentation. These scripts provide sample methods of automating various aspects of migration to reduce the chance of user errors. The scripts can reduce the overall demands on the IT staff responsible for a migration and they can speed up the overall process. These scripts are all drawn from actual migration projects performed by NetApp Professional Services and NetApp partners. Examples of their use are shown throughout this documentation.

Oracle database migration planning

Oracle data migration can occur at one of three levels: the database, the host, or the storage array.

The differences lie in which component of the overall solution is responsible for moving data: the database, the host operating system, or the storage system.

The figure below shows an example of the migration levels and the flow of data. In the case of database-level migration, the data is moved from the original storage system through the host and database layers into the new environment. Host-level migration is similar, but data does not pass through the application layer and is instead written to the new location by using host processes. Finally, with storage-level migration, an array such as a NetApp FAS system is responsible for data movement.



A database-level migration generally refers to the use of Oracle log shipping through a standby database to complete a migration at the Oracle layer. Host-level migrations are performed by using the native capability of the host operating system configuration. This configuration includes file copy operations using commands such as cp, tar, and Oracle Recovery Manager (RMAN) or using a logical volume manager (LVM) to relocate the underlying bytes of a file system. Oracle Automatic Storage Management (ASM) is categorized as a host- level capability because it runs below the level of the database application. ASM takes the place of the usual logical volume manager on a host. Finally, data can be migrated at the storage- array level, which means beneath the level of the operating system.

Planning considerations

The best option for migration depends on a combination of factors, including the scale of the environment to be migrated, the need to avoid downtime, and the overall effort required to perform the migration. Large databases obviously require more time and effort for migration, but the complexity of such a migration is minimal. Small databases can be migrated quickly, but, if there are thousands to be migrated, the scale of the effort can create complications. Finally, the larger the database, the more likely it is to be business-critical, which gives rise to a need to minimize downtime while preserving a back- out path.

Some of the considerations for planning a migration strategy are discussed here.

Data size

The sizes of the databases to be migrated obviously affect migration planning, although size does not necessarily affect the cutover time. When a large amount of data must be migrated, the primary consideration is bandwidth. Copy operations are usually performed with efficient sequential I/O. As a conservative estimate, assume 50% utilization of the available network bandwidth for copy operations. For example, an 8GB FC port can transfer about 800MBps in theory. Assuming 50% utilization, a database can be copied at a rate of about 400MBps. Therefore, a 10TB database can be copied in about seven hours at this rate.

Migration over longer distances usually requires a more creative approach, such as the log shipping process explained in Online datafile move. Long-distance IP networks rarely have bandwidth anywhere close to LAN or SAN speeds. In one case, NetApp assisted with the long-distance migration of a 220TB database with very high archive- log generation rates. The chosen approach for data transfer was daily shipment of tapes, because this method offered the maximum possible bandwidth.

Database count

In many cases, the problem with moving a large amount of data is not the data size, but rather it is the complexity of the configuration that supports the database. Simply knowing that 50TB of databases must be migrated is not sufficient information. It could be a single 50TB mission-critical database, a collection of 4, 000 legacy databases, or a mix of production and nonproduction data. In some cases, much of the data consists of clones of a source database. These clones do not need to be migrated at all because they can be easily recreated, especially when the new architecture is designed to leverage NetApp FlexClone volumes.

For migration planning, you must understand how many databases are in scope and how they must be prioritized. As the number of databases increases, the preferred migration option tends to be lower and lower in the stack. For example, copying a single database might be easily performed with RMAN and a short outage. This is host-level replication.

If there are 50 databases, it might be easier to avoid setting up a new file system structure to receive an RMAN copy and instead move the data in place. This process can be done by leveraging host-based LVM migration to relocate data from old LUNs to new LUNs. Doing so moves responsibility from the database administrator (DBA) team to the OS team, and, as a result, data is migrated transparently with respect to the database. The file system configuration is unchanged.

Finally, if 500 databases across 200 servers must be migrated, storage-based options such as the ONTAP Foreign LUN Import (FLI) capability can be used to perform a direct migration of the LUNs.

Rearchitecture requirements

Typically, a database file layout must be altered to leverage the features of the new storage array; however, this is not always the case. For example, the features of EF-Series all-flash arrays are directed primarily at SAN performance and SAN reliability. In most cases, databases can be migrated to an EF-Series array with no special considerations for data layout. The only requirements are high IOPS, low latency, and robust reliability. Although there are best practices relating to such factors as RAID configuration or Dynamic Disk Pools, EF-Series projects rarely require any significant changes to the overall storage architecture to leverage such features.

In contrast, migration to ONTAP generally requires more consideration of the database layout to make sure that the final configuration delivers maximum value. By itself, ONTAP offers many features for a database environment, even without any specific architecture effort. Most importantly, it delivers the ability to nondisruptively migrate to new hardware when the current hardware reaches its end of life. Generally speaking, a migration to ONTAP is the last migration that you would need to perform. Subsequent hardware is upgraded in place and data is nondisruptively migrated to new media.

With some planning, even more benefits are available. The most important considerations surround the use of snapshots. Snapshots are the basis for performing near-instantaneous backups, restores, and cloning operations. As an example of the power of snapshots, the largest known use is with a single database of 996TB running on about 250 LUNs on 6 controllers. This database can be backed up in 2 minutes, restored in 2 minutes, and cloned in 15 minutes. Additional benefits include the ability to move data around the cluster in response to changes in workload and the application of quality of service (QoS) controls to provide good, consistent performance in a multidatabase environment.

Technologies such as QoS controls, data relocation, snapshots, and cloning work in nearly any configuration.

However, some thought is generally required to maximize benefits. In some cases, database storage layouts can require design changes to maximize the investment in the new storage array. Such design changes can affect the migration strategy because host-based or storage-based migrations replicate the original data layout. Additional steps might be required to complete the migration and deliver a data layout optimized for ONTAP. The procedures shown in Oracle migration procedures overview and later demonstrate some of the methods to not just migrate a database, but to migrate it into the optimal final layout with minimal effort.

Cutover time

The maximum allowable service outage during cutover should be determined. It is a common mistake to assume that the entire migration process causes disruption. Many tasks can be completed before any service interruption begins, and many options enable the completion of migration without disruption or outage. Even when disruption is unavoidable, you must still define the maximum allowable service outage because the duration of the cutover time varies from procedure to procedure.

For example, copying a 10TB database typically requires approximately seven hours to complete. If business needs allow a seven- hour outage, file copying is an easy and safe option for migration. If five hours is unacceptable, a simple log- shipping process (see Oracle log shipping) can be set up with minimal effort to reduce the cutover time to approximately 15 minutes. During this time, a database administrator can complete the process. If 15 minutes is unacceptable, the final cutover process can be automated through scripting to reduce the cutover time to just a few minutes. You can always speed up a migration, but doing so comes at the cost of time and effort. The cutover time targets should be based on what is acceptable to the business.

Back-out path

No migration is completely risk free. Even if technology operates perfectly, there is always a possibility of user error. The risk associated with a chosen migration path must be considered alongside the consequences of a failed migration. For example, the transparent online storage migration capability of Oracle ASM is one of its key features, and this method is one of the most reliable known. However, data is being irreversibly copied with this method. In the highly unlikely event that a problem occurs with ASM, there is no easy back- out path. The only option is to either restore the original environment or use ASM to reverse the migration back to the original LUNs. The risk can be minimized, but not eliminated, by performing a snapshot-type backup on the original storage system, assuming the system is capable of performing such an operation.

Rehearsal

Some migration procedures must be fully verified before execution. A need for migration and rehearsal of the cutover process is a common request with mission-critical databases for which migration must be successful and downtime must be minimized. In addition, user- acceptance tests are frequently included as part of the postmigration work, and the overall system can be returned to production only after these tests are complete.

If there is a need for rehearsal, several ONTAP capabilities can make the process much easier. In particular, snapshots can reset a test environment and quickly create multiple space-efficient copies of a database environment.

Procedures

Oracle migration procedures overview

Many procedures are available for Oracle migration database. The right one depends on your business needs.

In many cases, system administrators and DBAs have their own preferred methods of relocating physical volume data, mirroring and demirroring, or leveraging Oracle RMAN to copy data.

These procedures are provided primarily as guidance for IT staff less familiar with some of the available options. In addition, the procedures illustrate the tasks, time requirements, and skillset demands for each migration approach. This allows other parties such as NetApp and partner professional services or IT management to more fully appreciate the requirements for each procedure.

There is no single best practice for creating a migration strategy. Creating a plan requires first understanding the availability options and then selecting the method that best suits the needs of the business. The figure below illustrates the basic considerations and typical conclusions made by customers, but it is not universally applicable to all situations.

For example, one step raises the issue of the total database size. The next step depends on whether the database is more or less than 1TB. The recommended steps are just that—recommendations based on typical customer practices. Most customers would not use DataGuard to copy a small database, but some might. Most customers would not attempt to copy a 50TB database because of the time required, but some might have a sufficiently large maintenance window to permit such an operation.

You can find a flowchart of the types of considerations on which migration path is best here.

Online datafile move

Oracle 12cR1 and higher include the ability to move a datafile while the database remains online. It furthermore works between different filesystem types. For example, a datafile can be relocated from an xfs filesystem to ASM. This method is not generally used at scale because of the number of individual datafile move operations that would be required, but it is an option worth considering with smaller databases with fewer datafiles.

In addition, simply moving a datafile is a good option for migrating parts of existing databases. For example, less-active datafiles could be relocated to more cost-efficient storage, such as a FabricPool volume which can store idle blocks in Object Store.

Database-level migration

Migration at the database level means allowing the database to relocate data. Specifically, this means log shipping. Technologies such as RMAN and ASM are Oracle products, but, for the purposes of migration, they operate at the host level where they copy files and manage volumes.

Log shipping

The foundation for database- level migration is the Oracle archive log, which contains a log of changes to the database. Most of the time, an archive log is part of a backup and recovery strategy. The recovery process begins with the restoration of a database and then the replaying of one or more archive logs to bring the database to the desired state. This same basic technology can be used to perform a migration with little to no interruption of operations. More importantly, this technology enables migration while leaving the original database untouched, preserving a back- out path.

The migration process begins with restoration of a database backup to a secondary server. You can do so in a variety of ways, but most customers use their normal backup application to restore the data files. After the data files are restored, users establish a method for log shipping. The goal is to create a constant feed of archive logs generated by the primary database and replay them on the restored database to keep them both close to the same state. When the cutover time arrives, the source database is completely shut down and the final archive logs, and in some cases the redo logs, are copied over and replayed. It is critical that the redo logs are also considered because they might contain some of the final transactions committed.

After these logs have been transferred and replayed, both databases are consistent with one another. At this point, most customers perform some basic testing. If any errors are made during the migration process, then

the log replay should report errors and fail. It is still advisable to perform some quick tests based on known queries or application-driven activities to verify that the configuration is optimal. It is also a common practice to create one final test table before shutting down the original database to verify whether it is present in the migrated database. This step makes sure that no errors were made during the final log synchronization.

A simple log- shipping migration can be configured out of band with respect to the original database, which makes it particularly useful for mission-critical databases. No configuration changes are required for the source database, and the restoration and initial configuration of the migration environment have no effect on production operations. After log shipping is configured, it places some I/O demands on the production servers. However, log shipping consists of simple sequential reads of the archive logs, which is unlikely to have any effect on production database performance.

Log shipping has proven to be particularly useful for long-distance, high- change-rate migration projects. In one instance, a single 220TB database was migrated to a new location approximately 500 miles away. The change rate was extremely high and security restrictions prevented the use of a network connection. Log shipping was performed by using tape and courier. A copy of the source database was initially restored by using procedures outlined below. The logs were then shipped on a weekly basis by courier until the time of cutover when the final set of tapes was delivered and the logs were applied to the replica database.

Oracle DataGuard

In some cases, a complete DataGuard environment is warranted. It is incorrect to use the term DataGuard to refer to any log shipping or standby database configuration. Oracle DataGuard is a comprehensive framework for managing database replication, but it is not a replication technology. The primary benefit of a complete DataGuard environment in a migration effort is the transparent switchover from one database to another. DataGuard also enables a transparent switchover back to the original database if a problem is discovered, such as a performance or network connectivity issue with the new environment. A fully configured DataGuard environment requires configuration of not only the database layer but also applications so that applications are able to detect a change in the primary database location. In general, it is not necessary to use DataGuard to complete a migration, but some customers have extensive DataGuard expertise in-house and already rely on it for migration work.

Rearchitecture

As discussed before, leveraging the advanced features of storage arrays sometimes requires changing the database layout. Furthermore, a change in storage protocol such as moving from ASM to an NFS file system necessarily alters the file system layout.

One of the principal advantages of log shipping methods, including DataGuard, is that the replication destination does not have to match the source. There are no issues with using a log-shipping approach to migrate from ASM to a regular file system or vice versa. The precise layout of data files can be changed at the destination to optimize the use of Pluggable Database (PDB) technology or to set QoS controls selectively on certain files. In other words, a migration process based on log shipping allows you to optimize the database storage layout easily and safely.

Server resources

One limitation to database-level migration is the need for a second server. There are two ways this second server can be used:

1. You can use the second server as a permanent new home for the database.
2. You can use the second server as a temporary staging server. After data migration to the new storage array is complete and tested, the LUN or NFS file systems are disconnected from the staging server and reconnected to the original server.

The first option is the easiest, but using it might not be feasible in very large environments requiring very powerful servers. The second option requires extra work to relocate the file systems back to the original location. This can be a simple operation in which NFS is used as the storage protocol because the file systems can be unmounted from the staging server and remounted on the original server.

Block-based file systems require extra work to update FC zoning or iSCSI initiators. With most logical volume managers (including ASM), the LUNs are automatically detected and brought online after they are made available on the original server. However, some file system and LVM implementations might require more work to export and import the data. The precise procedure might vary, but it is generally easy to establish a simple, repeatable procedure to complete the migration and rehome the data on the original server.

Although it is possible to set up log shipping and replicate a database within a single server environment, the new instance must have a different process SID to replay the logs. It is possible to temporarily bring up the database under a different set of process IDs with a different SID and change it later. However, doing so can lead to a lot of complicated management activities, and it puts the database environment at risk of user error.

Host-level migration

Migrating data at the host level means using the host operating system and associated utilities to complete the migration. This process includes any utility that copies data, including Oracle RMAN and Oracle ASM.

Data copying

The value of a simple copy operation should not be underestimated. Modern network infrastructures can move data at rates measured in gigabytes per second, and file copy operations are based on efficient sequential read and write I/O. More disruption is unavoidable with a host copy operation when compared to log shipping, but a migration is more than just the data movement. It generally includes changes to networking, the database restart time, and postmigration testing.

The actual time required to copy data might not be significant. Furthermore, a copy operation preserves a guaranteed back- out path because the original data remains untouched. If any problems are encountered during the migration process, the original file systems with the original data can be reactivated.

Replatforming

Replatforming refers to a change in the CPU type. When a database is migrated from a traditional Solaris, AIX, or HP-UX platform to x86 Linux, the data must be reformatted because of changes in the CPU architecture. SPARC, IA64, and POWER CPUs are known as big endian processors, while the x86 and x86_64 architectures are known as little endian. As a result, some data within Oracle data files is ordered differently depending on the processor in use.

Traditionally, customers have used DataPump to replicate data across platforms. DataPump is a utility that creates a special type of logical data export that can be more rapidly imported at the destination database. Because it creates a logical copy of the data, DataPump leaves the dependencies of processor endianness behind. DataPump is still used by some customers for replatforming, but a faster option has become available with Oracle 11g: cross-platform transportable tablespaces. This advance allows a tablespace to be converted to a different endian format in place. This is a physical transformation that offers better performance than a DataPump export, which must convert physical bytes to logical data and then convert back to physical bytes.

A complete discussion of DataPump and transportable tablespaces is beyond the scope NetApp documentation, but NetApp has some recommendations based on our experience assisting customers during migration to a new storage array log with a new CPU architecture:

- If DataPump is being used, the time required to complete the migration should be measured in a test environment. Customers are sometimes surprised at the time required to complete the migration. This

unexpected additional downtime can cause disruption.

- Many customers mistakenly believe that cross-platform transportable tablespaces do not require data conversion. When a CPU with a different endian is used, an RMAN `convert` operation must be performed on the data files beforehand. This is not an instantaneous operation. In some cases, the conversion process can be sped up by having multiple threads operating on different data files, but the conversion process cannot be avoided.

Logical volume manager-driven migration

LVMs work by taking a group of one or more LUNs and breaking them into small units generally referred to as extents. The pool of extents is then used as a source to create logical volumes that are essentially virtualized. This virtualization layer delivers value in various ways:

- Logical volumes can use extents drawn from multiple LUNs. When a file system is created on a logical volume, it can use the full performance capabilities of all LUNs. It also promotes the even loading of all LUNs in the volume group, delivering more predictable performance.
- Logical volumes can be resized by adding and, in some cases, removing extents. Resizing a file system on a logical volume is generally nondisruptive.
- Logical volumes can be nondisruptively migrated by moving the underlying extents.

Migration using an LVM works in one of two ways: moving an extent or mirroring/demirroring an extent. LVM migration uses efficient large-block sequential I/O and only rarely creates any performance concerns. If this does become an issue, there are usually options for throttling the I/O rate. Doing so increases the time required to complete the migration and yet reduces the I/O burden on the host and storage systems.

Mirror and demirror

Some volume managers, such as AIX LVM, allow the user to specify the number of copies for each extent and to control which devices host each copy. Migration is accomplished by taking an existing logical volume, mirroring the underlying extents to the new volumes, waiting for the copies to synchronize, and then dropping the old copy. If a back- out path is desired, a snapshot of the original data can be created before the point at which the mirror copy is dropped. Alternatively, the server can be shut down briefly to mask original LUNs before forcibly deleting the contained mirror copies. Doing so preserves a recoverable copy of the data in its original location.

Extent migration

Almost all volume managers allow extents to be migrated, and sometimes multiple options exist. For example, some volume managers allow an administrator to relocate the individual extents for a specific logical volume from old to new storage. Volume managers such as Linux LVM2 offer the `pvmove` command, which relocates all extents on the specified LUN device to a new LUN. After the old LUN is evacuated, it can be removed.

> The primary risk to operations is the removal of old, unused LUNs from the configuration. Great care must be taken when changing FC zoning and removing stale LUN devices.

Oracle Automatic Storage Management

Oracle ASM is a combined logical volume manager and file system. At a high level, Oracle ASM takes a collection of LUNs, breaks them into small units of allocation, and presents them as a single volume known as an ASM disk group. ASM also includes the ability to mirror the disk group by setting the redundancy level. A volume can be unmirrored (external redundancy), mirrored (normal redundancy), or three-way mirrored (high redundancy). Care must be taken when configuring the redundancy level because it cannot be changed after creation.

ASM also provides file system functionality. Although the file system is not visible directly from the host, the Oracle database can create, move, and delete files and directories on an ASM disk group. Also, the structure can be navigated by using the asmcmd utility.

As with other LVM implementations, Oracle ASM optimizes I/O performance by striping and load-balancing the I/O of each file across all available LUNs. Second, the underlying extents can be relocated to enable both resizing of the ASM disk group as well as migration. Oracle ASM automates the process through the rebalancing operation. New LUNs are added to an ASM disk group and old LUNs are dropped, which triggers extent relocation and subsequent drop of the evacuated LUN from the disk group. This process is one of the most proven methods of migration, and the reliability of ASM at delivering transparent migration is possibly its most important feature.

> Because the mirroring level of Oracle ASM is fixed, it cannot be used with the mirror and demirror method of migration.

Storage-level migration

Storage-level migration means performing the migration below both the application and operating system level. In the past, this sometimes meant using specialized devices that would copy LUNs at the network level, but these capabilities are now found natively in ONTAP.

SnapMirror

Migration of databases from between NetApp systems is almost universally performed with the NetApp SnapMirror data replication software. The process involves setting up a mirror relationship for the volumes to be migrated, allowing them to synchronize, and then waiting for the cutover window. When it arrives, the source database is shut down, one final mirror update is performed, and the mirror is broken. The replica volumes are then ready for use, either by mounting a contained NFS file system directory or by discovering the contained LUNs and starting the database.

Relocating volumes within a single ONTAP cluster is not considered migration, but rather a routine `volume move` operation. SnapMirror is used as the data replication engine within the cluster. This process is fully automated. There are no additional migration steps to be performed when attributes of the volume, such as LUN mapping or the NFS export permissions, are moved with the volume itself. The relocation is nondisruptive to host operations. In some cases, network access must be updated to make sure that the newly relocated data is accessed in the most efficient way possible, but these tasks are also nondisruptive.

Foreign LUN Import (FLI)

FLI is a feature that allows a Data ONTAP system running 8.3 or higher to migrate an existing LUN from another storage array. The procedure is simple: The ONTAP system is zoned to the existing storage array as if it was any other SAN host. Data ONTAP then takes control of the desired legacy LUNs and migrates the underlying data. In addition, the import process uses the efficiency settings of the new volume as data is migrated, meaning that data can be compressed and deduplicated inline during the migration process.

The first implementation of FLI in Data ONTAP 8.3 permitted only offline migration. This was an extremely fast transfer, but it still meant that the LUN data was unavailable until the migration was complete. Online migration was introduced in Data ONTAP 8.3.1. This kind of migration minimizes disruption by allowing ONTAP to serve LUN data during the transfer process. There is a brief disruption while the host is rezoned to use the LUNs through ONTAP. However, as soon as those changes are made, the data is once again accessible and remains accessible throughout the migration process.

Read I/O is proxied through ONTAP until the copy operation is complete, while write I/O is synchronously written to both the foreign and ONTAP LUN. The two LUN copies are kept in sync in this manner until the

administrator executes a complete cutover that releases the foreign LUN and no longer replicates writes.

FLI is designed to work with FC, but if there is a desire to change to iSCSI, then the migrated LUN can easily be remapped as an iSCSI LUN after migration is completed.

Among the features of FLI is automatic alignment detection and adjustment. In this context, the term alignment refers to a partition on a LUN device. Optimum performance requires that I/O be aligned to 4K blocks. If a partition is placed at an offset that is not a multiple of 4K, performance suffers.

There is a second aspect of alignment that cannot be corrected by adjusting a partition offset—the file system block size. For example, a ZFS file system generally defaults to an internal block size of 512 bytes. Other customers using AIX have occasionally created jfs2 file systems with a 512- or 1, 024- byte block size. Although the file system might be aligned to a 4K boundary, the files created within that file system are not and performance suffers.

FLI should not be used in these circumstances. Although the data is accessible after migration, the result is file systems with serious performance limitations. As a general principle, any file system supporting a random overwrite workload on ONTAP should use a 4K block size. This is primarily applicable to workloads such as database data files and VDI deployments. The block size can be identified using the relevant host operating system commands.

For example, on AIX, the block size can be viewed with `lsfs -q`. With Linux, `xfs_info` and `tune2fs` can be used for `xfs` and `ext3/ext4`, respectively. With `zfs`, the command is `zdb -C`.

The parameter that controls the block size is `ashift` and generally defaults to a value of 9, which means 2^9, or 512 bytes. For optimum performance, the `ashift` value must be 12 (2^12=4K). This value is set at the time the zpool is created and cannot be changed, which means that data zpools with an `ashift` other than 12 should be migrated by copying data to a newly created zpool.

Oracle ASM does not have a fundamental block size. The only requirement is that the partition on which the ASM disk is built must be properly aligned.

7-Mode Transition Tool

The 7-Mode Transition Tool (7MTT) is an automation utility used to migrate large 7- Mode configurations to ONTAP. Most database customers find other methods easier, in part because they usually migrate their environments database by database rather than relocating the entire storage footprint. Additionally, databases are frequently only a part of a larger storage environment. Therefore, databases are often migrated individually, and then the remaining environment can be moved with 7MTT.

There is a small but significant number of customers who have storage systems that are dedicated to complicated database environments. These environments might contain many volumes, snapshots, and numerous configuration details such as export permissions, LUN initiator groups, user permissions, and Lightweight Directory Access Protocol configuration. In such cases, the automation abilities of 7MTT can simplify a migration.

7MTT can operate in one of two modes:

- **Copy- based transition (CBT).** 7MTT with CBT sets up SnapMirror volumes from an existing 7- Mode system in the new environment. After the data is in sync, 7MTT orchestrates the cutover process.
- **Copy- free transition (CFT).** 7MTT with CFT is based on the in-place conversion of existing 7- Mode disk shelves. No data is copied, and the existing disk shelves can be reused. The existing data protection and storage efficiency configuration is preserved.

The primary difference between these two options is that copy-free transition is a big- bang approach in which all disk shelves attached to the original 7- Mode HA pair must be relocated to the new environment. There is no option to move a subset of shelves. The copy-based approach allows selected volumes to be moved. There is also potentially a longer cutover window with copy-free transition because of the tie required to recable disk shelves and convert metadata. Based on field experience, NetApp recommends allowing 1 hour for relocating and recabling disk shelves and between 15 minutes and 2 hours for metadata conversion.

Oracle datafile migration

Individual Oracle datafiles can be moved with a single command.

For example, the following command moves the datafile IOPST.dbf from filesystem `/oradata2` to filesystem `/oradata3`.

```
SQL> alter database move datafile  '/oradata2/NTAP/IOPS002.dbf' to
'/oradata3/NTAP/IOPS002.dbf';
Database altered.
```

Moving a datafile with this method can be slow, but it normally should not produce enough I/O that it interferes with the day-to-day database workloads. In contrast, migration via ASM rebalancing can run much faster but at the expense of slowing down the overall database while the data is being moved.

The time required to move datafiles can easily be measured by creating a test datafile and then moving it. The elapsed time for the operation is recorded in the v$session data:

```
SQL> set linesize 300;
SQL> select elapsed_seconds||':'||message from v$session_longops;
ELAPSED_SECONDS||':'||MESSAGE
------------------------------------------------------------------------
---------------
351:Online data file move: data file 8: 22548578304 out of 22548578304
bytes done
SQL> select bytes / 1024 / 1024 /1024 as GB from dba_data_files where
FILE_ID = 8;
        GB
----------
        21
```

In this example, the file that was moved was datafile 8, which was 21GB in size and required about 6 minutes to migrate. The time required obviously depends on the capabilities of the storage system, the storage network, and the overall database activity occurring at the time of migration.

Oracle database migration via log shipping

The goal of a migration using log shipping is to create a copy of the original data files at a new location and then establish a method of shipping changes into the new environment.

Once established, log shipment and replay can be automated to keep the replica database largely in sync with

the source. For example, a cron job can be scheduled to (a) copy the most recent logs to the new location and (b) replay them every 15 minutes. Doing so provides minimal disruption at the time of cutover because no more than 15 minutes of archive logs must be replayed.

The procedure shown below is also essentially a database clone operation. The logic shown is similar to the engine within NetApp SnapManager for Oracle (SMO) and the NetApp SnapCenter Oracle Plug- in. Some customers have used the procedure shown within scripts or WFA workflows for custom cloning operations. Although this procedure is more manual than using either SMO or SnapCenter, it is still readily scripted and the data management APIs within ONTAP further simplify the process.

Log shipping - file system to file system

This example demonstrates the migration of a database called WAFFLE from an ordinary file system to another ordinary file system located on a different server. It also illustrates the use of SnapMirror to make a rapid copy of data files, but this is not an integral part of the overall procedure.

Create database backup

The first step is to create a database backup. Specifically, this procedure requires a set of data files that can be used for archive log replay.

Environment

In this example, the source database is on an ONTAP system. The simplest method to create a backup of a database is by using a snapshot. The database is placed in hot backup mode for a few seconds while a `snapshot create` operation is executed on the volume hosting the data files.

```
SQL> alter database begin backup;
Database altered.
```

```
Cluster01::*> snapshot create -vserver vserver1 -volume jfsc1_oradata
hotbackup
Cluster01::*>
```

```
SQL> alter database end backup;
Database altered.
```

The result is a snapshot on disk called `hotbackup` that contains an image of the data files while in hot backup mode. When combined with the appropriate archive logs to make the data files consistent, the data in this snapshot can be used as the basis of a restore or a clone. In this case, it is replicated to the new server.

Restore to new environment

The backup must now be restored in the new environment. This can be done in a number of ways, including Oracle RMAN, restoration from a backup application like NetBackup, or a simple copy operation of data files that were placed in hot backup mode.

In this example, SnapMirror is used to replicate the snapshot hotbackup to a new location.

1. Create a new volume to receive the snapshot data. Initialize the mirroring from `jfsc1_oradata` to `vol_oradata`.

```
Cluster01::*> volume create -vserver vserver1 -volume vol_oradata
-aggregate data_01 -size 20g -state online -type DP -snapshot-policy
none -policy jfsc3
[Job 833] Job succeeded: Successful
```

```
Cluster01::*> snapmirror initialize -source-path vserver1:jfsc1_oradata
-destination-path vserver1:vol_oradata
Operation is queued: snapmirror initialize of destination
"vserver1:vol_oradata".
Cluster01::*> volume mount -vserver vserver1 -volume vol_oradata
-junction-path /vol_oradata
Cluster01::*>
```

2. After the state is set by SnapMirror, indicating that synchronization is complete, update the mirror based specifically on the desired snapshot.

```
Cluster01::*> snapmirror show -destination-path vserver1:vol_oradata
-fields state
source-path             destination-path        state
----------------------- ----------------------- ------------
vserver1:jfsc1_oradata vserver1:vol_oradata SnapMirrored
```

```
Cluster01::*> snapmirror update -destination-path vserver1:vol_oradata
-source-snapshot hotbackup
Operation is queued: snapmirror update of destination
"vserver1:vol_oradata".
```

3. Successful synchronization can be verified by viewing the `newest-snapshot` field on the mirror volume.

```
Cluster01::*> snapmirror show -destination-path vserver1:vol_oradata
-fields newest-snapshot
source-path             destination-path        newest-snapshot
----------------------- ----------------------- ---------------
vserver1:jfsc1_oradata vserver1:vol_oradata hotbackup
```

4. The mirror can then be broken.

```
Cluster01::> snapmirror break -destination-path vserver1:vol_oradata
Operation succeeded: snapmirror break for destination
"vserver1:vol_oradata".
Cluster01::>
```

5. Mount the new file system.With block-based file systems, the precise procedures vary based on the LVM in use. FC zoning or iSCSI connections must be configured. After connectivity to the LUNs is established, commands such as Linux `pvscan` might be needed to discover which volume groups or LUNs need to be properly configured to be discoverable by ASM.

   In this example, a simple NFS file system is used. This file system can be mounted directly.

```
fas8060-nfs1:/vol_oradata        19922944   1639360   18283584   9%
/oradata
fas8060-nfs1:/vol_logs            9961472       128    9961344   1%
/logs
```

Create controlfile creation template

You must next create a controlfile template. The `backup controlfile to trace` command creates text commands to recreate a controlfile. This function can be useful for restoring a database from backup under some circumstances, and it is often used with scripts that perform tasks such as database cloning.

1. The output of the following command is used to recreate the controlfiles for the migrated database.

```
SQL> alter database backup controlfile to trace as '/tmp/waffle.ctrl';
Database altered.
```

2. After the controlfiles have been created, copy the file to the new server.

```
[oracle@jfsc3 tmp]$ scp oracle@jfsc1:/tmp/waffle.ctrl /tmp/
oracle@jfsc1's password:
waffle.ctrl                                           100% 5199
5.1KB/s   00:00
```

Backup parameter file

A parameter file is also required in the new environment. The simplest method is to create a pfile from the current spfile or pfile. In this example, the source database is using an spfile.

```
SQL> create pfile='/tmp/waffle.tmp.pfile' from spfile;
File created.
```

Create oratab entry

The creation of an oratab entry is required for the proper functioning of utilities such as oraenv. To create an oratab entry, complete the following step.

```
WAFFLE:/orabin/product/12.1.0/dbhome_1:N
```

Prepare directory structure

If the required directories were not already present, you must create them or the database startup procedure fails. To prepare the directory structure, complete the following minimum requirements.

```
[oracle@jfsc3 ~]$ . oraenv
ORACLE_SID = [oracle] ? WAFFLE
The Oracle base has been set to /orabin
[oracle@jfsc3 ~]$ cd $ORACLE_BASE
[oracle@jfsc3 orabin]$ cd admin
[oracle@jfsc3 admin]$ mkdir WAFFLE
[oracle@jfsc3 admin]$ cd WAFFLE
[oracle@jfsc3 WAFFLE]$ mkdir adump dpdump pfile scripts xdb_wallet
```

Parameter file updates

1. To copy the parameter file to the new server, run the following commands. The default location is the `$ORACLE_HOME/dbs` directory. In this case, the pfile can be placed anywhere. It is only being used as an intermediate step in the migration process.

```
[oracle@jfsc3 admin]$ scp oracle@jfsc1:/tmp/waffle.tmp.pfile
$ORACLE_HOME/dbs/waffle.tmp.pfile
oracle@jfsc1's password:
waffle.pfile                                          100%  916
0.9KB/s   00:00
```

1. Edit the file as required. For example, if the archive log location has changed, the pfile must be altered to reflect the new location. In this example, only the controlfiles are being relocated, in part to distribute them between the log and data file systems.

```
[root@jfsc1 tmp]# cat waffle.pfile
WAFFLE.__data_transfer_cache_size=0
WAFFLE.__db_cache_size=507510784
WAFFLE.__java_pool_size=4194304
WAFFLE.__large_pool_size=20971520
WAFFLE.__oracle_base='/orabin'#ORACLE_BASE set from environment
WAFFLE.__pga_aggregate_target=268435456
WAFFLE.__sga_target=805306368
WAFFLE.__shared_io_pool_size=29360128
WAFFLE.__shared_pool_size=234881024
WAFFLE.__streams_pool_size=0
*.audit_file_dest='/orabin/admin/WAFFLE/adump'
*.audit_trail='db'
*.compatible='12.1.0.2.0'
*.control_files='/oradata//WAFFLE/control01.ctl','/oradata//WAFFLE/contr
ol02.ctl'
*.control_files='/oradata/WAFFLE/control01.ctl','/logs/WAFFLE/control02.
ctl'
*.db_block_size=8192
*.db_domain=''
*.db_name='WAFFLE'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=WAFFLEXDB)'
*.log_archive_dest_1='LOCATION=/logs/WAFFLE/arch'
*.log_archive_format='%t_%s_%r.dbf'
*.open_cursors=300
*.pga_aggregate_target=256m
*.processes=300
*.remote_login_passwordfile='EXCLUSIVE'
*.sga_target=768m
*.undo_tablespace='UNDOTBS1'
```

2. After the edits are complete, create an spfile based on this pfile.

```
SQL> create spfile from pfile='waffle.tmp.pfile';
File created.
```

Recreate controlfiles

In a previous step, the output of `backup controlfile to trace` was copied to the new server. The specific portion of the output required is the `controlfile recreation` command. This information can be found in the file under the section marked `Set #1. NORESETLOGS`. It starts with the line `create controlfile reuse database` and should include the word `noresetlogs`. It ends with the semicolon (;) character.

1. In this example procedure, the file reads as follows.

```
CREATE CONTROLFILE REUSE DATABASE "WAFFLE" NORESETLOGS  ARCHIVELOG
    MAXLOGFILES 16
    MAXLOGMEMBERS 3
    MAXDATAFILES 100
    MAXINSTANCES 8
    MAXLOGHISTORY 292
LOGFILE
  GROUP 1 '/logs/WAFFLE/redo/redo01.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 2 '/logs/WAFFLE/redo/redo02.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 3 '/logs/WAFFLE/redo/redo03.log'  SIZE 50M BLOCKSIZE 512
-- STANDBY LOGFILE
DATAFILE
  '/oradata/WAFFLE/system01.dbf',
  '/oradata/WAFFLE/sysaux01.dbf',
  '/oradata/WAFFLE/undotbs01.dbf',
  '/oradata/WAFFLE/users01.dbf'
CHARACTER SET WE8MSWIN1252
;
```

2. Edit this script as desired to reflect the new location of the various files. For example, certain data files known to support high I/O might be redirected to a file system on a high- performance storage tier. In other cases, the changes might be purely for administrator reasons, such as isolating the data files of a given PDB in dedicated volumes.
3. In this example, the `DATAFILE` stanza is left unchanged, but the redo logs are moved to a new location in `/redo` rather than sharing space with archive logs in `/logs`.

```
CREATE CONTROLFILE REUSE DATABASE "WAFFLE" NORESETLOGS  ARCHIVELOG
    MAXLOGFILES 16
    MAXLOGMEMBERS 3
    MAXDATAFILES 100
    MAXINSTANCES 8
    MAXLOGHISTORY 292
LOGFILE
  GROUP 1 '/redo/redo01.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 2 '/redo/redo02.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 3 '/redo/redo03.log'  SIZE 50M BLOCKSIZE 512
-- STANDBY LOGFILE
DATAFILE
  '/oradata/WAFFLE/system01.dbf',
  '/oradata/WAFFLE/sysaux01.dbf',
  '/oradata/WAFFLE/undotbs01.dbf',
  '/oradata/WAFFLE/users01.dbf'
CHARACTER SET WE8MSWIN1252
;
```

```
SQL> startup nomount;
ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             331353200 bytes
Database Buffers          465567744 bytes
Redo Buffers                5455872 bytes
SQL> CREATE CONTROLFILE REUSE DATABASE "WAFFLE" NORESETLOGS  ARCHIVELOG
  2      MAXLOGFILES 16
  3      MAXLOGMEMBERS 3
  4      MAXDATAFILES 100
  5      MAXINSTANCES 8
  6      MAXLOGHISTORY 292
  7  LOGFILE
  8    GROUP 1 '/redo/redo01.log'  SIZE 50M BLOCKSIZE 512,
  9    GROUP 2 '/redo/redo02.log'  SIZE 50M BLOCKSIZE 512,
 10    GROUP 3 '/redo/redo03.log'  SIZE 50M BLOCKSIZE 512
 11  -- STANDBY LOGFILE
 12  DATAFILE
 13    '/oradata/WAFFLE/system01.dbf',
 14    '/oradata/WAFFLE/sysaux01.dbf',
 15    '/oradata/WAFFLE/undotbs01.dbf',
 16    '/oradata/WAFFLE/users01.dbf'
 17  CHARACTER SET WE8MSWIN1252
 18  ;
Control file created.
SQL>
```

If any files are misplaced or parameters are misconfigured, errors are generated that indicate what must be fixed. The database is mounted, but it is not yet open and cannot be opened because the data files in use are still marked as being in hot backup mode. Archive logs must first be applied to make the database consistent.

Initial log replication

At least one log reply operation is required to make the data files consistent. Many options are available to replay logs. In some cases, the original archive log location on the original server can be shared through NFS, and log reply can be done directly. In other cases, the archive logs must be copied.

For example, a simple `scp` operation can copy all current logs from the source server to the migration server:

```
[oracle@jfsc3 arch]$ scp jfsc1:/logs/WAFFLE/arch/* ./
oracle@jfsc1's password:
1_22_912662036.dbf                                          100%   47MB
47.0MB/s   00:01
1_23_912662036.dbf                                          100%   40MB
40.4MB/s   00:00
1_24_912662036.dbf                                          100%   45MB
45.4MB/s   00:00
1_25_912662036.dbf                                          100%   41MB
40.9MB/s   00:01
1_26_912662036.dbf                                          100%   39MB
39.4MB/s   00:00
1_27_912662036.dbf                                          100%   39MB
38.7MB/s   00:00
1_28_912662036.dbf                                          100%   40MB
40.1MB/s   00:01
1_29_912662036.dbf                                          100%   17MB
16.9MB/s   00:00
1_30_912662036.dbf                                          100%  636KB
636.0KB/s   00:00
```

Initial log replay

After the files are in the archive log location, they can be replayed by issuing the command `recover database until cancel` followed by the response `AUTO` to automatically replay all available logs.

```
SQL> recover database until cancel;
ORA-00279: change 382713 generated at 05/24/2016 09:00:54 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_23_912662036.dbf
ORA-00280: change 382713 for thread 1 is in sequence #23
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00279: change 405712 generated at 05/24/2016 15:01:05 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_24_912662036.dbf
ORA-00280: change 405712 for thread 1 is in sequence #24
ORA-00278: log file '/logs/WAFFLE/arch/1_23_912662036.dbf' no longer
needed for
this recovery
...
ORA-00279: change 713874 generated at 05/26/2016 04:26:43 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_31_912662036.dbf
ORA-00280: change 713874 for thread 1 is in sequence #31
ORA-00278: log file '/logs/WAFFLE/arch/1_30_912662036.dbf' no longer
needed for
this recovery
ORA-00308: cannot open archived log '/logs/WAFFLE/arch/1_31_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

The final archive log reply reports an error, but this is normal. The log indicates that `sqlplus` was seeking a particular log file and did not find it. The reason is, most likely, that the log file does not exist yet.

If the source database can be shut down before copying archive logs, this step must be performed only once. The archive logs are copied and replayed, and then the process can continue directly to the cutover process that replicates the critical redo logs.

Incremental log replication and replay

In most cases, migration is not performed right away. It could be days or even weeks before the migration process is completed, which means that the logs must be continuously shipped to the replica database and replayed. Therefore, when cutover arrives, minimal data must be transferred and replayed.

Doing so can be scripted in many ways, but one of the more popular methods is using rsync, a common file replication utility. The safest way to use this utility is to configure it as a daemon. For example, the `rsyncd.conf` file that follows shows how to create a resource called `waffle.arch` that is accessed with Oracle user credentials and is mapped to `/logs/WAFFLE/arch`. Most importantly, the resource is set to read-only, which allows the production data to be read but not altered.

```
[root@jfsc1 arch]# cat /etc/rsyncd.conf
[waffle.arch]
   uid=oracle
   gid=dba
   path=/logs/WAFFLE/arch
   read only = true
[root@jfsc1 arch]# rsync --daemon
```

The following command synchronizes the new server's archive log destination against the rsync resource `waffle.arch` on the original server. The `t` argument in `rsync - potg` causes the file list to be compared based on timestamp, and only new files are copied. This process provides an incremental update of the new server. This command can also be scheduled in cron to run on a regular basis.

```
[oracle@jfsc3 arch]$ rsync -potg --stats --progress jfsc1::waffle.arch/*
/logs/WAFFLE/arch/
1_31_912662036.dbf
      650240 100%  124.02MB/s    0:00:00 (xfer#1, to-check=8/18)
1_32_912662036.dbf
     4873728 100%  110.67MB/s    0:00:00 (xfer#2, to-check=7/18)
1_33_912662036.dbf
     4088832 100%   50.64MB/s    0:00:00 (xfer#3, to-check=6/18)
1_34_912662036.dbf
     8196096 100%   54.66MB/s    0:00:00 (xfer#4, to-check=5/18)
1_35_912662036.dbf
    19376128 100%   57.75MB/s    0:00:00 (xfer#5, to-check=4/18)
1_36_912662036.dbf
       71680 100%  201.15kB/s    0:00:00 (xfer#6, to-check=3/18)
1_37_912662036.dbf
     1144320 100%    3.06MB/s    0:00:00 (xfer#7, to-check=2/18)
1_38_912662036.dbf
    35757568 100%   63.74MB/s    0:00:00 (xfer#8, to-check=1/18)
1_39_912662036.dbf
      984576 100%    1.63MB/s    0:00:00 (xfer#9, to-check=0/18)
Number of files: 18
Number of files transferred: 9
Total file size: 399653376 bytes
Total transferred file size: 75143168 bytes
Literal data: 75143168 bytes
Matched data: 0 bytes
File list size: 474
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 204
Total bytes received: 75153219
sent 204 bytes  received 75153219 bytes  150306846.00 bytes/sec
total size is 399653376  speedup is 5.32
```

After the logs have been received, they must be replayed. Previous examples show the use of sqlplus to manually run `recover database until cancel`, a process that can easily be automated. The example shown here uses the script described in Replay Logs on Database. The scripts accept an argument that specifies the database requiring a replay operation. This permits the same script to be used in a multidatabase migration effort.

```
[oracle@jfsc3 logs]$ ./replay.logs.pl WAFFLE
ORACLE_SID = [WAFFLE] ? The Oracle base remains unchanged with value
/orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu May 26 10:47:16 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> ORA-00279: change 713874 generated at 05/26/2016 04:26:43 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_31_912662036.dbf
ORA-00280: change 713874 for thread 1 is in sequence #31
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 814256 generated at 05/26/2016 04:52:30 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_32_912662036.dbf
ORA-00280: change 814256 for thread 1 is in sequence #32
ORA-00278: log file '/logs/WAFFLE/arch/1_31_912662036.dbf' no longer
needed for
this recovery
ORA-00279: change 814780 generated at 05/26/2016 04:53:04 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_33_912662036.dbf
ORA-00280: change 814780 for thread 1 is in sequence #33
ORA-00278: log file '/logs/WAFFLE/arch/1_32_912662036.dbf' no longer
needed for
this recovery
...
ORA-00279: change 1120099 generated at 05/26/2016 09:59:21 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_40_912662036.dbf
ORA-00280: change 1120099 for thread 1 is in sequence #40
ORA-00278: log file '/logs/WAFFLE/arch/1_39_912662036.dbf' no longer
needed for
this recovery
ORA-00308: cannot open archived log '/logs/WAFFLE/arch/1_40_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
```

Cutover

When you are ready to cut over to the new environment, you must perform one final synchronization that includes both archive logs and the redo logs. If the original redo log location is not already known, it can be identified as follows:

```
SQL> select member from v$logfile;
MEMBER
------------------------------------------------------------------------
------
/logs/WAFFLE/redo/redo01.log
/logs/WAFFLE/redo/redo02.log
/logs/WAFFLE/redo/redo03.log
```

1. Shut down the source database.
2. Perform one final synchronization of the archive logs on the new server with the desired method.
3. The source redo logs must be copied to the new server. In this example, the redo logs were relocated to a new directory at `/redo`.

   ```
   [oracle@jfsc3 logs]$ scp jfsc1:/logs/WAFFLE/redo/* /redo/
   oracle@jfsc1's password:
   redo01.log
   100%   50MB  50.0MB/s   00:01
   redo02.log
   100%   50MB  50.0MB/s   00:00
   redo03.log
   100%   50MB  50.0MB/s   00:00
   ```

4. At this stage, the new database environment contains all of the files required to bring it to the exact same state as the source. The archive logs must be replayed one final time.

```
SQL> recover database until cancel;
ORA-00279: change 1120099 generated at 05/26/2016 09:59:21 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_40_912662036.dbf
ORA-00280: change 1120099 for thread 1 is in sequence #40
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00308: cannot open archived log
'/logs/WAFFLE/arch/1_40_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
ORA-00308: cannot open archived log
'/logs/WAFFLE/arch/1_40_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

5. Once complete, the redo logs must be replayed. If the message `Media recovery complete` is returned, the process is successful and the databases are synchronized and can be opened.

```
SQL> recover database;
Media recovery complete.
SQL> alter database open;
Database altered.
```

Log shipping - ASM to file system

This example demonstrates the use of Oracle RMAN to migrate a database. It is very similar to the prior example of file system to file system log shipping, but the files on ASM are not visible to the host. The only options for migrating data located on ASM devices is either by relocating the ASM LUN or by using Oracle RMAN to perform the copy operations.

Although RMAN is a requirement for copying files from Oracle ASM, the use of RMAN is not limited to ASM. RMAN can be used to migrate from any type of storage to any other type.

This example shows the relocation of a database called PANCAKE from ASM storage to a regular file system located on a different server at paths `/oradata` and `/logs`.

Create database backup

The first step is to create a backup of the database to be migrated to an alternate server. Because the source uses Oracle ASM, RMAN must be used. A simple RMAN backup can be performed as follows. This method creates a tagged backup that can be easily identified by RMAN later in the procedure.

The first command defines the type of destination for the backup and the location to be used. The second initiates the backup of the data files only.

```
RMAN> configure channel device type disk format '/rman/pancake/%U';
using target database control file instead of recovery catalog
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT   '/rman/pancake/%U';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT   '/rman/pancake/%U';
new RMAN configuration parameters are successfully stored
RMAN> backup database tag 'ONTAP_MIGRATION';
Starting backup at 24-MAY-16
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=251 device type=DISK
channel ORA_DISK_1: starting full datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
input datafile file number=00001 name=+ASM0/PANCAKE/system01.dbf
input datafile file number=00002 name=+ASM0/PANCAKE/sysaux01.dbf
input datafile file number=00003 name=+ASM0/PANCAKE/undotbs101.dbf
input datafile file number=00004 name=+ASM0/PANCAKE/users01.dbf
channel ORA_DISK_1: starting piece 1 at 24-MAY-16
channel ORA_DISK_1: finished piece 1 at 24-MAY-16
piece handle=/rman/pancake/1gr6c161_1_1 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:03
channel ORA_DISK_1: starting full datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current control file in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 24-MAY-16
channel ORA_DISK_1: finished piece 1 at 24-MAY-16
piece handle=/rman/pancake/1hr6c164_1_1 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 24-MAY-16
```

Backup controlfile

A backup controlfile is required later in the procedure for the `duplicate database` operation.

```
RMAN> backup current controlfile format '/rman/pancake/ctrl.bkp';
Starting backup at 24-MAY-16
using channel ORA_DISK_1
channel ORA_DISK_1: starting full datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current control file in backup set
channel ORA_DISK_1: starting piece 1 at 24-MAY-16
channel ORA_DISK_1: finished piece 1 at 24-MAY-16
piece handle=/rman/pancake/ctrl.bkp tag=TAG20160524T032651 comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 24-MAY-16
```

Backup parameter file

A parameter file is also required in the new environment. The simplest method is to create a pfile from the current spfile or pfile. In this example, the source database uses an spfile.

```
RMAN> create pfile='/rman/pancake/pfile' from spfile;
Statement processed
```

ASM file rename script

Several file locations currently defined in the controlfiles change when the database is moved. The following script creates an RMAN script to make the process easier. This example shows a database with a very small number of data files, but typically databases contain hundreds or even thousands of data files.

This script can be found in ASM to File System Name Conversion and it does two things.

First, it creates a parameter to redefine the redo log locations called `log_file_name_convert`. It is essentially a list of alternating fields. The first field is the location of a current redo log, and the second field is the location on the new server. The pattern is then repeated.

The second function is to supply a template for data file renaming. The script loops through the data files, pulls the name and file number information, and formats it as an RMAN script. Then it does the same with the temp files. The result is a simple rman script that can be edited as desired to make sure that the files are restored to the desired location.

```
SQL> @/rman/mk.rename.scripts.sql
Parameters for log file conversion:
*.log_file_name_convert = '+ASM0/PANCAKE/redo01.log',
'/NEW_PATH/redo01.log','+ASM0/PANCAKE/redo02.log',
'/NEW_PATH/redo02.log','+ASM0/PANCAKE/redo03.log', '/NEW_PATH/redo03.log'
rman duplication script:
run
{
set newname for datafile 1 to '+ASM0/PANCAKE/system01.dbf';
set newname for datafile 2 to '+ASM0/PANCAKE/sysaux01.dbf';
set newname for datafile 3 to '+ASM0/PANCAKE/undotbs101.dbf';
set newname for datafile 4 to '+ASM0/PANCAKE/users01.dbf';
set newname for tempfile 1 to '+ASM0/PANCAKE/temp01.dbf';
duplicate target database for standby backup location INSERT_PATH_HERE;
}
PL/SQL procedure successfully completed.
```

Capture the output of this screen. The `log_file_name_convert` parameter is placed in the pfile as described below. The RMAN data file rename and duplicate script must be edited accordingly to place the data files in the desired locations. In this example, they are all placed in `/oradata/pancake`.

```
run
{
set newname for datafile 1 to '/oradata/pancake/pancake.dbf';
set newname for datafile 2 to '/oradata/pancake/sysaux.dbf';
set newname for datafile 3 to '/oradata/pancake/undotbs1.dbf';
set newname for datafile 4 to '/oradata/pancake/users.dbf';
set newname for tempfile 1 to '/oradata/pancake/temp.dbf';
duplicate target database for standby backup location '/rman/pancake';
}
```

Prepare directory structure

The scripts are almost ready to execute, but first the directory structure must be in place. If the required directories are not already present, they must be created or the database startup procedure fails. The example below reflects the minimum requirements.

```
[oracle@jfsc2 ~]$ mkdir /oradata/pancake
[oracle@jfsc2 ~]$ mkdir /logs/pancake
[oracle@jfsc2 ~]$ cd /orabin/admin
[oracle@jfsc2 admin]$ mkdir PANCAKE
[oracle@jfsc2 admin]$ cd PANCAKE
[oracle@jfsc2 PANCAKE]$ mkdir adump dpdump pfile scripts xdb_wallet
```

Create oratab entry

The following command is required for utilities such as oraenv to work properly.

```
PANCAKE:/orabin/product/12.1.0/dbhome_1:N
```

Parameter updates

The saved pfile must be updated to reflect any path changes on the new server. The data file path changes are changed by the RMAN duplication script, and nearly all databases require changes to the `control_files` and `log_archive_dest` parameters. There might also be audit file locations that must be changed, and parameters such as `db_create_file_dest` might not be relevant outside of ASM. An experienced DBA should carefully review the proposed changes before proceeding.

In this example, the key changes are the controlfile locations, the log archive destination, and the addition of the `log_file_name_convert` parameter.

```
PANCAKE.__data_transfer_cache_size=0
PANCAKE.__db_cache_size=545259520
PANCAKE.__java_pool_size=4194304
PANCAKE.__large_pool_size=25165824
PANCAKE.__oracle_base='/orabin'#ORACLE_BASE set from environment
PANCAKE.__pga_aggregate_target=268435456
PANCAKE.__sga_target=805306368
PANCAKE.__shared_io_pool_size=29360128
PANCAKE.__shared_pool_size=192937984
PANCAKE.__streams_pool_size=0
*.audit_file_dest='/orabin/admin/PANCAKE/adump'
*.audit_trail='db'
*.compatible='12.1.0.2.0'
*.control_files='+ASM0/PANCAKE/control01.ctl','+ASM0/PANCAKE/control02.ctl
'
*.control_files='/oradata/pancake/control01.ctl','/logs/pancake/control02.
ctl'
*.db_block_size=8192
*.db_domain=''
*.db_name='PANCAKE'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=PANCAKEXDB)'
*.log_archive_dest_1='LOCATION=+ASM1'
*.log_archive_dest_1='LOCATION=/logs/pancake'
*.log_archive_format='%t_%s_%r.dbf'
'/logs/path/redo02.log'
*.log_file_name_convert = '+ASM0/PANCAKE/redo01.log',
'/logs/pancake/redo01.log', '+ASM0/PANCAKE/redo02.log',
'/logs/pancake/redo02.log', '+ASM0/PANCAKE/redo03.log',
'/logs/pancake/redo03.log'
*.open_cursors=300
*.pga_aggregate_target=256m
*.processes=300
*.remote_login_passwordfile='EXCLUSIVE'
*.sga_target=768m
*.undo_tablespace='UNDOTBS1'
```

After the new parameters are confirmed, the parameters must be put into effect. Multiple options exist, but most customers create an spfile based on the text pfile.

```
bash-4.1$ sqlplus / as sysdba
SQL*Plus: Release 12.1.0.2.0 Production on Fri Jan 8 11:17:40 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> create spfile from pfile='/rman/pancake/pfile';
File created.
```

Startup nomount

The final step before replicating the database is to bring up the database processes but not mount the files. In this step, problems with the spfile might become evident. If the `startup nomount` command fails because of a parameter error, it is simple to shut down, correct the pfile template, reload it as an spfile, and try again.

```
SQL> startup nomount;
ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             373296240 bytes
Database Buffers          423624704 bytes
Redo Buffers                5455872 bytes
```

Duplicate the database

Restoring the prior RMAN backup to the new location consumes more time than other steps in this process. The database must be duplicated without a change to the database ID (DBID) or resetting the logs. This prevents logs from being applied, which is a required step to fully synchronize the copies.

Connect to the database with RMAN as aux and issue the duplicate database command by using the script created in a previous step.

```
[oracle@jfsc2 pancake]$ rman auxiliary /
Recovery Manager: Release 12.1.0.2.0 - Production on Tue May 24 03:04:56
2016
Copyright (c) 1982, 2014, Oracle and/or its affiliates.  All rights
reserved.
connected to auxiliary database: PANCAKE (not mounted)
RMAN> run
2> {
3> set newname for datafile 1 to '/oradata/pancake/pancake.dbf';
4> set newname for datafile 2 to '/oradata/pancake/sysaux.dbf';
5> set newname for datafile 3 to '/oradata/pancake/undotbs1.dbf';
6> set newname for datafile 4 to '/oradata/pancake/users.dbf';
7> set newname for tempfile 1 to '/oradata/pancake/temp.dbf';
8> duplicate target database for standby backup location '/rman/pancake';
9> }
```

```
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
Starting Duplicate Db at 24-MAY-16
contents of Memory Script:
{
   restore clone standby controlfile from  '/rman/pancake/ctrl.bkp';
}
executing Memory Script
Starting restore at 24-MAY-16
allocated channel: ORA_AUX_DISK_1
channel ORA_AUX_DISK_1: SID=243 device type=DISK
channel ORA_AUX_DISK_1: restoring control file
channel ORA_AUX_DISK_1: restore complete, elapsed time: 00:00:01
output file name=/oradata/pancake/control01.ctl
output file name=/logs/pancake/control02.ctl
Finished restore at 24-MAY-16
contents of Memory Script:
{
   sql clone 'alter database mount standby database';
}
executing Memory Script
sql statement: alter database mount standby database
released channel: ORA_AUX_DISK_1
allocated channel: ORA_AUX_DISK_1
channel ORA_AUX_DISK_1: SID=243 device type=DISK
contents of Memory Script:
{
   set newname for tempfile  1 to
 "/oradata/pancake/temp.dbf";
   switch clone tempfile all;
   set newname for datafile  1 to
 "/oradata/pancake/pancake.dbf";
   set newname for datafile  2 to
 "/oradata/pancake/sysaux.dbf";
   set newname for datafile  3 to
 "/oradata/pancake/undotbs1.dbf";
   set newname for datafile  4 to
 "/oradata/pancake/users.dbf";
   restore
   clone database
   ;
}
executing Memory Script
```

```
executing command: SET NEWNAME
renamed tempfile 1 to /oradata/pancake/temp.dbf in control file
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
Starting restore at 24-MAY-16
using channel ORA_AUX_DISK_1
channel ORA_AUX_DISK_1: starting datafile backup set restore
channel ORA_AUX_DISK_1: specifying datafile(s) to restore from backup set
channel ORA_AUX_DISK_1: restoring datafile 00001 to
/oradata/pancake/pancake.dbf
channel ORA_AUX_DISK_1: restoring datafile 00002 to
/oradata/pancake/sysaux.dbf
channel ORA_AUX_DISK_1: restoring datafile 00003 to
/oradata/pancake/undotbs1.dbf
channel ORA_AUX_DISK_1: restoring datafile 00004 to
/oradata/pancake/users.dbf
channel ORA_AUX_DISK_1: reading from backup piece
/rman/pancake/1gr6c161_1_1
channel ORA_AUX_DISK_1: piece handle=/rman/pancake/1gr6c161_1_1
tag=ONTAP_MIGRATION
channel ORA_AUX_DISK_1: restored backup piece 1
channel ORA_AUX_DISK_1: restore complete, elapsed time: 00:00:07
Finished restore at 24-MAY-16
contents of Memory Script:
{
   switch clone datafile all;
}
executing Memory Script
datafile 1 switched to datafile copy
input datafile copy RECID=5 STAMP=912655725 file
name=/oradata/pancake/pancake.dbf
datafile 2 switched to datafile copy
input datafile copy RECID=6 STAMP=912655725 file
name=/oradata/pancake/sysaux.dbf
datafile 3 switched to datafile copy
input datafile copy RECID=7 STAMP=912655725 file
name=/oradata/pancake/undotbs1.dbf
datafile 4 switched to datafile copy
input datafile copy RECID=8 STAMP=912655725 file
name=/oradata/pancake/users.dbf
Finished Duplicate Db at 24-MAY-16
```

Initial log replication

You must now ship the changes from the source database to a new location. Doing so might require a combination of steps. The simplest method would be to have RMAN on the source database write out archive logs to a shared network connection. If a shared location is not available, an alternative method is using RMAN to write to a local file system and then using rcp or rsync to copy the files.

In this example, the `/rman` directory is an NFS share that is available to both the original and migrated database.

One important issue here is the `disk format` clause. The disk format of the backup is `%h_%e_%a.dbf`, which means that you must use the format of thread number, sequence number, and activation ID for the database. Although the letters are different, this matches the `log_archive_format='%t_%s_%r.dbf` parameter in the pfile. This parameter also specifies archive logs in the format of thread number, sequence number, and activation ID. The end result is that the log file backups on the source use a naming convention that is expected by the database. Doing so makes operations such as `recover database` much simpler because sqlplus correctly anticipates the names of the archive logs to be replayed.

```
RMAN> configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/arch/%h_%e_%a.dbf';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters are successfully stored
released channel: ORA_DISK_1
RMAN> backup as copy archivelog from time 'sysdate-2';
Starting backup at 24-MAY-16
current log archived
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=373 device type=DISK
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=54 RECID=70 STAMP=912658508
output file name=/rman/pancake/logship/1_54_912576125.dbf RECID=123
STAMP=912659482
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=41 RECID=29 STAMP=912654101
output file name=/rman/pancake/logship/1_41_912576125.dbf RECID=124
STAMP=912659483
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
...
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=45 RECID=33 STAMP=912654688
output file name=/rman/pancake/logship/1_45_912576125.dbf RECID=152
STAMP=912659514
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=47 RECID=36 STAMP=912654809
output file name=/rman/pancake/logship/1_47_912576125.dbf RECID=153
STAMP=912659515
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
Finished backup at 24-MAY-16
```

Initial log replay

After the files are in the archive log location, they can be replayed by issuing the command `recover database until cancel` followed by the response `AUTO` to automatically replay all available logs. The parameter file is currently directing archive logs to `/logs/archive`, but this does not match the location where RMAN was used to save logs. The location can be temporarily redirected as follows before recovering the database.

```
SQL> alter system set log_archive_dest_1='LOCATION=/rman/pancake/logship'
scope=memory;
System altered.
SQL> recover standby database until cancel;
ORA-00279: change 560224 generated at 05/24/2016 03:25:53 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_49_912576125.dbf
ORA-00280: change 560224 for thread 1 is in sequence #49
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00279: change 560353 generated at 05/24/2016 03:29:17 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_50_912576125.dbf
ORA-00280: change 560353 for thread 1 is in sequence #50
ORA-00278: log file '/rman/pancake/logship/1_49_912576125.dbf' no longer
needed
for this recovery
...
ORA-00279: change 560591 generated at 05/24/2016 03:33:56 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_54_912576125.dbf
ORA-00280: change 560591 for thread 1 is in sequence #54
ORA-00278: log file '/rman/pancake/logship/1_53_912576125.dbf' no longer
needed
for this recovery
ORA-00308: cannot open archived log
'/rman/pancake/logship/1_54_912576125.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

The final archive log reply reports an error, but this is normal. The error indicates that sqlplus was seeking a particular log file and did not find it. The reason is most likely that the log file does not yet exist.

If the source database can be shut down before copying archive logs, this step must be performed only once. The archive logs are copied and replayed, and then the process can continue directly to the cutover process that replicates the critical redo logs.

Incremental log replication and replay

In most cases, migration is not performed right away. It could be days or even weeks before the migration process is complete, which means that the logs must be continuously shipped to the replica database and replayed. Doing so makes sure that minimal data must be transferred and replayed when the cutover arrives.

This process can easily be scripted. For example, the following command can be scheduled on the original database to make sure that the location used for log shipping is continuously updated.

```
[oracle@jfsc1 pancake]$ cat copylogs.rman
configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
backup as copy archivelog from time 'sysdate-2';
```

```
[oracle@jfsc1 pancake]$ rman target / cmdfile=copylogs.rman
Recovery Manager: Release 12.1.0.2.0 - Production on Tue May 24 04:36:19
2016
Copyright (c) 1982, 2014, Oracle and/or its affiliates.  All rights
reserved.
connected to target database: PANCAKE (DBID=3574534589)
RMAN> configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
2> backup as copy archivelog from time 'sysdate-2';
3>
4>
using target database control file instead of recovery catalog
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters are successfully stored
Starting backup at 24-MAY-16
current log archived
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=369 device type=DISK
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=54 RECID=123 STAMP=912659482
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:22
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_54_912576125.dbf
continuing other job steps, job failed will not be re-run
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=41 RECID=124 STAMP=912659483
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:23
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_41_912576125.dbf
continuing other job steps, job failed will not be re-run
...
channel ORA_DISK_1: starting archived log copy
```

```
input archived log thread=1 sequence=45 RECID=152 STAMP=912659514
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:55
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_45_912576125.dbf
continuing other job steps, job failed will not be re-run
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=47 RECID=153 STAMP=912659515
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:57
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_47_912576125.dbf
Recovery Manager complete.
```

After the logs have been received, they must be replayed. Previous examples showed the use of sqlplus to manually run `recover database until cancel`, which can be easily automated. The example shown here uses the script described in Replay Logs on Standby Database. The script accepts an argument that specifies the database requiring a replay operation. This process permits the same script to be used in a multidatabase migration effort.

```
[root@jfsc2 pancake]# ./replaylogs.pl PANCAKE
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Tue May 24 04:47:10 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> ORA-00279: change 560591 generated at 05/24/2016 03:33:56 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_54_912576125.dbf
ORA-00280: change 560591 for thread 1 is in sequence #54
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 562219 generated at 05/24/2016 04:15:08 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_55_912576125.dbf
ORA-00280: change 562219 for thread 1 is in sequence #55
ORA-00278: log file '/rman/pancake/logship/1_54_912576125.dbf' no longer
needed for this recovery
ORA-00279: change 562370 generated at 05/24/2016 04:19:18 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_56_912576125.dbf
ORA-00280: change 562370 for thread 1 is in sequence #56
ORA-00278: log file '/rman/pancake/logship/1_55_912576125.dbf' no longer
needed for this recovery
...
ORA-00279: change 563137 generated at 05/24/2016 04:36:20 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_65_912576125.dbf
ORA-00280: change 563137 for thread 1 is in sequence #65
ORA-00278: log file '/rman/pancake/logship/1_64_912576125.dbf' no longer
needed for this recovery
ORA-00308: cannot open archived log
'/rman/pancake/logship/1_65_912576125.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
```

Cutover

When you are ready to cut over to the new environment, you must perform one final synchronization. When working with regular file systems, it is easy to make sure that the migrated database is 100% synchronized against the original because the original redo logs are copied and replayed. There is no good way to do this with ASM. Only the archive logs can be easily recopied. To make sure that no data is lost, the final shutdown of the original database must be performed carefully.

1. First, the database must be quiesced, ensuring that no changes are being made. This quiescing might include disabling scheduled operations, shutting down listeners, and/or shutting down applications.
2. After this step is taken, most DBAs create a dummy table to serve as a marker of the shutdown.
3. Force a log archiving to make sure that the creation of the dummy table is recorded within the archive logs. To do so, run the following commands:

```
SQL> create table cutovercheck as select * from dba_users;
Table created.
SQL> alter system archive log current;
System altered.
SQL> shutdown immediate;
Database closed.
Database dismounted.
ORACLE instance shut down.
```

4. To copy the last of the archive logs, run the following commands. The database must be available but not open.

```
SQL> startup mount;
ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             331353200 bytes
Database Buffers          465567744 bytes
Redo Buffers                5455872 bytes
Database mounted.
```

5. To copy the archive logs, run the following commands:

```
RMAN> configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
2> backup as copy archivelog from time 'sysdate-2';
3>
4>
using target database control file instead of recovery catalog
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters are successfully stored
Starting backup at 24-MAY-16
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=8 device type=DISK
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=54 RECID=123 STAMP=912659482
RMAN-03009: failure of backup command on ORA_DISK_1 channel at
05/24/2016 04:58:24
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_54_912576125.dbf
continuing other job steps, job failed will not be re-run
...
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=45 RECID=152 STAMP=912659514
RMAN-03009: failure of backup command on ORA_DISK_1 channel at
05/24/2016 04:58:58
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_45_912576125.dbf
continuing other job steps, job failed will not be re-run
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=47 RECID=153 STAMP=912659515
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03009: failure of backup command on ORA_DISK_1 channel at
05/24/2016 04:59:00
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_47_912576125.dbf
```

6. Finally, replay the remaining archive logs on the new server.

```
[root@jfsc2 pancake]# ./replaylogs.pl PANCAKE
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Tue May 24 05:00:53 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> ORA-00279: change 563137 generated at 05/24/2016 04:36:20 needed
for thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_65_912576125.dbf
ORA-00280: change 563137 for thread 1 is in sequence #65
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 563629 generated at 05/24/2016 04:55:20 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_66_912576125.dbf
ORA-00280: change 563629 for thread 1 is in sequence #66
ORA-00278: log file '/rman/pancake/logship/1_65_912576125.dbf' no longer
needed
for this recovery
ORA-00308: cannot open archived log
'/rman/pancake/logship/1_66_912576125.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
```

7. At this stage, replicate all data. The database is ready to be converted from a standby database to an active operational database and then opened.

```
SQL> alter database activate standby database;
Database altered.
SQL> alter database open;
Database altered.
```

8. Confirm the presence of the dummy table and then drop it.

```
SQL> desc cutovercheck
 Name                                      Null?    Type
 ----------------------------------------- --------
----------------------------
 USERNAME                                  NOT NULL VARCHAR2(128)
 USER_ID                                   NOT NULL NUMBER
 PASSWORD                                           VARCHAR2(4000)
 ACCOUNT_STATUS                            NOT NULL VARCHAR2(32)
 LOCK_DATE                                          DATE
 EXPIRY_DATE                                        DATE
 DEFAULT_TABLESPACE                        NOT NULL VARCHAR2(30)
 TEMPORARY_TABLESPACE                      NOT NULL VARCHAR2(30)
 CREATED                                   NOT NULL DATE
 PROFILE                                   NOT NULL VARCHAR2(128)
 INITIAL_RSRC_CONSUMER_GROUP                        VARCHAR2(128)
 EXTERNAL_NAME                                      VARCHAR2(4000)
 PASSWORD_VERSIONS                                  VARCHAR2(12)
 EDITIONS_ENABLED                                   VARCHAR2(1)
 AUTHENTICATION_TYPE                                VARCHAR2(8)
 PROXY_ONLY_CONNECT                                 VARCHAR2(1)
 COMMON                                             VARCHAR2(3)
 LAST_LOGIN                                         TIMESTAMP(9) WITH
TIME ZONE
 ORACLE_MAINTAINED                                  VARCHAR2(1)
SQL> drop table cutovercheck;
Table dropped.
```

Nondisruptive redo log migration

There are times when a database is correctly organized overall with the exception of the redo logs. This can happen for many reasons, the most common of which is related to snapshots. Products such as SnapManager for Oracle, SnapCenter, and the NetApp Snap Creator storage management framework enable near-instantaneous recovery of a database, but only if you revert the state of the data file volumes. If redo logs share space with the data files, reversion cannot be performed safely because it would result in destruction of the redo logs, likely meaning data loss. Therefore, the redo logs must be relocated.

This procedure is simple and can be performed nondisruptively.

Current redo log configuration

1. Identify the number of redo log groups and their respective group numbers.

```
SQL> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
---------------------------------------------------------------------------
--------
1 /redo0/NTAP/redo01a.log
1 /redo1/NTAP/redo01b.log
2 /redo0/NTAP/redo02a.log
2 /redo1/NTAP/redo02b.log
3 /redo0/NTAP/redo03a.log
3 /redo1/NTAP/redo03b.log
rows selected.
```

2. Enter the size of the redo logs.

```
SQL> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
---------------------------------------------------------------------------
--------
1 524288000
2 524288000
3 524288000
```

Create new logs

1. For each redo log, create a new group with a matching size and number of members.

```
SQL> alter database add logfile ('/newredo0/redo01a.log',
'/newredo1/redo01b.log') size 500M;
Database altered.
SQL> alter database add logfile ('/newredo0/redo02a.log',
'/newredo1/redo02b.log') size 500M;
Database altered.
SQL> alter database add logfile ('/newredo0/redo03a.log',
'/newredo1/redo03b.log') size 500M;
Database altered.
SQL>
```

2. Verify the new configuration.

```
SQL> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
---------------------------------------------------------------------------
--------
1 /redo0/NTAP/redo01a.log
1 /redo1/NTAP/redo01b.log
2 /redo0/NTAP/redo02a.log
2 /redo1/NTAP/redo02b.log
3 /redo0/NTAP/redo03a.log
3 /redo1/NTAP/redo03b.log
4 /newredo0/redo01a.log
4 /newredo1/redo01b.log
5 /newredo0/redo02a.log
5 /newredo1/redo02b.log
6 /newredo0/redo03a.log
6 /newredo1/redo03b.log
12 rows selected.
```

Drop old logs

1. Drop the old logs (groups 1, 2, and 3).

```
SQL> alter database drop logfile group 1;
Database altered.
SQL> alter database drop logfile group 2;
Database altered.
SQL> alter database drop logfile group 3;
Database altered.
```

2. If you encounter an error that prevents you from dropping an active log, force a switch to the next log to release the lock and force a global checkpoint. See the following example of this process. The attempt to drop logfile group 2, which was located on the old location, was denied because there was still active data in this logfile.

```
SQL> alter database drop logfile group 2;
alter database drop logfile group 2
*
ERROR at line 1:
ORA-01623: log 2 is current log for instance NTAP (thread 1) - cannot
drop
ORA-00312: online log 2 thread 1: '/redo0/NTAP/redo02a.log'
ORA-00312: online log 2 thread 1: '/redo1/NTAP/redo02b.log'
```

3. A log archiving followed by a checkpoint enables you to drop the logfile.

```
SQL> alter system archive log current;
System altered.
SQL> alter system checkpoint;
System altered.
SQL> alter database drop logfile group 2;
Database altered.
```

4. Then delete the logs from the file system. You should perform this process with extreme care.

Oracle database host data copy

As with database-level migration, migration at the host layer provides a storage vendor–independent approach.

In other words, sometime "just copy the files" is the best option.

Although this low-tech approach might seem too basic, it does offer significant benefits because no special software is required and the original data remains safely untouched during the process. The primary limitation is the fact that a file-copy data migration is a disruptive process, because the database must be shut down before the copy operation begins. There is no good way to synchronize changes within a file, so the files must be completely quiesced before copying begins.

If the shutdown required by a copy operation is not desirable, the next best host-based option is leveraging a logical volume manager (LVM). Many LVM options exist, including Oracle ASM, all with similar capabilities, but also with some limitations that must be taken into account. In most cases, the migration can be accomplished without downtime and disruption.

Filesystem to filesystem copying

The usefulness of a simple copy operation should not be underestimated. This operation requires downtime during the copy process, but it is a highly reliable process and requires no special expertise with operating systems, databases, or storage systems. Furthermore, it is very safe because it does not affect the original data. Typically, a system administrator changes the source file systems to be mounted as read-only and then reboots a server to guarantee that nothing can damage the current data. The copy process can be scripted to make sure that it runs as quickly as possible without risk of user error. Because the type of I/O is a simple sequential transfer of data, it is highly bandwidth efficient.

The following example demonstrates one option for a safe and rapid migration.

Environment

The environment to be migrated is as follows:

- Current file systems

```
ontap-nfs1:/host1_oradata       52428800  16196928  36231872  31%
/oradata
ontap-nfs1:/host1_logs          49807360    548032  49259328   2% /logs
```

- New file systems

```
ontap-nfs1:/host1_logs_new      49807360       128  49807232   1%
/new/logs
ontap-nfs1:/host1_oradata_new   49807360       128  49807232   1%
/new/oradata
```

Overview

The database can be migrated by a DBA by simply shutting down the database and copying the files, but the process is easily scripted if many databases must be migrated or minimizing downtime is critical. The use of scripts also reduces the chance for user error.

The example scripts shown automate the following operations:

- Shutting down the database
- Converting the existing file systems to a read-only state
- Copying all data from the source to target file systems, which preserves all file permissions
- Unmounting the old and new file systems
- Remounting the new file systems at the same paths as the prior file systems

Procedure

1. Shut down the database.

```
[root@host1 current]# ./dbshut.pl NTAP
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 15:58:48 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> Database closed.
Database dismounted.
ORACLE instance shut down.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP shut down
```

2. Convert the file systems to read-only. This can be done more quickly by using a script, as shown in Convert File System to Read Only.

```
[root@host1 current]# ./mk.fs.readonly.pl /oradata
/oradata unmounted
/oradata mounted read-only
[root@host1 current]# ./mk.fs.readonly.pl /logs
/logs unmounted
/logs mounted read-only
```

3. Confirm that the file systems are now read-only.

```
ontap-nfs1:/host1_oradata on /oradata type nfs
(ro,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
ontap-nfs1:/host1_logs on /logs type nfs
(ro,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
```

4. Synchronize file system contents with the `rsync` command.

```
[root@host1 current]# rsync -rlpogt --stats --progress
--exclude=.snapshot /oradata/ /new/oradata/
sending incremental file list
./
NTAP/
NTAP/IOPS.dbf
```

```
 10737426432 100%  153.50MB/s    0:01:06 (xfer#1, to-check=10/13)
NTAP/iops.dbf.zip
    22823573 100%   12.09MB/s    0:00:01 (xfer#2, to-check=9/13)
...
NTAP/undotbs02.dbf
  1073750016 100%  131.60MB/s    0:00:07 (xfer#10, to-check=1/13)
NTAP/users01.dbf
     5251072 100%    3.95MB/s    0:00:01 (xfer#11, to-check=0/13)
Number of files: 13
Number of files transferred: 11
Total file size: 18570092218 bytes
Total transferred file size: 18570092218 bytes
Literal data: 18570092218 bytes
Matched data: 0 bytes
File list size: 277
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 18572359828
Total bytes received: 228
sent 18572359828 bytes  received 228 bytes  162204017.96 bytes/sec
total size is 18570092218  speedup is 1.00
[root@host1 current]# rsync -rlpogt --stats --progress
--exclude=.snapshot /logs/ /new/logs/
sending incremental file list
./
NTAP/
NTAP/1_22_897068759.dbf
    45523968 100%   95.98MB/s    0:00:00 (xfer#1, to-check=15/18)
NTAP/1_23_897068759.dbf
    40601088 100%   49.45MB/s    0:00:00 (xfer#2, to-check=14/18)
...
NTAP/redo/redo02.log
    52429312 100%   44.68MB/s    0:00:01 (xfer#12, to-check=1/18)
NTAP/redo/redo03.log
    52429312 100%   68.03MB/s    0:00:00 (xfer#13, to-check=0/18)
Number of files: 18
Number of files transferred: 13
Total file size: 527032832 bytes
Total transferred file size: 527032832 bytes
Literal data: 527032832 bytes
Matched data: 0 bytes
File list size: 413
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 527098156
Total bytes received: 278
```

```
sent 527098156 bytes  received 278 bytes  95836078.91 bytes/sec
total size is 527032832  speedup is 1.00
```

5. Unmount the old file systems and relocate the copied data. This can be done more quickly by using a script, as shown in Replace File System.

```
[root@host1 current]# ./swap.fs.pl /logs,/new/logs
/new/logs unmounted
/logs unmounted
Updated /logs mounted
[root@host1 current]# ./swap.fs.pl /oradata,/new/oradata
/new/oradata unmounted
/oradata unmounted
Updated /oradata mounted
```

6. Confirm that the new file systems are in position.

```
ontap-nfs1:/host1_logs_new on /logs type nfs
(rw,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
ontap-nfs1:/host1_oradata_new on /oradata type nfs
(rw,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
```

7. Start the database.

```
[root@host1 current]# ./dbstart.pl NTAP
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 16:10:07 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             390073456 bytes
Database Buffers          406847488 bytes
Redo Buffers                5455872 bytes
Database mounted.
Database opened.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP started
```

Fully automated cutover

This sample script accepts arguments of the database SID followed by common-delimited pairs of file systems. For the example shown above, the command is issued as follows:

```
[root@host1 current]# ./migrate.oracle.fs.pl NTAP /logs,/new/logs
/oradata,/new/oradata
```

When executed, the example script attempts to perform the following sequence. It terminates if it encounters an error in any step:

1. Shut down the database.
2. Convert the current file systems to read-only status.
3. Use each comma-delimited pair of file system arguments and synchronize the first file system to the second.
4. Dismount the prior file systems.
5. Update the `/etc/fstab` file as follows:
   a. Create a backup at `/etc/fstab.bak`.
   b. Comment out the prior entries for the prior and new file systems.
   c. Create a new entry for the new file system that uses the old mountpoint.
6. Mount the file systems.
7. Start the database.

The following text provides an execution example for this script:

```
[root@host1 current]# ./migrate.oracle.fs.pl NTAP /logs,/new/logs
/oradata,/new/oradata
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 17:05:50 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> Database closed.
Database dismounted.
ORACLE instance shut down.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP shut down
sending incremental file list
```

```
./
NTAP/
NTAP/1_22_897068759.dbf
    45523968 100%  185.40MB/s    0:00:00 (xfer#1, to-check=15/18)
NTAP/1_23_897068759.dbf
    40601088 100%   81.34MB/s    0:00:00 (xfer#2, to-check=14/18)
...
NTAP/redo/redo02.log
    52429312 100%   70.42MB/s    0:00:00 (xfer#12, to-check=1/18)
NTAP/redo/redo03.log
    52429312 100%   47.08MB/s    0:00:01 (xfer#13, to-check=0/18)
Number of files: 18
Number of files transferred: 13
Total file size: 527032832 bytes
Total transferred file size: 527032832 bytes
Literal data: 527032832 bytes
Matched data: 0 bytes
File list size: 413
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 527098156
Total bytes received: 278
sent 527098156 bytes  received 278 bytes  150599552.57 bytes/sec
total size is 527032832  speedup is 1.00
Succesfully replicated filesystem /logs to /new/logs
sending incremental file list
./
NTAP/
NTAP/IOPS.dbf
 10737426432 100%  176.55MB/s    0:00:58 (xfer#1, to-check=10/13)
NTAP/iops.dbf.zip
    22823573 100%    9.48MB/s    0:00:02 (xfer#2, to-check=9/13)
... NTAP/undotbs01.dbf
   309338112 100%   70.76MB/s    0:00:04 (xfer#9, to-check=2/13)
NTAP/undotbs02.dbf
  1073750016 100%  187.65MB/s    0:00:05 (xfer#10, to-check=1/13)
NTAP/users01.dbf
     5251072 100%    5.09MB/s    0:00:00 (xfer#11, to-check=0/13)
Number of files: 13
Number of files transferred: 11
Total file size: 18570092218 bytes
Total transferred file size: 18570092218 bytes
Literal data: 18570092218 bytes
Matched data: 0 bytes
File list size: 277
File list generation time: 0.001 seconds
```

```
File list transfer time: 0.000 seconds
Total bytes sent: 18572359828
Total bytes received: 228
sent 18572359828 bytes  received 228 bytes  177725933.55 bytes/sec
total size is 18570092218  speedup is 1.00
Succesfully replicated filesystem /oradata to /new/oradata
swap 0 /logs /new/logs
/new/logs unmounted
/logs unmounted
Mounted updated /logs
Swapped filesystem /logs for /new/logs
swap 1 /oradata /new/oradata
/new/oradata unmounted
/oradata unmounted
Mounted updated /oradata
Swapped filesystem /oradata for /new/oradata
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 17:08:59 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             390073456 bytes
Database Buffers          406847488 bytes
Redo Buffers                5455872 bytes
Database mounted.
Database opened.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP started
[root@host1 current]#
```

Oracle ASM spfile and passwd migration

One difficulty in completing migration involving ASM is the ASM-specific spfile and the password file. By default, these critical metadata files are created on the first ASM disk group defined. If a particular ASM disk group must be evacuated and removed, the spfile and password file that govern that ASM instance must be relocated.

Another use case in which these files might need to be relocated is during a deployment of database management software such as SnapManager for Oracle or the SnapCenter Oracle plug- in. One of the features of these products is to rapidly restore a database by reverting the state of the ASM LUNs hosting the data files. Doing so requires taking the ASM disk group offline before performing a restore. This is not a problem as long as a given database's data files are isolated in a dedicated ASM disk group.

When that disk group also contains the ASM spfile/passwd file, the only way the disk group can be brought offline is to shut down the entire ASM instance. This is a disruptive process, which means that the spfile/passwd file would need to be relocated.

Environment

1. Database SID = TOAST
2. Current data files on `+DATA`
3. Current logfiles and controlfiles on `+LOGS`
4. New ASM disk groups established as `+NEWDATA` and `+NEWLOGS`

ASM spfile/passwd file locations

Relocating these files can be done nondisruptively. However, for safety, NetApp recommends shutting down the database environment so that you can be certain that the files have been relocated and the configuration is properly updated. This procedure must be repeated if multiple ASM instances are present on a server.

Identify ASM instances

Identify the ASM instances based on the data recorded in the `oratab` file. The ASM instances are denoted by a + symbol.

```
-bash-4.1$ cat /etc/oratab | grep '^+'
+ASM:/orabin/grid:N             # line added by Agent
```

There is one ASM instance called +ASM on this server.

Make sure all databases are shut down

The only smon process visible should be the smon for the ASM instance in use. The presence of another smon process indicates that a database is still running.

```
-bash-4.1$ ps -ef | grep smon
oracle     857     1  0 18:26 ?        00:00:00 asm_smon_+ASM
```

The only smon process is the ASM instance itself. This means that no other databases are running, and it is safe to proceed without risk of disrupting database operations.

Locate files

Identify the current location of the ASM spfile and password file by using the `spget` and `pwget` commands.

```
bash-4.1$ asmcmd
ASMCMD> spget
+DATA/spfile.ora
```

```
ASMCMD> pwget --asm
+DATA/orapwasm
```

The files are both located at the base of the `+DATA` disk group.

Copy files

Copy the files to the new ASM disk group with the `spcopy` and `pwcopy` commands. If the new disk group was recently created and is currently empty, it might need to be mounted first.

```
ASMCMD> mount NEWDATA
```

```
ASMCMD> spcopy +DATA/spfile.ora +NEWDATA/spfile.ora
copying +DATA/spfile.ora -> +NEWDATA/spfilea.ora
```

```
ASMCMD> pwcopy +DATA/orapwasm +NEWDATA/orapwasm
copying +DATA/orapwasm -> +NEWDATA/orapwasm
```

The files have now been copied from `+DATA` to `+NEWDATA`.

Update ASM instance

The ASM instance must now be updated to reflect the change in location. The `spset` and `pwset` commands update the ASM metadata required for starting the ASM disk group.

```
ASMCMD> spset +NEWDATA/spfile.ora
ASMCMD> pwset --asm +NEWDATA/orapwasm
```

Activate ASM using updated files

At this point, the ASM instance still uses the prior locations of these files. The instance must be restarted to force a reread of the files from their new locations and to release locks on the prior files.

```
-bash-4.1$ sqlplus / as sysasm
SQL> shutdown immediate;
ASM diskgroups volume disabled
ASM diskgroups dismounted
ASM instance shutdown
```

```
SQL> startup
ASM instance started
Total System Global Area 1140850688 bytes
Fixed Size                  2933400 bytes
Variable Size            1112751464 bytes
ASM Cache                  25165824 bytes
ORA-15032: not all alterations performed
ORA-15017: diskgroup "NEWDATA" cannot be mounted
ORA-15013: diskgroup "NEWDATA" is already mounted
```

Remove old spfile and password files

If the procedure has been performed successfully, the prior files are no longer locked and can now be removed.

```
-bash-4.1$ asmcmd
ASMCMD> rm +DATA/spfile.ora
ASMCMD> rm +DATA/orapwasm
```

Oracle ASM to ASM copy

Oracle ASM is essentially a lightweight combined volume manager and file system. Because the file system is not readily visible, RMAN must be used to perform copy operations. Although a copy-based migration process is safe and simple, it results in some disruption. The disruption can be minimized, but not fully eliminated.

If you want nondisruptive migration of an ASM-based database, the best option is to leverage ASM's capability to rebalance ASM extents to new LUNs while dropping the old LUNs. Doing so is generally safe and nondisruptive to operations, but it offers no back- out path. If functional or performance problems are encountered, the only option is to migrate the data back to the source.

This risk can be avoided by copying the database to the new location rather than moving data, so that the original data is untouched. The database can be fully tested in its new location before going live, and the original database is available as a fall- back option if problems are found.

This procedure is one of many options involving RMAN. It is designed to allow a two-step process in which the initial backup is created and then later synchronized through log replay. This process is desirable to minimize downtime because it allows the database to remain operational and serving data during the initial baseline copy.

Copy database

Oracle RMAN creates a level 0 (complete) copy of the source database currently located on the ASM disk group `+DATA` to the new location on `+NEWDATA`.

```
-bash-4.1$ rman target /
Recovery Manager: Release 12.1.0.2.0 - Production on Sun Dec 6 17:40:03
2015
Copyright (c) 1982, 2014, Oracle and/or its affiliates.  All rights
reserved.
connected to target database: TOAST (DBID=2084313411)
RMAN> backup as copy incremental level 0 database format '+NEWDATA' tag
'ONTAP_MIGRATION';
Starting backup at 06-DEC-15
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=302 device type=DISK
channel ORA_DISK_1: starting datafile copy
input datafile file number=00001
name=+DATA/TOAST/DATAFILE/system.262.897683141
...
input datafile file number=00004
name=+DATA/TOAST/DATAFILE/users.264.897683151
output file name=+NEWDATA/TOAST/DATAFILE/users.258.897759623
tag=ONTAP_MIGRATION RECID=5 STAMP=897759622
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting incremental level 0 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 06-DEC-15
channel ORA_DISK_1: finished piece 1 at 06-DEC-15
piece
handle=+NEWDATA/TOAST/BACKUPSET/2015_12_06/nnsnn0_ontap_migration_0.262.89
7759623 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 06-DEC-15
```

Force archive log switch

You must force an archive log switch to make sure that the archive logs contain all data required to make the copy fully consistent. Without this command, key data might still be present in the redo logs.

```
RMAN> sql 'alter system archive log current';
sql statement: alter system archive log current
```

Shut down source database

Disruption begins in this step because the database is shut down and placed in a limited-access, read-only mode. To shut down the source database, run the following commands:

```
RMAN> shutdown immediate;
using target database control file instead of recovery catalog
database closed
database dismounted
Oracle instance shut down
RMAN> startup mount;
connected to target database (not started)
Oracle instance started
database mounted
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                390073456 bytes
Database Buffers             406847488 bytes
Redo Buffers                   5455872 bytes
```

Controlfile backup

You must back up the controlfile in case you must abort the migration and revert to the original storage location. A copy of the backup controlfile isn't 100% required, but it does make the process of resetting the database file locations back to the original location easier.

```
RMAN> backup as copy current controlfile format '/tmp/TOAST.ctrl';
Starting backup at 06-DEC-15
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=358 device type=DISK
channel ORA_DISK_1: starting datafile copy
copying current control file
output file name=/tmp/TOAST.ctrl tag=TAG20151206T174753 RECID=6
STAMP=897760073
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
Finished backup at 06-DEC-15
```

Parameter updates

The current spfile contains references to the controlfiles on their current locations within the old ASM disk group. It must be edited, which is easily done by editing an intermediate pfile version.

```
RMAN> create pfile='/tmp/pfile' from spfile;
Statement processed
```

Update pfile

Update any parameters referring to old ASM disk groups to reflect the new ASM disk group names. Then save the updated pfile. Make sure that the `db_create` parameters are present.

In the example below, the references to `+DATA` that were changed to `+NEWDATA` are highlighted in yellow. Two key parameters are the `db_create` parameters that create any new files at the correct location.

```
*.compatible='12.1.0.2.0'
*.control_files='+NEWLOGS/TOAST/CONTROLFILE/current.258.897683139'
*.db_block_size=8192
*. db_create_file_dest='+NEWDATA'
*. db_create_online_log_dest_1='+NEWLOGS'
*.db_domain=''
*.db_name='TOAST'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=TOASTXDB)'
*.log_archive_dest_1='LOCATION=+NEWLOGS'
*.log_archive_format='%t_%s_%r.dbf'
```

Update init.ora file

Most ASM-based databases use an `init.ora` file located in the `$ORACLE_HOME/dbs` directory, which is a point to the spfile on the ASM disk group. This file must be redirected to a location on the new ASM disk group.

```
-bash-4.1$ cd $ORACLE_HOME/dbs
-bash-4.1$ cat initTOAST.ora
SPFILE='+DATA/TOAST/spfileTOAST.ora'
```

Change this file as follows:

```
SPFILE=+NEWLOGS/TOAST/spfileTOAST.ora
```

Parameter file recreation

The spfile is now ready to be populated by the data in the edited pfile.

```
RMAN> create spfile from pfile='/tmp/pfile';
Statement processed
```

Start database to start using new spfile

Start the database to make sure that it now uses the newly created spfile and that any further changes to system parameters are correctly recorded.

```
RMAN> startup nomount;
connected to target database (not started)
Oracle instance started
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                373296240 bytes
Database Buffers             423624704 bytes
Redo Buffers                   5455872 bytes
```

Restore controlfile

The backup controlfile created by RMAN can also be restored by RMAN directly to the location specified in the new spfile.

```
RMAN> restore controlfile from
'+DATA/TOAST/CONTROLFILE/current.258.897683139';
Starting restore at 06-DEC-15
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=417 device type=DISK
channel ORA_DISK_1: copied control file copy
output file name=+NEWLOGS/TOAST/CONTROLFILE/current.273.897761061
Finished restore at 06-DEC-15
```

Mount the database and verify the use of the new controlfile.

```
RMAN> alter database mount;
using target database control file instead of recovery catalog
Statement processed
```

```
SQL> show parameter control_files;
NAME                                 TYPE        VALUE
------------------------------------ -----------
------------------------------
control_files                        string
+NEWLOGS/TOAST/CONTROLFILE/cur
                                                 rent.273.897761061
```

Log replay

The database currently uses the data files in the old location. Before the copy can be used, they must be synchronized. Time has passed during the initial copy process, and the changes have been logged primarily in the archive logs. These changes are replicated as follows:

1. Perform an RMAN incremental backup, which contains the archive logs.

```
RMAN> backup incremental level 1 format '+NEWLOGS' for recover of copy
with tag 'ONTAP_MIGRATION' database;
Starting backup at 06-DEC-15
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=62 device type=DISK
channel ORA_DISK_1: starting incremental level 1 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
input datafile file number=00001
name=+DATA/TOAST/DATAFILE/system.262.897683141
input datafile file number=00002
name=+DATA/TOAST/DATAFILE/sysaux.260.897683143
input datafile file number=00003
name=+DATA/TOAST/DATAFILE/undotbs1.257.897683145
input datafile file number=00004
name=+DATA/TOAST/DATAFILE/users.264.897683151
channel ORA_DISK_1: starting piece 1 at 06-DEC-15
channel ORA_DISK_1: finished piece 1 at 06-DEC-15
piece
handle=+NEWLOGS/TOAST/BACKUPSET/2015_12_06/nnndn1_ontap_migration_0.268.
897762693 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting incremental level 1 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current control file in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 06-DEC-15
channel ORA_DISK_1: finished piece 1 at 06-DEC-15
piece
handle=+NEWLOGS/TOAST/BACKUPSET/2015_12_06/ncsnn1_ontap_migration_0.267.
897762697 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 06-DEC-15
```

2. Replay the log.

```
RMAN> recover copy of database with tag 'ONTAP_MIGRATION';
Starting recover at 06-DEC-15
using channel ORA_DISK_1
channel ORA_DISK_1: starting incremental datafile backup set restore
channel ORA_DISK_1: specifying datafile copies to recover
recovering datafile copy file number=00001
name=+NEWDATA/TOAST/DATAFILE/system.259.897759609
recovering datafile copy file number=00002
name=+NEWDATA/TOAST/DATAFILE/sysaux.263.897759615
recovering datafile copy file number=00003
name=+NEWDATA/TOAST/DATAFILE/undotbs1.264.897759619
recovering datafile copy file number=00004
name=+NEWDATA/TOAST/DATAFILE/users.258.897759623
channel ORA_DISK_1: reading from backup piece
+NEWLOGS/TOAST/BACKUPSET/2015_12_06/nnndn1_ontap_migration_0.268.8977626
93
channel ORA_DISK_1: piece
handle=+NEWLOGS/TOAST/BACKUPSET/2015_12_06/nnndn1_ontap_migration_0.268.
897762693 tag=ONTAP_MIGRATION
channel ORA_DISK_1: restored backup piece 1
channel ORA_DISK_1: restore complete, elapsed time: 00:00:01
Finished recover at 06-DEC-15
```

Activation

The controlfile that was restored still references the data files at the original location, and it also contains the path information for the copied data files.

1. To change the active data files, run the `switch database to copy` command.

```
RMAN> switch database to copy;
datafile 1 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/system.259.897759609"
datafile 2 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/sysaux.263.897759615"
datafile 3 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/undotbs1.264.897759619"
datafile 4 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/users.258.897759623"
```

   The active data files are now the copied data files, but there still might be changes contained within the final redo logs.

2. To replay all of the remaining logs, run the `recover database` command. If the message `media recovery complete` appears, the process was successful.

```
RMAN> recover database;
Starting recover at 06-DEC-15
using channel ORA_DISK_1
starting media recovery
media recovery complete, elapsed time: 00:00:01
Finished recover at 06-DEC-15
```

This process only changed the location of the normal data files. The temporary data files must be renamed, but they do not need to be copied because they are only temporary. The database is currently down, so there is no active data in the temporary data files.

3. To relocate the temporary data files, first identify their location.

```
RMAN> select file#||' '||name from v$tempfile;
FILE#||''||NAME
---------------------------------------------------------------------------
--------
1 +DATA/TOAST/TEMPFILE/temp.263.897683145
```

4. Relocate temporary data files by using an RMAN command that sets the new name for each data file. With Oracle Managed Files (OMF), the complete name is not necessary; the ASM disk group is sufficient. When the database is opened, OMF links to the appropriate location on the ASM disk group. To relocate files, run the following commands:

```
run {
set newname for tempfile 1 to '+NEWDATA';
switch tempfile all;
}
```

```
RMAN> run {
2> set newname for tempfile 1 to '+NEWDATA';
3> switch tempfile all;
4> }
executing command: SET NEWNAME
renamed tempfile 1 to +NEWDATA in control file
```

Redo log migration

The migration process is nearly complete, but the redo logs are still located on the original ASM disk group. Redo logs cannot be directly relocated. Instead, a new set of redo logs is created and added to the configuration, followed by a drop of the old logs.

1. Identify the number of redo log groups and their respective group numbers.

```
RMAN> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
---------------------------------------------------------------------------
--------
1 +DATA/TOAST/ONLINELOG/group_1.261.897683139
2 +DATA/TOAST/ONLINELOG/group_2.259.897683139
3 +DATA/TOAST/ONLINELOG/group_3.256.897683139
```

2. Enter the size of the redo logs.

```
RMAN> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
---------------------------------------------------------------------------
--------
1 52428800
2 52428800
3 52428800
```

3. For each redo log, create a new group with a matching configuration. If you are not using OMF, you must specify the full path. This is also an example that uses the `db_create_online_log` parameters. As was shown previously, this parameter was set to +NEWLOGS. This configuration allows you to use the following commands to create new online logs without the need to specify a file location or even a specific ASM disk group.

```
RMAN> alter database add logfile size 52428800;
Statement processed
RMAN> alter database add logfile size 52428800;
Statement processed
RMAN> alter database add logfile size 52428800;
Statement processed
```

4. Open the database.

```
SQL> alter database open;
Database altered.
```

5. Drop the old logs.

```
RMAN> alter database drop logfile group 1;
Statement processed
```

6. If you encounter an error that prevents you from dropping an active log, force a switch to the next log to

release the lock and force a global checkpoint. An example is shown below. The attempt to drop logfile group 3, which was located on the old location, was denied because there was still active data in this logfile. A log archiving following a checkpoint allows you to delete the logfile.

```
RMAN> alter database drop logfile group 3;
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03002: failure of sql statement command at 12/08/2015 20:23:51
ORA-01623: log 3 is current log for instance TOAST (thread 4) - cannot
drop
ORA-00312: online log 3 thread 1:
'+LOGS/TOAST/ONLINELOG/group_3.259.897563549'
RMAN> alter system switch logfile;
Statement processed
RMAN> alter system checkpoint;
Statement processed
RMAN> alter database drop logfile group 3;
Statement processed
```

7. Review the environment to make sure that all location-based parameters are updated.

```
SQL> select name from v$datafile;
SQL> select member from v$logfile;
SQL> select name from v$tempfile;
SQL> show parameter spfile;
SQL> select name, value from v$parameter where value is not null;
```

8. The following script demonstrates how to simplify this process:

```
[root@host1 current]# ./checkdbdata.pl TOAST
TOAST datafiles:
+NEWDATA/TOAST/DATAFILE/system.259.897759609
+NEWDATA/TOAST/DATAFILE/sysaux.263.897759615
+NEWDATA/TOAST/DATAFILE/undotbs1.264.897759619
+NEWDATA/TOAST/DATAFILE/users.258.897759623
TOAST redo logs:
+NEWLOGS/TOAST/ONLINELOG/group_4.266.897763123
+NEWLOGS/TOAST/ONLINELOG/group_5.265.897763125
+NEWLOGS/TOAST/ONLINELOG/group_6.264.897763125
TOAST temp datafiles:
+NEWDATA/TOAST/TEMPFILE/temp.260.897763165
TOAST spfile
spfile                                   string
+NEWDATA/spfiletoast.ora
TOAST key parameters
control_files +NEWLOGS/TOAST/CONTROLFILE/current.273.897761061
log_archive_dest_1 LOCATION=+NEWLOGS
db_create_file_dest +NEWDATA
db_create_online_log_dest_1 +NEWLOGS
```

9. If the ASM disk groups were completely evacuated, they can now be unmounted with `asmcmd`. However, in many cases the files belonging to other databases or the ASM spfile/passwd file might still be present.

```
-bash-4.1$ . oraenv
ORACLE_SID = [TOAST] ? +ASM
The Oracle base remains unchanged with value /orabin
-bash-4.1$ asmcmd
ASMCMD> umount DATA
ASMCMD>
```

Oracle ASM to file system copy

The Oracle ASM to file system copy procedure is very similar to the ASM to ASM copy procedure, with similar benefits and restrictions. The primary difference is the syntax of the various commands and configuration parameters when using a visible file system as opposed to an ASM disk group.

Copy database

Oracle RMAN is used to create a level 0 (complete) copy of the source database currently located on the ASM disk group `+DATA` to the new location on `/oradata`.

```
RMAN> backup as copy incremental level 0 database format
'/oradata/TOAST/%U' tag 'ONTAP_MIGRATION';
Starting backup at 13-MAY-16
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=377 device type=DISK
channel ORA_DISK_1: starting datafile copy
input datafile file number=00001 name=+ASM0/TOAST/system01.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSTEM_FNO-
1_01r5fhjg tag=ONTAP_MIGRATION RECID=1 STAMP=911722099
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:07
channel ORA_DISK_1: starting datafile copy
input datafile file number=00002 name=+ASM0/TOAST/sysaux01.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSAUX_FNO-
2_02r5fhjo tag=ONTAP_MIGRATION RECID=2 STAMP=911722106
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:07
channel ORA_DISK_1: starting datafile copy
input datafile file number=00003 name=+ASM0/TOAST/undotbs101.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-UNDOTBS1_FNO-
3_03r5fhjt tag=ONTAP_MIGRATION RECID=3 STAMP=911722113
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:07
channel ORA_DISK_1: starting datafile copy
copying current control file
output file name=/oradata/TOAST/cf_D-TOAST_id-2098173325_04r5fhk5
tag=ONTAP_MIGRATION RECID=4 STAMP=911722118
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting datafile copy
input datafile file number=00004 name=+ASM0/TOAST/users01.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-USERS_FNO-
4_05r5fhk6 tag=ONTAP_MIGRATION RECID=5 STAMP=911722118
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting incremental level 0 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 13-MAY-16
channel ORA_DISK_1: finished piece 1 at 13-MAY-16
piece handle=/oradata/TOAST/06r5fhk7_1_1 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 13-MAY-16
```

Force archive log switch

Forcing the archive log switch is required to make sure that the archive logs contain all of the data required to make the copy fully consistent. Without this command, key data might still be present in the redo logs. To force an archive log switch, run the following command:

```
RMAN> sql 'alter system archive log current';
sql statement: alter system archive log current
```

Shut down source database

Disruption begins in this step because the database is shut down and placed in a limited-access read-only mode. To shut down the source database, run the following commands:

```
RMAN> shutdown immediate;
using target database control file instead of recovery catalog
database closed
database dismounted
Oracle instance shut down
RMAN> startup mount;
connected to target database (not started)
Oracle instance started
database mounted
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                331353200 bytes
Database Buffers             465567744 bytes
Redo Buffers                   5455872 bytes
```

Controlfile backup

Back up controlfiles in case you must abort the migration and revert to the original storage location. A copy of the backup controlfile isn't 100% required, but it does make the process of resetting the database file locations back to the original location easier.

```
RMAN> backup as copy current controlfile format '/tmp/TOAST.ctrl';
Starting backup at 08-DEC-15
using channel ORA_DISK_1
channel ORA_DISK_1: starting datafile copy
copying current control file
output file name=/tmp/TOAST.ctrl tag=TAG20151208T194540 RECID=30
STAMP=897939940
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
Finished backup at 08-DEC-15
```

Parameter updates

```
RMAN> create pfile='/tmp/pfile' from spfile;
Statement processed
```

Update pfile

Any parameters referring to old ASM disk groups should be updated and, in some cases, deleted when they are no longer relevant. Update them to reflect the new file system paths and save the updated pfile. Make sure that the complete target path is listed. To update these parameters, run the following commands:

```
*.audit_file_dest='/orabin/admin/TOAST/adump'
*.audit_trail='db'
*.compatible='12.1.0.2.0'
*.control_files='/logs/TOAST/arch/control01.ctl','/logs/TOAST/redo/control
02.ctl'
*.db_block_size=8192
*.db_domain=''
*.db_name='TOAST'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=TOASTXDB)'
*.log_archive_dest_1='LOCATION=/logs/TOAST/arch'
*.log_archive_format='%t_%s_%r.dbf'
*.open_cursors=300
*.pga_aggregate_target=256m
*.processes=300
*.remote_login_passwordfile='EXCLUSIVE'
*.sga_target=768m
*.undo_tablespace='UNDOTBS1'
```

Disable the original init.ora file

This file is located in the `$ORACLE_HOME/dbs` directory and is usually in a pfile that serves as a pointer to the spfile on the ASM disk group. To make sure that the original spfile is no longer used, rename it. Do not delete it, however, because this file is needed if the migration must be aborted.

```
[oracle@jfsc1 ~]$ cd $ORACLE_HOME/dbs
[oracle@jfsc1 dbs]$ cat initTOAST.ora
SPFILE='+ASM0/TOAST/spfileTOAST.ora'
[oracle@jfsc1 dbs]$ mv initTOAST.ora initTOAST.ora.prev
[oracle@jfsc1 dbs]$
```

Parameter file recreation

This is the final step in spfile relocation. The original spfile is no longer used and the database is currently started (but not mounted) using the intermediate file. The contents of this file can be written out to the new spfile location as follows:

```
RMAN> create spfile from pfile='/tmp/pfile';
Statement processed
```

Start database to start using new spfile

You must start the database to release the locks on the intermediate file and start the database by using only the new spfile file. Starting the database also proves that the new spfile location is correct and its data is valid.

```
RMAN> shutdown immediate;
Oracle instance shut down
RMAN> startup nomount;
connected to target database (not started)
Oracle instance started
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                331353200 bytes
Database Buffers             465567744 bytes
Redo Buffers                   5455872 bytes
```

Restore controlfile

A backup controlfile was created at the path `/tmp/TOAST.ctrl` earlier in the procedure. The new spfile defines the controlfile locations as `/logfs/TOAST/ctrl/ctrlfile1.ctrl` and `/logfs/TOAST/redo/ctrlfile2.ctrl`. However, those files do not yet exist.

1. This command restores the controlfile data to the paths defined in the spfile.

   ```
   RMAN> restore controlfile from '/tmp/TOAST.ctrl';
   Starting restore at 13-MAY-16
   using channel ORA_DISK_1
   channel ORA_DISK_1: copied control file copy
   output file name=/logs/TOAST/arch/control01.ctl
   output file name=/logs/TOAST/redo/control02.ctl
   Finished restore at 13-MAY-16
   ```

2. Issue the mount command so that the controlfiles are discovered correctly and contain valid data.

   ```
   RMAN> alter database mount;
   Statement processed
   released channel: ORA_DISK_1
   ```

   To validate the `control_files` parameter, run the following command:

```
SQL> show parameter control_files;
NAME                                 TYPE        VALUE
------------------------------------ -----------
------------------------------
control_files                        string
/logs/TOAST/arch/control01.ctl
                                                 ,
/logs/TOAST/redo/control02.c
                                                 tl
```

Log replay

The database is currently using the data files in the old location. Before the copy can be used, the data files must be synchronized. Time has passed during the initial copy process, and changes were logged primarily in the archive logs. These changes are replicated in the following two steps.

1. Perform an RMAN incremental backup, which contains the archive logs.

```
RMAN>  backup incremental level 1 format '/logs/TOAST/arch/%U' for
recover of copy with tag 'ONTAP_MIGRATION' database;
Starting backup at 13-MAY-16
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=124 device type=DISK
channel ORA_DISK_1: starting incremental level 1 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
input datafile file number=00001 name=+ASM0/TOAST/system01.dbf
input datafile file number=00002 name=+ASM0/TOAST/sysaux01.dbf
input datafile file number=00003 name=+ASM0/TOAST/undotbs101.dbf
input datafile file number=00004 name=+ASM0/TOAST/users01.dbf
channel ORA_DISK_1: starting piece 1 at 13-MAY-16
channel ORA_DISK_1: finished piece 1 at 13-MAY-16
piece handle=/logs/TOAST/arch/09r5fj8i_1_1 tag=ONTAP_MIGRATION
comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 13-MAY-16
RMAN-06497: WARNING: control file is not current, control file
AUTOBACKUP skipped
```

2. Replay the logs.

```
RMAN> recover copy of database with tag 'ONTAP_MIGRATION';
Starting recover at 13-MAY-16
using channel ORA_DISK_1
channel ORA_DISK_1: starting incremental datafile backup set restore
channel ORA_DISK_1: specifying datafile copies to recover
recovering datafile copy file number=00001 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-SYSTEM_FNO-1_01r5fhjg
recovering datafile copy file number=00002 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-SYSAUX_FNO-2_02r5fhjo
recovering datafile copy file number=00003 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-UNDOTBS1_FNO-3_03r5fhjt
recovering datafile copy file number=00004 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-USERS_FNO-4_05r5fhk6
channel ORA_DISK_1: reading from backup piece
/logs/TOAST/arch/09r5fj8i_1_1
channel ORA_DISK_1: piece handle=/logs/TOAST/arch/09r5fj8i_1_1
tag=ONTAP_MIGRATION
channel ORA_DISK_1: restored backup piece 1
channel ORA_DISK_1: restore complete, elapsed time: 00:00:01
Finished recover at 13-MAY-16
RMAN-06497: WARNING: control file is not current, control file
AUTOBACKUP skipped
```

Activation

The controlfile that was restored still references the data files at the original location, and it also contains the path information for the copied data files.

1. To change the active data files, run the `switch database to copy` command:

```
RMAN> switch database to copy;
datafile 1 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-SYSTEM_FNO-1_01r5fhjg"
datafile 2 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-SYSAUX_FNO-2_02r5fhjo"
datafile 3 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-UNDOTBS1_FNO-3_03r5fhjt"
datafile 4 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-USERS_FNO-4_05r5fhk6"
```

2. Although the data files should be fully consistent, one final step is required to replay the remaining changes recorded in the online redo logs. Use the `recover database` command to replay these changes and make the copy 100% identical to the original. The copy is not yet open, however.

```
RMAN> recover database;
Starting recover at 13-MAY-16
using channel ORA_DISK_1
starting media recovery
archived log for thread 1 with sequence 28 is already on disk as file
+ASM0/TOAST/redo01.log
archived log file name=+ASM0/TOAST/redo01.log thread=1 sequence=28
media recovery complete, elapsed time: 00:00:00
Finished recover at 13-MAY-16
```

Relocate Temporary Data Files

1. Identify the location of temporary data files still in use on the original disk group.

```
RMAN> select file#||' '||name from v$tempfile;
FILE#||''||NAME
------------------------------------------------------------------------
--------
1 +ASM0/TOAST/temp01.dbf
```

2. To relocate the data files, run the following commands. If there are many tempfiles, use a text editor to create the RMAN command and then cut and paste it.

```
RMAN> run {
2> set newname for tempfile 1 to '/oradata/TOAST/temp01.dbf';
3> switch tempfile all;
4> }
executing command: SET NEWNAME
renamed tempfile 1 to /oradata/TOAST/temp01.dbf in control file
```

Redo log migration

The migration process is nearly complete, but the redo logs are still located on the original ASM disk group. Redo logs cannot be directly relocated. Instead, a new set of redo logs is created and added to the configuration, following by a drop of the old logs.

1. Identify the number of redo log groups and their respective group numbers.

```
RMAN> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
---------------------------------------------------------------------------
--------
1 +ASM0/TOAST/redo01.log
2 +ASM0/TOAST/redo02.log
3 +ASM0/TOAST/redo03.log
```

2. Enter the size of the redo logs.

```
RMAN> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
---------------------------------------------------------------------------
--------
1 52428800
2 52428800
3 52428800
```

3. For each redo log, create a new group by using the same size as the current redo log group using the new file system location.

```
RMAN> alter database add logfile '/logs/TOAST/redo/log00.rdo' size
52428800;
Statement processed
RMAN> alter database add logfile '/logs/TOAST/redo/log01.rdo' size
52428800;
Statement processed
RMAN> alter database add logfile '/logs/TOAST/redo/log02.rdo' size
52428800;
Statement processed
```

4. Remove the old logfile groups that are still located on the prior storage.

```
RMAN> alter database drop logfile group 4;
Statement processed
RMAN> alter database drop logfile group 5;
Statement processed
RMAN> alter database drop logfile group 6;
Statement processed
```

5. If an error is encountered that blocks dropping an active log, force a switch to the next log to release the lock and force a global checkpoint. An example is shown below. The attempt to drop logfile group 3, which was located on the old location, was denied because there was still active data in this logfile. A log

archiving followed by a checkpoint enables logfile deletion.

```
RMAN> alter database drop logfile group 4;
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03002: failure of sql statement command at 12/08/2015 20:23:51
ORA-01623: log 4 is current log for instance TOAST (thread 4) - cannot
drop
ORA-00312: online log 4 thread 1:
'+NEWLOGS/TOAST/ONLINELOG/group_4.266.897763123'
RMAN> alter system switch logfile;
Statement processed
RMAN> alter system checkpoint;
Statement processed
RMAN> alter database drop logfile group 4;
Statement processed
```

6. Review the environment to make sure that all location-based parameters are updated.

```
SQL> select name from v$datafile;
SQL> select member from v$logfile;
SQL> select name from v$tempfile;
SQL> show parameter spfile;
SQL> select name, value from v$parameter where value is not null;
```

7. The following script demonstrates how to make this process easier.

```
[root@jfsc1 current]# ./checkdbdata.pl TOAST
TOAST datafiles:
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSTEM_FNO-1_01r5fhjg
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSAUX_FNO-2_02r5fhjo
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-UNDOTBS1_FNO-3_03r5fhjt
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-USERS_FNO-4_05r5fhk6
TOAST redo logs:
/logs/TOAST/redo/log00.rdo
/logs/TOAST/redo/log01.rdo
/logs/TOAST/redo/log02.rdo
TOAST temp datafiles:
/oradata/TOAST/temp01.dbf
TOAST spfile
spfile                                  string
/orabin/product/12.1.0/dbhome_
                                                        1/dbs/spfileTOAST.ora
TOAST key parameters
control_files /logs/TOAST/arch/control01.ctl,
/logs/TOAST/redo/control02.ctl
log_archive_dest_1 LOCATION=/logs/TOAST/arch
```

8. If the ASM disk groups were completely evacuated, they can now be unmounted with `asmcmd`. In many cases, files belonging to other databases or the ASM spfile/passwd file can still be present.

```
-bash-4.1$ . oraenv
ORACLE_SID = [TOAST] ? +ASM
The Oracle base remains unchanged with value /orabin
-bash-4.1$ asmcmd
ASMCMD> umount DATA
ASMCMD>
```

Data file cleanup procedure

The migration process might result in data files with long or cryptic syntax, depending on how Oracle RMAN was used. In the example shown here, the backup was performed with the file format of `/oradata/TOAST/%U`. `%U` indicates that RMAN should create a default unique name for each data file. The result is similar to what is shown in the following text. The traditional names for the data files are embedded within the names. This can be cleaned up by using the scripted approach shown in ASM Migration Cleanup.

```
[root@jfsc1 current]# ./fixuniquenames.pl TOAST
#sqlplus Commands
shutdown immediate;
startup mount;
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSTEM_FNO-1_01r5fhjg
/oradata/TOAST/system.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSAUX_FNO-2_02r5fhjo
/oradata/TOAST/sysaux.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-UNDOTBS1_FNO-
3_03r5fhjt /oradata/TOAST/undotbs1.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-USERS_FNO-4_05r5fhk6
/oradata/TOAST/users.dbf
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
SYSTEM_FNO-1_01r5fhjg' to '/oradata/TOAST/system.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
SYSAUX_FNO-2_02r5fhjo' to '/oradata/TOAST/sysaux.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
UNDOTBS1_FNO-3_03r5fhjt' to '/oradata/TOAST/undotbs1.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
USERS_FNO-4_05r5fhk6' to '/oradata/TOAST/users.dbf';
alter database open;
```

Oracle ASM rebalance

As discussed previously, an Oracle ASM disk group can be transparently migrated to a new storage system by using the rebalancing process. In summary, the rebalancing process requires the addition of equal-sized LUNs to the existing group of LUNs followed by a drop operation of the prior LUN. Oracle ASM automatically relocates the underlying data to new storage in an optimal layout and then releases the old LUNs when complete.

The migration process uses efficient sequential I/O and does not generally cause any performance disruption, but the migration rate can be throttled when needed.

Identify data to be migrated

```
SQL> select name||' '||group_number||' '||total_mb||' '||path||'
'||header_status from v$asm_disk;
NEWDATA_0003 1 10240 /dev/mapper/3600a098038303537762b47594c315864 MEMBER
NEWDATA_0002 1 10240 /dev/mapper/3600a098038303537762b47594c315863 MEMBER
NEWDATA_0000 1 10240 /dev/mapper/3600a098038303537762b47594c315861 MEMBER
NEWDATA_0001 1 10240 /dev/mapper/3600a098038303537762b47594c315862 MEMBER
SQL> select group_number||' '||name from v$asm_diskgroup;
1 NEWDATA
```

Create new LUNs

Create new LUNs of the same size, and set user and group membership as required. The LUNs should appear as `CANDIDATE` disks.

```
SQL> select name||' '||group_number||' '||total_mb||' '||path||'
'||header_status from v$asm_disk;
 0 0 /dev/mapper/3600a098038303537762b47594c31586b CANDIDATE
 0 0 /dev/mapper/3600a098038303537762b47594c315869 CANDIDATE
 0 0 /dev/mapper/3600a098038303537762b47594c315858 CANDIDATE
 0 0 /dev/mapper/3600a098038303537762b47594c31586a CANDIDATE
NEWDATA_0003 1 10240 /dev/mapper/3600a098038303537762b47594c315864 MEMBER
NEWDATA_0002 1 10240 /dev/mapper/3600a098038303537762b47594c315863 MEMBER
NEWDATA_0000 1 10240 /dev/mapper/3600a098038303537762b47594c315861 MEMBER
NEWDATA_0001 1 10240 /dev/mapper/3600a098038303537762b47594c315862 MEMBER
```

Add new LUNS

While the add and drop operations can be performed together, it is generally easier to add new LUNs in two steps. First, add the new LUNs to the disk group. This step results in half of the extents being migrated from the current ASM LUNs to the new LUNs.

The rebalance power indicates the rate at which data is being transferred. The higher the number, the higher the parallelism of the data transfer. The migration is performed with efficient sequential I/O operations that are unlikely to cause performance problems. However, if desired, the rebalance power of an ongoing migration can be adjusted with the `alter diskgroup [name] rebalance power [level]` command. Typical migrations use a value of 5.

```
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c31586b' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c315869' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c315858' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c31586a' rebalance power 5;
Diskgroup altered.
```

Monitor operation

A rebalancing operation can be monitored and managed in multiple ways. We used the following command for this example.

```
SQL> select group_number,operation,state from v$asm_operation;
GROUP_NUMBER OPERA STAT
------------ ----- ----
           1 REBAL RUN
           1 REBAL WAIT
```

When migration is complete, no rebalancing operations are reported.

```
SQL> select group_number,operation,state from v$asm_operation;
no rows selected
```

Drop old LUNs

The migration is now halfway complete. It might be desirable to perform some basic performance tests to make sure that the environment is healthy. After confirmation, the remaining data can be relocated by dropping the old LUNs. Note that this does not result in immediate release of the LUNs. The drop operation signals Oracle ASM to relocate the extents first and then release the LUN.

```
sqlplus / as sysasm
SQL> alter diskgroup NEWDATA drop disk NEWDATA_0000 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA drop disk NEWDATA_0001 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup newdata drop disk NEWDATA_0002 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup newdata drop disk NEWDATA_0003 rebalance power 5;
Diskgroup altered.
```

Monitor operation

The rebalancing operation can be monitored and managed in multiple ways. We used the following command for this example:

```
SQL> select group_number,operation,state from v$asm_operation;
GROUP_NUMBER OPERA STAT
------------ ----- ----
           1 REBAL RUN
           1 REBAL WAIT
```

When migration is complete, no rebalancing operations are reported.

```
SQL> select group_number,operation,state from v$asm_operation;
no rows selected
```

Remove old LUNs

Before you remove the old LUNs from the disk group, you should perform one final check on the header status. After a LUN is released from ASM, it no longer has a name listed and the header status is listed as `FORMER`. This indicates that these LUNs can safely be removed from the system.

```
SQL> select name||' '||group_number||' '||total_mb||' '||path||'
'||header_status from v$asm_disk;
NAME||''||GROUP_NUMBER||''||TOTAL_MB||''||PATH||''||HEADER_STATUS
------------------------------------------------------------------------
------
 0 0 /dev/mapper/3600a098038303537762b47594c315863 FORMER
 0 0 /dev/mapper/3600a098038303537762b47594c315864 FORMER
 0 0 /dev/mapper/3600a098038303537762b47594c315861 FORMER
 0 0 /dev/mapper/3600a098038303537762b47594c315862 FORMER
NEWDATA_0005 1 10240 /dev/mapper/3600a098038303537762b47594c315869 MEMBER
NEWDATA_0007 1 10240 /dev/mapper/3600a098038303537762b47594c31586a MEMBER
NEWDATA_0004 1 10240 /dev/mapper/3600a098038303537762b47594c31586b MEMBER
NEWDATA_0006 1 10240 /dev/mapper/3600a098038303537762b47594c315858 MEMBER
8 rows selected.
```

LVM migration

The procedure presented here shows the principles of an LVM-based migration of a volume group called `datavg`. The examples are drawn from the Linux LVM, but the principles apply equally to AIX, HP-UX, and VxVM. The precise commands might vary.

1. Identify the LUNs currently in the `datavg` volume group.

```
[root@host1 ~]# pvdisplay -C | grep datavg
  /dev/mapper/3600a098038303537762b47594c31582f datavg lvm2 a--  10.00g
10.00g
  /dev/mapper/3600a098038303537762b47594c31585a datavg lvm2 a--  10.00g
10.00g
  /dev/mapper/3600a098038303537762b47594c315859 datavg lvm2 a--  10.00g
10.00g
  /dev/mapper/3600a098038303537762b47594c31586c datavg lvm2 a--  10.00g
10.00g
```

2. Create new LUNs of the same or slightly larger physical size and define them as physical volumes.

```
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315864
  Physical volume "/dev/mapper/3600a098038303537762b47594c315864"
successfully created
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315863
  Physical volume "/dev/mapper/3600a098038303537762b47594c315863"
successfully created
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315862
  Physical volume "/dev/mapper/3600a098038303537762b47594c315862"
successfully created
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315861
  Physical volume "/dev/mapper/3600a098038303537762b47594c315861"
successfully created
```

3. Add the new volumes to the volume group.

```
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315864
  Volume group "datavg" successfully extended
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315863
  Volume group "datavg" successfully extended
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315862
  Volume group "datavg" successfully extended
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315861
  Volume group "datavg" successfully extended
```

4. Issue the `pvmove` command to relocate the extents of each current LUN to the new LUN. The `- i [seconds]` argument monitors the progress of the operation.

```
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c31582f
/dev/mapper/3600a098038303537762b47594c315864
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 14.2%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 28.4%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 42.5%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 57.1%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 72.3%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 87.3%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 100.0%
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c31585a
/dev/mapper/3600a098038303537762b47594c315863
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 14.9%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 29.9%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 44.8%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 60.1%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 75.8%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 90.9%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 100.0%
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c315859
/dev/mapper/3600a098038303537762b47594c315862
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 14.8%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 29.8%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 45.5%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 61.1%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 76.6%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 91.7%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 100.0%
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c31586c
/dev/mapper/3600a098038303537762b47594c315861
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 15.0%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 30.4%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 46.0%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 61.4%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 77.2%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 92.3%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 100.0%
```

5. When this process is complete, drop the old LUNs from the volume group by using the `vgreduce` command. If successful, the LUN can now be safely removed from the system.

```
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c31582f
Removed "/dev/mapper/3600a098038303537762b47594c31582f" from volume
group "datavg"
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c31585a
  Removed "/dev/mapper/3600a098038303537762b47594c31585a" from volume
group "datavg"
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c315859
  Removed "/dev/mapper/3600a098038303537762b47594c315859" from volume
group "datavg"
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c31586c
  Removed "/dev/mapper/3600a098038303537762b47594c31586c" from volume
group "datavg"
```

Foreign LUN import

Oracle migration with FLI - planning

The procedures to migrate SAN resources using FLI are documented in NetApp TR-4380: SAN Migration Using Foreign LUN Import.

From a database and host point of view, no special steps are required. After the FC zones are updated and the LUNs become available on ONTAP, the LVM should be able to read the LVM metadata from the LUNs. Also, the volume groups are ready for use with no further configuration steps. In rare cases, environments might include configuration files that were hard-coded with references to the prior storage array. For example, a Linux system that included `/etc/multipath.conf` rules that referenced a WWN of a given device must be updated to reflect the changes introduced by FLI.

> Reference the NetApp Compatibility Matrix for information on supported configurations. If your environment is not included, contact your NetApp representative for assistance.

This example shows the migration of both ASM and LVM LUNs hosted on a Linux server. FLI is supported on other operating systems, and, although the host-side commands might differ, the principles are the same, and the ONTAP procedures are identical.

Identify LVM LUNs

The first step in preparation is to identify the LUNs to be migrated. In the example shown here, two SAN-based file systems are mounted at `/orabin` and `/backups`.

```
[root@host1 ~]# df -k
Filesystem                        1K-blocks      Used Available Use%
Mounted on
/dev/mapper/rhel-root              52403200   8811464  43591736  17% /
devtmpfs                           65882776         0  65882776   0% /dev
...
fas8060-nfs-public:/install      199229440 119368128  79861312  60%
/install
/dev/mapper/sanvg-lvorabin         20961280  12348476   8612804  59%
/orabin
/dev/mapper/sanvg-lvbackups        73364480  62947536  10416944  86%
/backups
```

The name of the volume group can be extracted from the device name, which uses the format (volume group name)-(logical volume name). In this case, the volume group is called `sanvg`.

The `pvdisplay` command can be used as follows to identify the LUNs that support this volume group. In this case, there are 10 LUNs that make up the `sanvg` volume group.

```
[root@host1 ~]# pvdisplay -C -o pv_name,pv_size,pv_fmt,vg_name
  PV                                            PSize   VG
  /dev/mapper/3600a0980383030445424487556574266  10.00g sanvg
  /dev/mapper/3600a0980383030445424487556574267  10.00g sanvg
  /dev/mapper/3600a0980383030445424487556574268  10.00g sanvg
  /dev/mapper/3600a0980383030445424487556574269  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426a  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426b  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426c  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426d  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426e  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426f  10.00g sanvg
  /dev/sda2                                     278.38g rhel
```

Identify ASM LUNs

ASM LUNs must also be migrated. To obtain the number of LUNs and LUN paths from sqlplus as the sysasm user, run the following command:

```
SQL> select path||' '||os_mb from v$asm_disk;
PATH||''||OS_MB
------------------------------------------------------------------------
------
/dev/oracleasm/disks/ASM0 10240
/dev/oracleasm/disks/ASM9 10240
/dev/oracleasm/disks/ASM8 10240
/dev/oracleasm/disks/ASM7 10240
/dev/oracleasm/disks/ASM6 10240
/dev/oracleasm/disks/ASM5 10240
/dev/oracleasm/disks/ASM4 10240
/dev/oracleasm/disks/ASM1 10240
/dev/oracleasm/disks/ASM3 10240
/dev/oracleasm/disks/ASM2 10240
10 rows selected.
SQL>
```

FC network changes

The current environment contains 20 LUNs to be migrated. Update the current SAN so that ONTAP can access the current LUNs. Data is not migrated yet, but ONTAP must read configuration information from the current LUNs to create the new home for that data.

At a minimum, at least one HBA port on the AFF/FAS system must be configured as an initiator port. In addition, the FC zones must be updated so that ONTAP can access the LUNs on the foreign storage array. Some storage arrays have LUN masking configured, which limits which WWNs can access a given LUN. In such cases, LUN masking must also be updated to grant access to the ONTAP WWNs.

After this step is completed, ONTAP should be able to view the foreign storage array with the `storage array show` command. The key field it returns is the prefix that is used to identify the foreign LUN on the system. In the example below, the LUNs on the foreign array `FOREIGN_1` appear within ONTAP using the prefix of `FOR-1`.

Identify foreign array

```
Cluster01::> storage array show -fields name,prefix
name          prefix
------------- ------
FOREIGN_1     FOR-1
Cluster01::>
```

Identify foreign LUNs

The LUNs can be listed by passing the `array-name` to the `storage disk show` command. The data returned is referenced multiple times during the migration procedure.

```
Cluster01::> storage disk show -array-name FOREIGN_1 -fields disk,serial
disk     serial-number
-------- -------------
FOR-1.1  800DT$HuVWBX
FOR-1.2  800DT$HuVWBZ
FOR-1.3  800DT$HuVWBW
FOR-1.4  800DT$HuVWBY
FOR-1.5  800DT$HuVWB/
FOR-1.6  800DT$HuVWBa
FOR-1.7  800DT$HuVWBd
FOR-1.8  800DT$HuVWBb
FOR-1.9  800DT$HuVWBc
FOR-1.10 800DT$HuVWBe
FOR-1.11 800DT$HuVWBf
FOR-1.12 800DT$HuVWBg
FOR-1.13 800DT$HuVWBi
FOR-1.14 800DT$HuVWBh
FOR-1.15 800DT$HuVWBj
FOR-1.16 800DT$HuVWBk
FOR-1.17 800DT$HuVWBm
FOR-1.18 800DT$HuVWBl
FOR-1.19 800DT$HuVWBo
FOR-1.20 800DT$HuVWBn
20 entries were displayed.
Cluster01::>
```

Register foreign array LUNs as import candidates

The foreign LUNs are initially classified as any particular LUN type. Before data can be imported, the LUNs must be tagged as foreign and therefore a candidate for the import process. This step is completed by passing the serial number to the `storage disk modify` command, as shown in the following example. Note that this process tags only the LUN as foreign within ONTAP. No data is written to the foreign LUN itself.

```
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBW} -is
-foreign true
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBX} -is
-foreign true
...
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBn} -is
-foreign true
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBo} -is
-foreign true
Cluster01::*>
```

Create volumes to host migrated LUNs

A volume is needed to host the migrated LUNs. The exact volume configuration depends on the overall plan to leverage ONTAP features. In this example, the ASM LUNs are placed into one volume and the LVM LUNs are placed in a second volume. Doing so allows you to manage the LUNs as independent groups for purposes such as tiering, creation of snapshots, or setting QoS controls.

Set the `snapshot-policy `to `none`. The migration process can include a great deal of data turnover. Therefore, there might be a large increase in space consumption if snapshots are created by accident because unwanted data is captured in the snapshots.

```
Cluster01::> volume create -volume new_asm -aggregate data_02 -size 120G
-snapshot-policy none
[Job 1152] Job succeeded: Successful
Cluster01::> volume create -volume new_lvm -aggregate data_02 -size 120G
-snapshot-policy none
[Job 1153] Job succeeded: Successful
Cluster01::>
```

Create ONTAP LUNs

After the volumes are created, the new LUNs must be created. Normally, the creation of a LUN requires the user to specify such information as the LUN size, but in this case the foreign-disk argument is passed to the command. As a result, ONTAP replicates the current LUN configuration data from the specified serial number. It also uses the LUN geometry and partition table data to adjust LUN alignment and establish optimum performance.

In this step, serial numbers must be cross-referenced against the foreign array to make sure that the correct foreign LUN is matched to the correct new LUN.

```
Cluster01::*> lun create -vserver vserver1 -path /vol/new_asm/LUN0 -ostype
linux -foreign-disk 800DT$HuVWBW
Created a LUN of size 10g (10737418240)
Cluster01::*> lun create -vserver vserver1 -path /vol/new_asm/LUN1 -ostype
linux -foreign-disk 800DT$HuVWBX
Created a LUN of size 10g (10737418240)
...
Created a LUN of size 10g (10737418240)
Cluster01::*> lun create -vserver vserver1 -path /vol/new_lvm/LUN8 -ostype
linux -foreign-disk 800DT$HuVWBn
Created a LUN of size 10g (10737418240)
Cluster01::*> lun create -vserver vserver1 -path /vol/new_lvm/LUN9 -ostype
linux -foreign-disk 800DT$HuVWBo
Created a LUN of size 10g (10737418240)
```

Create import relationships

The LUNs have now been created but are not configured as a replication destination. Before this step can be taken, the LUNs must first be placed offline. This extra step is designed to protect data from user errors. If ONTAP allowed a migration to be performed on an online LUN, it would create a risk that a typographical error could result in overwriting active data. The extra step of forcing the user to first take a LUN offline helps verify that the correct target LUN is used as a migration destination.

```
Cluster01::*> lun offline -vserver vserver1 -path /vol/new_asm/LUN0
Warning: This command will take LUN "/vol/new_asm/LUN0" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
Cluster01::*> lun offline -vserver vserver1 -path /vol/new_asm/LUN1
Warning: This command will take LUN "/vol/new_asm/LUN1" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
...
Warning: This command will take LUN "/vol/new_lvm/LUN8" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
Cluster01::*> lun offline -vserver vserver1 -path /vol/new_lvm/LUN9
Warning: This command will take LUN "/vol/new_lvm/LUN9" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
```

After the LUNs are offline, you can establish the import relationship by passing the foreign LUN serial number to the `lun import create` command.

```
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_asm/LUN0
-foreign-disk 800DT$HuVWBW
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_asm/LUN1
-foreign-disk 800DT$HuVWBX
...
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_lvm/LUN8
-foreign-disk 800DT$HuVWBn
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_lvm/LUN9
-foreign-disk 800DT$HuVWBo
Cluster01::*>
```

After all import relationships are established, the LUNs can be placed back online.

```
Cluster01::*> lun online -vserver vserver1 -path /vol/new_asm/LUN0
Cluster01::*> lun online -vserver vserver1 -path /vol/new_asm/LUN1
...
Cluster01::*> lun online -vserver vserver1 -path /vol/new_lvm/LUN8
Cluster01::*> lun online -vserver vserver1 -path /vol/new_lvm/LUN9
Cluster01::*>
```

Create initiator group

An initiator group (igroup) is part of the ONTAP LUN masking architecture. A newly created LUN is not accessible unless a host is first granted access. This is done by creating an igroup that lists either the FC WWNs or iSCSI initiator names that should be granted access. At the time this report was written, FLI was supported only for FC LUNs. However, converting to iSCSI postmigration is a simple task, as shown in Protocol Conversion.

In this example, an igroup is created that contains two WWNs that correspond to the two ports available on the host's HBA.

```
Cluster01::*> igroup create linuxhost -protocol fcp -ostype linux
-initiator 21:00:00:0e:1e:16:63:50 21:00:00:0e:1e:16:63:51
```

Map new LUNs to host

Following igroup creation, the LUNs are then mapped to the defined igroup. These LUNs are available only to the WWNs included in this igroup. NetApp assumes at this stage in the migration process that the host has not been zoned to ONTAP. This is important because if the host is simultaneously zoned to the foreign array and the new ONTAP system, then there is a risk that LUNs bearing the same serial number could be discovered on each array. This situation could lead to multipath malfunctions or damage to data.

```
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN0 -igroup
linuxhost
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN1 -igroup
linuxhost
...
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN8 -igroup
linuxhost
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN9 -igroup
linuxhost
Cluster01::*>
```

Oracle migration with FLI - cutover

Some disruption during a foreign LUN import is unavoidable because of the need to change the FC network configuration. However, the disruption does not have to last much longer than the time required to restart the database environment and update FC zoning

to switch the host FC connectivity from the foreign LUN to ONTAP.

This process can be summarized as follows:

1. Quiesce all LUN activity on the foreign LUNs.
2. Redirect host FC connections to the new ONTAP system.
3. Trigger the import process.
4. Rediscover the LUNs.
5. Restart the database.

You do not need to wait for the migration process to complete. As soon as the migration for a given LUN begins, it is available on ONTAP and can serve data while the data copy process continues. All reads are passed through to the foreign LUN, and all writes are synchronously written to both arrays. The copy operation is very fast and the overhead of redirecting FC traffic is minimal, so any impact on performance should be transient and minimal. If there is concern, you can delay restarting the environment until after the migration process is complete and the import relationships have been deleted.

Shut down database

The first step in quiescing the environment in this example is to shut down the database.

```
[oracle@host1 bin]$ . oraenv
ORACLE_SID = [oracle] ? FLIDB
The Oracle base remains unchanged with value /orabin
[oracle@host1 bin]$ sqlplus / as sysdba
SQL*Plus: Release 12.1.0.2.0
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, Automatic Storage Management, OLAP, Advanced
Analytics
and Real Application Testing options
SQL> shutdown immediate;
Database closed.
Database dismounted.
ORACLE instance shut down.
SQL>
```

Shut down grid services

One of the SAN-based file systems being migrated also includes the Oracle ASM services. Quiescing the underlying LUNs requires dismounting the file systems, which in turn means stopping any processes with open files on this file system.

```
[oracle@host1 bin]$ ./crsctl stop has -f
CRS-2791: Starting shutdown of Oracle High Availability Services-managed
resources on 'host1'
CRS-2673: Attempting to stop 'ora.evmd' on 'host1'
CRS-2673: Attempting to stop 'ora.DATA.dg' on 'host1'
CRS-2673: Attempting to stop 'ora.LISTENER.lsnr' on 'host1'
CRS-2677: Stop of 'ora.DATA.dg' on 'host1' succeeded
CRS-2673: Attempting to stop 'ora.asm' on 'host1'
CRS-2677: Stop of 'ora.LISTENER.lsnr' on 'host1' succeeded
CRS-2677: Stop of 'ora.evmd' on 'host1' succeeded
CRS-2677: Stop of 'ora.asm' on 'host1' succeeded
CRS-2673: Attempting to stop 'ora.cssd' on 'host1'
CRS-2677: Stop of 'ora.cssd' on 'host1' succeeded
CRS-2793: Shutdown of Oracle High Availability Services-managed resources
on 'host1' has completed
CRS-4133: Oracle High Availability Services has been stopped.
[oracle@host1 bin]$
```

Dismount file systems

If all the processes are shut down, the umount operation succeeds. If permission is denied, there must be a process with a lock on the file system. The `fuser` command can help identify these processes.

```
[root@host1 ~]# umount /orabin
[root@host1 ~]# umount /backups
```

Deactivate volume groups

After all file systems in a given volume group are dismounted, the volume group can be deactivated.

```
[root@host1 ~]# vgchange --activate n sanvg
  0 logical volume(s) in volume group "sanvg" now active
[root@host1 ~]#
```

FC network changes

The FC zones can now be updated to remove all access from the host to the foreign array and establish access to ONTAP.

Start import process

To start the LUN import processes, run the `lun import start` command.

```
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_asm/LUN0
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_asm/LUN1
...
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_lvm/LUN8
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_lvm/LUN9
Cluster01::lun import*>
```

Monitor import progress

The import operation can be monitored with the `lun import show` command. As shown below, the import of all 20 LUNs is underway, which means that data is now accessible through ONTAP even though the data copy operation still progresses.

```
Cluster01::lun import*> lun import show -fields path,percent-complete
vserver   foreign-disk path              percent-complete
--------- ------------ ----------------- ----------------
vserver1  800DT$HuVWB/ /vol/new_asm/LUN4 5
vserver1  800DT$HuVWBW /vol/new_asm/LUN0 5
vserver1  800DT$HuVWBX /vol/new_asm/LUN1 6
vserver1  800DT$HuVWBY /vol/new_asm/LUN2 6
vserver1  800DT$HuVWBZ /vol/new_asm/LUN3 5
vserver1  800DT$HuVWBa /vol/new_asm/LUN5 4
vserver1  800DT$HuVWBb /vol/new_asm/LUN6 4
vserver1  800DT$HuVWBc /vol/new_asm/LUN7 4
vserver1  800DT$HuVWBd /vol/new_asm/LUN8 4
vserver1  800DT$HuVWBe /vol/new_asm/LUN9 4
vserver1  800DT$HuVWBf /vol/new_lvm/LUN0 5
vserver1  800DT$HuVWBg /vol/new_lvm/LUN1 4
vserver1  800DT$HuVWBh /vol/new_lvm/LUN2 4
vserver1  800DT$HuVWBi /vol/new_lvm/LUN3 3
vserver1  800DT$HuVWBj /vol/new_lvm/LUN4 3
vserver1  800DT$HuVWBk /vol/new_lvm/LUN5 3
vserver1  800DT$HuVWBl /vol/new_lvm/LUN6 4
vserver1  800DT$HuVWBm /vol/new_lvm/LUN7 3
vserver1  800DT$HuVWBn /vol/new_lvm/LUN8 2
vserver1  800DT$HuVWBo /vol/new_lvm/LUN9 2
20 entries were displayed.
```

If you require an offline process, delay rediscovering or restarting services until the `lun import show` command indicates that all migration is successful and complete. You can then complete the migration process as described in Foreign LUN Import—Completion.

If you require an online migration, proceed to rediscover the LUNs in their new home and bring up the services.

Scan for SCSI device changes

In most cases, the simplest option to rediscover new LUNs is to restart the host. Doing so automatically removes old stale devices, properly discovers all new LUNs, and builds associated devices such as multipathing devices. The example here shows a wholly online process for demonstration purposes.

Caution: Before restarting a host, make sure that all entries in `/etc/fstab` that reference migrated SAN resources are commented out. If this is not done and there are problems with LUN access, the OS might not boot. This situation does not damage data. However, it can be very inconvenient to boot into rescue mode or a similar mode and correct the `/etc/fstab` so that the OS can be booted to enable troubleshooting.

The LUNs on the version of Linux used in this example can be rescanned with the `rescan-scsi-bus.sh` command. If the command is successful, each LUN path should appear in the output. The output can be difficult to interpret, but, if the zoning and igroup configuration was correct, many LUNs should appear that include a `NETAPP` vendor string.

```
[root@host1 /]# rescan-scsi-bus.sh
Scanning SCSI subsystem for new devices
Scanning host 0 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
 Scanning for device 0 2 0 0 ...
OLD: Host: scsi0 Channel: 02 Id: 00 Lun: 00
      Vendor: LSI      Model: RAID SAS 6G 0/1  Rev: 2.13
      Type:   Direct-Access                    ANSI SCSI revision: 05
Scanning host 1 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
 Scanning for device 1 0 0 0 ...
OLD: Host: scsi1 Channel: 00 Id: 00 Lun: 00
      Vendor: Optiarc  Model: DVD RW AD-7760H  Rev: 1.41
      Type:   CD-ROM                           ANSI SCSI revision: 05
Scanning host 2 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 3 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 4 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 5 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 6 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 7 for  all SCSI target IDs, all LUNs
 Scanning for device 7 0 0 10 ...
OLD: Host: scsi7 Channel: 00 Id: 00 Lun: 10
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
 Scanning for device 7 0 0 11 ...
OLD: Host: scsi7 Channel: 00 Id: 00 Lun: 11
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
 Scanning for device 7 0 0 12 ...
...
OLD: Host: scsi9 Channel: 00 Id: 01 Lun: 18
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
 Scanning for device 9 0 1 19 ...
OLD: Host: scsi9 Channel: 00 Id: 01 Lun: 19
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
0 new or changed device(s) found.
0 remapped or resized device(s) found.
0 device(s) removed.
```

Check for multipath devices

The LUN discovery process also triggers the recreation of multipath devices, but the Linux multipathing driver is known to have occasional problems. The output of `multipath - ll` should be checked to verify that the output looks as expected. For example, the output below shows multipath devices associated with a `NETAPP` vendor string. Each device has four paths, with two at a priority of 50 and two at a priority of 10. Although the exact output can vary with different versions of Linux, this output looks as expected.

Reference the host utilities documentation for the version of Linux you use to verify that the `/etc/multipath.conf` settings are correct.

```
[root@host1 /]# multipath -ll
3600a098038303558735d493762504b36 dm-5 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:4  sdat 66:208 active ready running
| `- 9:0:1:4  sdbn 68:16  active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:4  sdf  8:80   active ready running
  `- 9:0:0:4  sdz  65:144 active ready running
3600a098038303558735d493762504b2d dm-10 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:8  sdax 67:16  active ready running
| `- 9:0:1:8  sdbr 68:80  active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:8  sdj  8:144  active ready running
  `- 9:0:0:8  sdad 65:208 active ready running
...
3600a098038303558735d493762504b37 dm-8 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:5  sdau 66:224 active ready running
| `- 9:0:1:5  sdbo 68:32  active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:5  sdg  8:96   active ready running
  `- 9:0:0:5  sdaa 65:160 active ready running
3600a098038303558735d493762504b4b dm-22 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:19 sdbi 67:192 active ready running
| `- 9:0:1:19 sdcc 69:0   active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:19 sdu  65:64  active ready running
  `- 9:0:0:19 sdao 66:128 active ready running
```

Reactivate LVM volume group

If the LVM LUNs have been properly discovered, the `vgchange --activate y` command should succeed.

This is a good example of the value of a logical volume manager. A change in the WWN of a LUN or even a serial number is unimportant because the volume group metadata is written on the LUN itself.

The OS scanned the LUNs and discovered a small amount of data written on the LUN that identifies it as a physical volume belonging to the `sanvg volumegroup`. It then built all of the required devices. All that is required is to reactivate the volume group.

```
[root@host1 /]# vgchange --activate y sanvg
  Found duplicate PV fpCzdLTuKfy2xDZjai1NliJh3TjLUBiT: using
/dev/mapper/3600a098038303558735d493762504b46 not /dev/sdp
  Using duplicate PV /dev/mapper/3600a098038303558735d493762504b46 from
subsystem DM, ignoring /dev/sdp
  2 logical volume(s) in volume group "sanvg" now active
```

Remount file systems

After the volume group is reactivated, the file systems can be mounted with all of the original data intact. As discussed previously, the file systems are fully operational even if data replication is still active in the back group.

```
[root@host1 /]# mount /orabin
[root@host1 /]# mount /backups
[root@host1 /]# df -k
Filesystem                      1K-blocks      Used Available Use%
Mounted on
/dev/mapper/rhel-root             52403200   8837100  43566100  17% /
devtmpfs                          65882776         0  65882776   0% /dev
tmpfs                              6291456        84   6291372   1%
/dev/shm
tmpfs                             65898668      9884  65888784   1% /run
tmpfs                             65898668         0  65898668   0%
/sys/fs/cgroup
/dev/sda1                           505580    224828    280752  45% /boot
fas8060-nfs-public:/install      199229440 119368256  79861184  60%
/install
fas8040-nfs-routable:/snapomatic   9961472     30528   9930944   1%
/snapomatic
tmpfs                             13179736        16  13179720   1%
/run/user/42
tmpfs                             13179736         0  13179736   0%
/run/user/0
/dev/mapper/sanvg-lvorabin        20961280  12357456   8603824  59%
/orabin
/dev/mapper/sanvg-lvbackups       73364480  62947536  10416944  86%
/backups
```

Rescan for ASM devices

The ASMlib devices should have been rediscovered when the SCSI devices were rescanned. Rediscovery can be verified online by restarting ASMlib and then scanning the disks.

> This step is only relevant to ASM configurations where ASMlib is used.

Caution: Where ASMlib is not used, the `/dev/mapper` devices should have been automatically recreated. However, the permissions might not be correct. You must set special permissions on the underlying devices for ASM in the absence of ASMlib. Doing so is usually accomplished through special entries in either the `/etc/multipath.conf` or `udev` rules, or possibly in both rule sets. These files might need to be updated to reflect changes in the environment in terms of WWNs or serial numbers to make sure that the ASM devices still have the correct permissions.

In this example, restarting ASMlib and scanning for disks show the same 10 ASM LUNs as the original environment.

```
[root@host1 /]# oracleasm exit
Unmounting ASMlib driver filesystem: /dev/oracleasm
Unloading module "oracleasm": oracleasm
[root@host1 /]# oracleasm init
Loading module "oracleasm": oracleasm
Configuring "oracleasm" to use device physical block size
Mounting ASMlib driver filesystem: /dev/oracleasm
[root@host1 /]# oracleasm scandisks
Reloading disk partitions: done
Cleaning any stale ASM disks...
Scanning system for ASM disks...
Instantiating disk "ASM0"
Instantiating disk "ASM1"
Instantiating disk "ASM2"
Instantiating disk "ASM3"
Instantiating disk "ASM4"
Instantiating disk "ASM5"
Instantiating disk "ASM6"
Instantiating disk "ASM7"
Instantiating disk "ASM8"
Instantiating disk "ASM9"
```

Restart grid services

Now that the LVM and ASM devices are online and available, the grid services can be restarted.

```
[root@host1 /]# cd /orabin/product/12.1.0/grid/bin
[root@host1 bin]# ./crsctl start has
```

Restart database

After the grid services have been restarted, the database can be brought up. It might be necessary to wait a few minutes for the ASM services to become fully available before trying to start the database.

```
[root@host1 bin]# su - oracle
[oracle@host1 ~]$ . oraenv
ORACLE_SID = [oracle] ? FLIDB
The Oracle base has been set to /orabin
[oracle@host1 ~]$ sqlplus / as sysdba
SQL*Plus: Release 12.1.0.2.0
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> startup
ORACLE instance started.
Total System Global Area 3221225472 bytes
Fixed Size                  4502416 bytes
Variable Size            1207962736 bytes
Database Buffers         1996488704 bytes
Redo Buffers               12271616 bytes
Database mounted.
Database opened.
SQL>
```

Oracle migration with FLI - completion

From a host point of view, the migration is complete, but I/O is still served from the foreign array until the import relationships are deleted.

Before deleting the relationships, you must confirm that the migration process is complete for all LUNs.

```
Cluster01::*> lun import show -vserver vserver1 -fields foreign-
disk,path,operational-state
vserver   foreign-disk path              operational-state
--------- ------------ ----------------- -----------------
vserver1 800DT$HuVWB/ /vol/new_asm/LUN4 completed
vserver1 800DT$HuVWBW /vol/new_asm/LUN0 completed
vserver1 800DT$HuVWBX /vol/new_asm/LUN1 completed
vserver1 800DT$HuVWBY /vol/new_asm/LUN2 completed
vserver1 800DT$HuVWBZ /vol/new_asm/LUN3 completed
vserver1 800DT$HuVWBa /vol/new_asm/LUN5 completed
vserver1 800DT$HuVWBb /vol/new_asm/LUN6 completed
vserver1 800DT$HuVWBc /vol/new_asm/LUN7 completed
vserver1 800DT$HuVWBd /vol/new_asm/LUN8 completed
vserver1 800DT$HuVWBe /vol/new_asm/LUN9 completed
vserver1 800DT$HuVWBf /vol/new_lvm/LUN0 completed
vserver1 800DT$HuVWBg /vol/new_lvm/LUN1 completed
vserver1 800DT$HuVWBh /vol/new_lvm/LUN2 completed
vserver1 800DT$HuVWBi /vol/new_lvm/LUN3 completed
vserver1 800DT$HuVWBj /vol/new_lvm/LUN4 completed
vserver1 800DT$HuVWBk /vol/new_lvm/LUN5 completed
vserver1 800DT$HuVWBl /vol/new_lvm/LUN6 completed
vserver1 800DT$HuVWBm /vol/new_lvm/LUN7 completed
vserver1 800DT$HuVWBn /vol/new_lvm/LUN8 completed
vserver1 800DT$HuVWBo /vol/new_lvm/LUN9 completed
20 entries were displayed.
```

Delete import relationships

When the migration process is complete, delete the migration relationship. After you have done so, I/O is served exclusively from the drives on ONTAP.

```
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_asm/LUN0
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_asm/LUN1
...
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_lvm/LUN8
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_lvm/LUN9
```

Deregister foreign LUNs

Finally, modify the disk to remove the `is-foreign` designation.

```
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBW} -is
-foreign false
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBX} -is
-foreign false
...
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBn} -is
-foreign false
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBo} -is
-foreign false
Cluster01::*>
```

Oracle migration with FLI - protocol conversion

Changing the protocol used to access a LUN is a common requirement.

In some cases, it is part of an overall strategy to migrate data to the cloud. TCP/IP is the protocol of the cloud, and changing from FC to iSCSI allows easier migration into various cloud environments. In other cases, iSCSI might be desirable to leverage the decreased costs of an IP SAN. On occasion, a migration might use a different protocol as a temporary measure. For example, if a foreign array and ONTAP based LUNs cannot coexist on the same HBAs, you can use iSCSI LUNs long enough to copy data from the old array. You can then convert back to FC after the old LUNs are removed from the system.

The following procedure demonstrates conversion from FC to iSCSI, but the overall principles apply to a reverse iSCSI to FC conversion.

Install iSCSI initiator

Most operating systems include a software iSCSI initiator by default, but if one is not included, it can be easily installed.

```
[root@host1 /]# yum install -y iscsi-initiator-utils
Loaded plugins: langpacks, product-id, search-disabled-repos,
subscription-
              : manager
Resolving Dependencies
--> Running transaction check
---> Package iscsi-initiator-utils.x86_64 0:6.2.0.873-32.el7 will be
updated
--> Processing Dependency: iscsi-initiator-utils = 6.2.0.873-32.el7 for
package: iscsi-initiator-utils-iscsiuio-6.2.0.873-32.el7.x86_64
---> Package iscsi-initiator-utils.x86_64 0:6.2.0.873-32.0.2.el7 will be
an update
--> Running transaction check
---> Package iscsi-initiator-utils-iscsiuio.x86_64 0:6.2.0.873-32.el7 will
be updated
---> Package iscsi-initiator-utils-iscsiuio.x86_64 0:6.2.0.873-32.0.2.el7
will be an update
```

```
--> Finished Dependency Resolution
Dependencies Resolved
================================================================================
===
 Package                        Arch   Version              Repository
Size
================================================================================
===
Updating:
 iscsi-initiator-utils          x86_64 6.2.0.873-32.0.2.el7 ol7_latest 416
k
Updating for dependencies:
 iscsi-initiator-utils-iscsiuio x86_64 6.2.0.873-32.0.2.el7 ol7_latest  84
k
Transaction Summary
================================================================================
===
Upgrade  1 Package (+1 Dependent package)
Total download size: 501 k
Downloading packages:
No Presto metadata available for ol7_latest
(1/2): iscsi-initiator-utils-6.2.0.873-32.0.2.el7.x86_6 | 416 kB   00:00
(2/2): iscsi-initiator-utils-iscsiuio-6.2.0.873-32.0.2. |  84 kB   00:00
--------------------------------------------------------------------------------
---
Total                                           2.8 MB/s | 501 kB
00:00Cluster01
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  Updating   : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.0.2.el7.x86
1/4
  Updating   : iscsi-initiator-utils-6.2.0.873-32.0.2.el7.x86_64
2/4
  Cleanup    : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.el7.x86_64
3/4
  Cleanup    : iscsi-initiator-utils-6.2.0.873-32.el7.x86_64
4/4
rhel-7-server-eus-rpms/7Server/x86_64/productid       | 1.7 kB     00:00
rhel-7-server-rpms/7Server/x86_64/productid           | 1.7 kB     00:00
  Verifying  : iscsi-initiator-utils-6.2.0.873-32.0.2.el7.x86_64
1/4
  Verifying  : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.0.2.el7.x86
2/4
  Verifying  : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.el7.x86_64
```

```
3/4
  Verifying  : iscsi-initiator-utils-6.2.0.873-32.el7.x86_64
4/4
Updated:
  iscsi-initiator-utils.x86_64 0:6.2.0.873-32.0.2.el7
Dependency Updated:
  iscsi-initiator-utils-iscsiuio.x86_64 0:6.2.0.873-32.0.2.el7
Complete!
[root@host1 /]#
```

Identify iSCSI initiator name

A unique iSCSI initiator name is generated during the installation process. On Linux, it is located in the `/etc/iscsi/initiatorname.iscsi` file. This name is used to identify the host on the IP SAN.

```
[root@host1 /]# cat /etc/iscsi/initiatorname.iscsi
InitiatorName=iqn.1992-05.com.redhat:497bd66ca0
```

Create new initiator group

An initiator group (igroup) is part of the ONTAP LUN masking architecture. A newly created LUN is not accessible unless a host is first granted access. This step is accomplished by creating an igroup that lists either the FC WWNs or iSCSI initiator names that require access.

In this example, an igroup is created that contains the iSCSI initiator of the Linux host.

```
Cluster01::*> igroup create -igroup linuxiscsi -protocol iscsi -ostype
linux -initiator iqn.1994-05.com.redhat:497bd66ca0
```

Shut down environment

Before changing the LUN protocol, the LUNs must be fully quiesced. Any database on one of the LUNs being converted must be shut down, file systems must be dismounted, and volume groups must be deactivated. Where ASM is used, make sure that the ASM disk group is dismounted and shut down all grid services.

Unmap LUNs from FC network

After the LUNs are fully quiesced, remove the mappings from the original FC igroup.

```
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_asm/LUN0 -igroup
linuxhost
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_asm/LUN1 -igroup
linuxhost
...
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_lvm/LUN8 -igroup
linuxhost
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_lvm/LUN9 -igroup
linuxhost
```

Remap LUNs to IP network

Grant access to each LUN to the new iSCSI-based initiator group.

```
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN0 -igroup
linuxiscsi
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN1 -igroup
linuxiscsi
...
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN8 -igroup
linuxiscsi
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN9 -igroup
linuxiscsi
Cluster01::*>
```

Discover iSCSI targets

There are two phases to iSCSI discovery. The first is to discover the targets, which is not the same as discovering a LUN. The `iscsiadm` command shown below probes the portal group specified by the `-p argument` and stores a list of all IP addresses and ports that offer iSCSI services. In this case, there are four IP addresses that have iSCSI services on the default port 3260.

> This command can take several minutes to complete if any of the target IP addresses cannot be reached.

```
[root@host1 ~]# iscsiadm -m discovery -t st -p fas8060-iscsi-public1
10.63.147.197:3260,1033 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
10.63.147.198:3260,1034 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
172.20.108.203:3260,1030 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
172.20.108.202:3260,1029 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
```

Discover iSCSI LUNs

After the iSCSI targets are discovered, restart the iSCSI service to discover the available iSCSI LUNs and build associated devices such as multipath or ASMlib devices.

```
[root@host1 ~]# service iscsi restart
Redirecting to /bin/systemctl restart  iscsi.service
```

Restart environment

Restart the environment by reactivating volume groups, remounting file systems, restarting RAC services, and so on. As a precaution, NetApp recommends that you reboot the server after the conversion process is complete to be certain that all configuration files are correct and all stale devices are removed.

Caution: Before restarting a host, make sure that all entries in `/etc/fstab` that reference migrated SAN resources are commented out. If this step is not taken and there are problems with LUN access, the result can be an OS that does not boot. This issue does not damage data. However, it can be very inconvenient to boot into rescue mode or a similar mode and correct `/etc/fstab` so that the OS can be booted to allow troubleshooting efforts to begin.

Oracle migration procedure sample scripts

The scripts presented are provided as examples of how to script various OS and database tasks. They are supplied as is. If support is required for a particular procedure, contact NetApp or a NetApp reseller.

Database shutdown

The following Perl script takes a single argument of the Oracle SID and shuts down a database. It can be run as the Oracle user or as root.

```
#! /usr/bin/perl
use strict;
use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
my $uid=$<;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
77 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
shutdown immediate;
EOF2
'
`;}
else {
@out=`. oraenv << EOF1
$oraclesid
EOF4
sqlplus / as sysdba << EOF2
shutdown immediate;
EOF2
`;};
print @out;
if ("@out" =~ /ORACLE instance shut down/) {
print "$oraclesid shut down\n";
exit 0;}
elsif ("@out" =~ /Connected to an idle instance/) {
print "$oraclesid already shut down\n";
exit 0;}
else {
print "$oraclesid failed to shut down\n";
exit 1;}
```

Database startup

The following Perl script takes a single argument of the Oracle SID and shuts down a database. It can be run as the Oracle user or as root.

```
#! /usr/bin/perl
use strict;
use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
my $uid=$<;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
startup;
EOF2
'
`;}
else {
@out=`. oraenv << EOF3
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
startup;
EOF2
`;};
print @out;
if ("@out" =~ /Database opened/) {
print "$oraclesid started\n";
exit 0;}
elsif ("@out" =~ /cannot start already-running ORACLE/) {
print "$oraclesid already started\n";
exit 1;}
else {
78 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
print "$oraclesid failed to start\n";
exit 1;}
```

Convert file system to read-only

The following script takes a file- system argument and attempts to dismount and remount it as read-only. Doing so is useful during migration processes in which a file system must be kept available to replicate data and yet must be protected against accidental damage.

```
#! /usr/bin/perl
use strict;
#use warnings;
my $filesystem=$ARGV[0];
my @out=`umount '$filesystem'`;
if ($? == 0) {
 print "$filesystem unmounted\n";
 @out = `mount -o ro '$filesystem'`;
 if ($? == 0) {
 print "$filesystem mounted read-only\n";
 exit 0;}}
else {
 print "Unable to unmount $filesystem\n";
 exit 1;}
print @out;
```

Replace file system

The following script example is used to replace one file system with another. Because it edits the `/etc/fstab `file, it must run as root. It accepts a single comma-delimited argument of the old and new file systems.

1. To replace the file system, run the following script:

```
#! /usr/bin/perl
use strict;
#use warnings;
my $oldfs;
my $newfs;
my @oldfstab;
my @newfstab;
my $source;
my $mountpoint;
my $leftover;
my $oldfstabentry='';
my $newfstabentry='';
my $migratedfstabentry='';
($oldfs, $newfs) = split (',',$ARGV[0]);
open(my $filehandle, '<', '/etc/fstab') or die "Could not open
/etc/fstab\n";
while (my $line = <$filehandle>) {
 chomp $line;
 ($source, $mountpoint, $leftover) = split(/[ , ]/,$line, 3);
 if ($mountpoint eq $oldfs) {
 $oldfstabentry = "#Removed by swap script $source $oldfs $leftover";}
 elsif ($mountpoint eq $newfs) {
```

```
 $newfstabentry = "#Removed by swap script $source $newfs $leftover";
 $migratedfstabentry = "$source $oldfs $leftover";}
 else {
 push (@newfstab, "$line\n")}}
79 Migration of Oracle Databases to NetApp Storage Systems © 2021
NetApp, Inc. All rights reserved
push (@newfstab, "$oldfstabentry\n");
push (@newfstab, "$newfstabentry\n");
push (@newfstab, "$migratedfstabentry\n");
close($filehandle);
if ($oldfstabentry eq ''){
 die "Could not find $oldfs in /etc/fstab\n";}
if ($newfstabentry eq ''){
 die "Could not find $newfs in /etc/fstab\n";}
my @out=`umount '$newfs'`;
if ($? == 0) {
 print "$newfs unmounted\n";}
else {
 print "Unable to unmount $newfs\n";
 exit 1;}
@out=`umount '$oldfs'`;
if ($? == 0) {
 print "$oldfs unmounted\n";}
else {
 print "Unable to unmount $oldfs\n";
 exit 1;}
system("cp /etc/fstab /etc/fstab.bak");
open ($filehandle, ">", '/etc/fstab') or die "Could not open /etc/fstab
for writing\n";
for my $line (@newfstab) {
 print $filehandle $line;}
close($filehandle);
@out=`mount '$oldfs'`;
if ($? == 0) {
 print "Mounted updated $oldfs\n";
 exit 0;}
else{
 print "Unable to mount updated $oldfs\n";
 exit 1;}
exit 0;
```

As an example of this script's use, assume that data in `/oradata` is migrated to `/neworadata` and `/logs` is migrated to `/newlogs`. One of the simplest methods to perform this task is by using a simple file copy operation to relocate the new device back to the original mountpoint.

2. Assume that the old and new file systems are present in the `/etc/fstab` file as follows:

```
cluster01:/vol_oradata /oradata nfs rw,bg,vers=3,rsize=65536,wsize=65536
0 0
cluster01:/vol_logs /logs nfs rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_neworadata /neworadata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_newlogs /newlogs nfs rw,bg,vers=3,rsize=65536,wsize=65536
0 0
```

3. When run, this script unmounts the current file system and replaces it with the new:

```
[root@jfsc3 scripts]# ./swap.fs.pl /oradata,/neworadata
/neworadata unmounted
/oradata unmounted
Mounted updated /oradata
[root@jfsc3 scripts]# ./swap.fs.pl /logs,/newlogs
/newlogs unmounted
/logs unmounted
Mounted updated /logs
```

4. The script also updates the `/etc/fstab` file accordingly. In the example shown here, it includes the following changes:

```
#Removed by swap script cluster01:/vol_oradata /oradata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
#Removed by swap script cluster01:/vol_neworadata /neworadata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_neworadata /oradata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
#Removed by swap script cluster01:/vol_logs /logs nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
#Removed by swap script cluster01:/vol_newlogs /newlogs nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_newlogs /logs nfs rw,bg,vers=3,rsize=65536,wsize=65536 0
0
```

Automated database migration

This example demonstrates the use of shutdown, startup, and file system replacement scripts to fully automate a migration.

```
#! /usr/bin/perl
use strict;
#use warnings;
```

```
my $oraclesid=$ARGV[0];
my @oldfs;
my @newfs;
my $x=1;
while ($x < scalar(@ARGV)) {
   ($oldfs[$x-1], $newfs[$x-1]) = split (',',$ARGV[$x]);
   $x+=1;}
my @out=`./dbshut.pl '$oraclesid'`;
print @out;
if ($? ne 0) {
   print "Failed to shut down database\n";
   exit 0;}
$x=0;
while ($x < scalar(@oldfs)) {
   my @out=`./mk.fs.readonly.pl '$oldfs[$x]'`;
   if ($? ne 0) {
      print "Failed to make filesystem $oldfs[$x] readonly\n";
      exit 0;}
   $x+=1;}
$x=0;
while ($x < scalar(@oldfs)) {
   my @out=`rsync -rlpogt --stats --progress --exclude='.snapshot'
'$oldfs[$x]/' '/$newfs[$x]/'`;
   print @out;
   if ($? ne 0) {
      print "Failed to copy filesystem $oldfs[$x] to $newfs[$x]\n";
      exit 0;}
   else {
      print "Succesfully replicated filesystem $oldfs[$x] to
$newfs[$x]\n";}
   $x+=1;}
$x=0;
while ($x < scalar(@oldfs)) {
   print "swap $x $oldfs[$x] $newfs[$x]\n";
   my @out=`./swap.fs.pl '$oldfs[$x],$newfs[$x]'`;
   print @out;
   if ($? ne 0) {
      print "Failed to swap filesystem $oldfs[$x] for $newfs[$x]\n";
      exit 1;}
   else {
      print "Swapped filesystem $oldfs[$x] for $newfs[$x]\n";}
   $x+=1;}
my @out=`./dbstart.pl '$oraclesid'`;
print @out;
```

Display file locations

This script collects a number of critical database parameters and prints them in an easy-to-read format. This script can be useful when reviewing data layouts. In addition, the script can be modified for other uses.

```
#! /usr/bin/perl
#use strict;
#use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
sub dosql{
        my $command = @_[0];
        my @lines;
        my $uid=$<;
        if ($uid == 0) {
        @lines=`su - $oracleuser -c "export ORAENV_ASK=NO;export
ORACLE_SID=$oraclesid;. oraenv -s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
"
        `;}
        else {
        $command=~s/\\\\\\/\\/g;
        @lines=`export ORAENV_ASK=NO;export ORACLE_SID=$oraclesid;. oraenv
-s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
        `;};
return @lines}
print "\n";
@out=dosql('select name from v\\\\\$datafile;');
print "$oraclesid datafiles:\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('select member from v\\\\\$logfile;');
print "$oraclesid redo logs:\n";
for $line (@out) {
```

```
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('select name from v\\\\\$tempfile;');
print "$oraclesid temp datafiles:\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('show parameter spfile;');
print "$oraclesid spfile\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('select name||\' \'||value from v\\\\\$parameter where
isdefault=\'FALSE\';');
print "$oraclesid key parameters\n";
for $line (@out) {
        chomp($line);
        if ($line =~ /control_files/) {print "$line\n";}
        if ($line =~ /db_create/) {print "$line\n";}
        if ($line =~ /db_file_name_convert/) {print "$line\n";}
        if ($line =~ /log_archive_dest/) {print "$line\n";}}
        if ($line =~ /log_file_name_convert/) {print "$line\n";}
        if ($line =~ /pdb_file_name_convert/) {print "$line\n";}
        if ($line =~ /spfile/) {print "$line\n";}
print "\n";
```

ASM migration cleanup

```
#! /usr/bin/perl
#use strict;
#use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
sub dosql{
        my $command = @_[0];
        my @lines;
        my $uid=$<;
        if ($uid == 0) {
        @lines=`su - $oracleuser -c "export ORAENV_ASK=NO;export
ORACLE_SID=$oraclesid;. oraenv -s << EOF1
EOF1
```

```
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
"
        `;}
        else {
        $command=~s/\\\\\\/\\/g;
        @lines=`export ORAENV_ASK=NO;export ORACLE_SID=$oraclesid;. oraenv
-s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
        `;}
return @lines}
print "\n";
@out=dosql('select name from v\\\\\$datafile;');
print @out;
print "shutdown immediate;\n";
print "startup mount;\n";
print "\n";
for $line (@out) {
        if (length($line) > 1) {
                chomp($line);
                ($first, $second,$third,$fourth)=split('_',$line);
                $fourth =~ s/^TS-//;
                $newname=lc("$fourth.dbf");
                $path2file=$line;
                $path2file=~ /(^.*.\/)/;
                print "host mv $line $1$newname\n";}}
print "\n";
for $line (@out) {
        if (length($line) > 1) {
                chomp($line);
                ($first, $second,$third,$fourth)=split('_',$line);
                $fourth =~ s/^TS-//;
                $newname=lc("$fourth.dbf");
                $path2file=$line;
                $path2file=~ /(^.*.\/)/;
                print "alter database rename file '$line' to
'$1$newname';\n";}}
print "alter database open;\n";
print "\n";
```

ASM to file system name conversion

```
set serveroutput on;
set wrap off;
declare
    cursor df is select file#, name from v$datafile;
    cursor tf is select file#, name from v$tempfile;
    cursor lf is select member from v$logfile;
    firstline boolean := true;
begin
    dbms_output.put_line(CHR(13));
    dbms_output.put_line('Parameters for log file conversion:');
    dbms_output.put_line(CHR(13));
    dbms_output.put('*.log_file_name_convert = ');
    for lfrec in lf loop
        if (firstline = true) then
            dbms_output.put('''' || lfrec.member || ''', ');
            dbms_output.put('''/NEW_PATH/' ||
regexp_replace(lfrec.member,'^.*./','') || '''');
        else
            dbms_output.put(',''' || lfrec.member || ''', ');
            dbms_output.put('''/NEW_PATH/' ||
regexp_replace(lfrec.member,'^.*./','') || '''');
        end if;
        firstline:=false;
    end loop;
    dbms_output.put_line(CHR(13));
    dbms_output.put_line(CHR(13));
    dbms_output.put_line('rman duplication script:');
    dbms_output.put_line(CHR(13));
    dbms_output.put_line('run');
    dbms_output.put_line('{');
    for dfrec in df loop
        dbms_output.put_line('set newname for datafile ' ||
            dfrec.file# || ' to ''' || dfrec.name ||''';');
    end loop;
    for tfrec in tf loop
        dbms_output.put_line('set newname for tempfile ' ||
            tfrec.file# || ' to ''' || tfrec.name ||''';');
    end loop;
    dbms_output.put_line('duplicate target database for standby backup
location INSERT_PATH_HERE;');
    dbms_output.put_line('}');
end;
/
```

Replay logs on database

This script accepts a single argument of an Oracle SID for a database that is in mount mode and attempts to replay all currently available archive logs.

```
#! /usr/bin/perl
use strict;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
84 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
my $uid = $<;
my @out;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover database until cancel;
auto
EOF2
'
`;}
else {
@out=`. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover database until cancel;
auto
EOF2
`;
}
print @out;
```

Replay logs on standby database

This script is identical to the preceding script, except that it is designed for a standby database.

```
#! /usr/bin/perl
use strict;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my $uid = $<;
my @out;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover standby database until cancel;
auto
EOF2
'
`;}
else {
@out=`. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover standby database until cancel;
auto
EOF2
`;
}
print @out;
```

Additional notes

Oracle database performance optimization and benchmarking procedures

Accurate testing of database storage performance is an extremely complicated subject. It requires an understanding of the following issues:

- IOPS and throughput
- The difference between foreground and background I/O operations
- The effect of latency upon the database
- Numerous OS and network settings that also affect storage performance

In addition, there are nonstorage databases tasks to consider. There is a point at which optimizing storage performance yields no useful benefits because storage performance is no longer a limiting factor for performance.

A majority of database customers now select all-flash arrays, which creates some additional considerations.

For example, consider performance testing on a two-node AFF A900 system:

- With a 80/20 read/write ratio, two A900 nodes can deliver over 1M random database IOPS before latency even crosses the 150µs mark. This is so far beyond the current performance demands of most databases that it is difficult to predict the expected improvement. Storage would be largely erased as a bottleneck.
- Network bandwidth is an increasingly common source of performance limitations. For example, spinning disk solutions are often bottlenecks for database performance because the I/O latency is very high. When latency limitations are removed by an all-flash array, the barrier frequently shifts to the network. This is especially notable with virtualized environments and blade systems where the true network connectivity is difficult to visualize. This can complicate performance testing if the storage system itself cannot be fully utilized due to bandwidth limitations.
- Comparing the performance of an all-flash array with an array containing spinning disks is generally not possible because of the dramatically improved latency of all-flash arrays. Test results are typically not meaningful.
- Comparing peak IOPS performance with an all-flash array is frequently not a useful test because databases are not limited by storage I/O. For example, assume one array can sustain 500K random IOPS, whereas another can sustain 300K. The difference is irrelevant in the real world if a database is spending 99% of its time on CPU processing. The workloads never utilize the full capabilities of the storage array. In contrast, peak IOPS capabilities might be critical in a consolidation platform in which the storage array is expected to be loaded to its peak capabilities.
- Always consider latency as well as IOPS in any storage test. Many storage arrays in the market make claims of extreme levels of IOPS, but the latency renders those IOPS useless at such levels. The typical target with all-flash arrays is the 1ms mark. A better approach to testing is not to measure the maximum possible IOPS, but to determine how many IOPS a storage array can sustain before average latency is greater than 1ms.

Oracle Automatic Workload Repository and benchmarking

The gold standard for Oracle performance comparisons is an Oracle Automatic Workload Repository (AWR) report.

There are multiple types of AWR reports. From a storage point of view, a report generated by running the `awrrpt.sql` command is the most comprehensive and valuable because it targets a specific database instance and includes some detailed histograms that break down storage I/O events based on latency.

Comparing two performance arrays ideally involves running the same workload on each array and producing an AWR report that precisely targets the workload. In the case of a very long-running workload, a single AWR report with an elapsed time that encompasses the start and stop time can be used, but it is preferable to break out the AWR data as multiple reports. For example, if a batch job ran from midnight to 6 a.m., create a series of one-hour AWR reports from midnight–1 a.m., 1 a.m.–2 a.m., and so on.

In other cases, a very short query should be optimized. The best option is an AWR report based on an AWR snapshot created when the query begins and a second AWR snapshot created when the query ends. The database server should be otherwise quiet to minimize the background activity that would obscure the activity of the query under analysis.

> Where AWR reports are not available, Oracle statspack reports are a good alternative. They contain most of the same I/O statistics as an AWR report.

Oracle AWR and troubleshooting

An AWR report is also the most important tool for analyzing a performance problem.

As with benchmarking, performance troubleshooting requires that you precisely measure a particular workload. When possible, provide AWR data when reporting a performance problem to the NetApp support center or when working with a NetApp or partner account team about a new solution.

When providing AWR data, consider the following requirements:

- Run the `awrrpt.sql` command to generate the report. The output can be either text or HTML.
- If Oracle Real Application Clusters (RACs) are used, generate AWR reports for each instance in the cluster.
- Target the specific time the problem existed. The maximum acceptable elapsed time of an AWR report is generally one hour. If a problem persists for multiple hours or involves a multihour operation such as a batch job, provide multiple one-hour AWR reports that cover the entire period to be analyzed.
- If possible, adjust the AWR snapshot interval to 15 minutes. This setting allows a more detailed analysis to be performed. This also requires additional executions of `awrrpt.sql` to provide a report for each 15-minute interval.
- If the problem is a very short running query, provide an AWR report based on an AWR snapshot created when the operation begins and a second AWR snapshot created when the operation ends. The database server should be otherwise quiet to minimize the background activity that would obscure the activity of the operation under analysis.
- If a performance problem is reported at certain times but not others, provide additional AWR data that demonstrates good performance for comparison.

calibrate_io

The `calibrate_io` command should never be used to test, compare, or benchmark storage systems. As stated in the Oracle documentation, this procedure calibrates the I/O capabilities of storage.

Calibration is not the same as benchmarking. The purpose of this command is to issue I/O to help calibrate database operations and improve their efficiency by optimizing the level of I/O issued to the host. Because the type of I/O performed by the `calibrate_io` operation does not represent actual database user I/O, the results are not predictable and are frequently not even reproducible.

SLOB2

SLOB2, the Silly Little Oracle Benchmark, has become the preferred tool for evaluating database performance. It was developed by Kevin Closson and is available at https://kevinclosson.net/slob/. It takes minutes to install and configure, and it uses an actual Oracle database to generate I/O patterns on a user-definable tablespace. It is one of the few testing options available that can saturate an all-flash array with I/O. It is also useful for generating much lower levels of I/O to simulate storage workloads that are low IOPS but latency sensitive.

Swingbench

Swingbench can be useful for testing database performance, but it is extremely difficult to use Swingbench in a way that stresses storage. NetApp has not seen any tests from Swingbench that yielded enough I/O to be a significant load on any AFF array. In limited cases, the Order Entry Test (OET) can be used to evaluate storage from a latency point of view. This could be useful in situations where a database has a known latency dependency for particular queries. Care must be taken to make sure that the host and network are properly configured to realize the latency potentials of an all-flash array.

HammerDB

HammerDB is a database testing tool that simulates TPC-C and TPC-H benchmarks, among others. It can take a lot of time to construct a sufficiently large data set to properly execute a test, but it can be an effective tool for evaluating performance for OLTP and data warehouse applications.

Orion

The Oracle Orion tool was commonly used with Oracle 9, but it has not been maintained to ensure compatibility with changes in various host operation systems. It is rarely used with Oracle 10 or Oracle 11 due to incompatibilities with OS and storage configuration.

Oracle rewrote the tool, and it is installed by default with Oracle 12c. Although this product has been improved and uses many of the same calls that a real Oracle database uses, it does not use precisely the same code path or I/O behavior used by Oracle. For example, most Oracle I/Os are performed synchronously, meaning the database halts until the I/O is complete as the I/O operation completes in the foreground. Simply flooding a storage system with random I/Os is not a reproduction of real Oracle I/O and does not offer a direct method of comparing storage arrays or measuring the effect of configuration changes.

That said, there are some use cases for Orion, such as general measurement of the maximum possible performance of a particular host-network-storage configuration, or to gauge the health of a storage system. With careful testing, usable Orion tests could be devised to compare storage arrays or evaluate the effect of a configuration change so long as the parameters include consideration of IOPS, throughput, and latency and attempt to faithfully replicate a realistic workload.

Stale NFSv3 locks and Oracle databases

If an Oracle database server crashes, it might have problems with stale NFS locks upon restart. This problem is avoidable by paying careful attention to the configuration of name resolution on the server.

This problem arises because creating a lock and clearing a lock use two slightly different methods of name resolution. Two processes are involved, the Network Lock Manager (NLM) and the NFS client. The NLM uses `uname n` to determine the host name, while the `rpc.statd` process uses `gethostbyname()`. These host names must match for the OS to properly clear stale locks. For example, the host might be looking for locks owned by `dbserver5`, but the locks were registered by the host as `dbserver5.mydomain.org`. If `gethostbyname()` does not return the same value as `uname -a`, then the lock release process did not succeed.

The following sample script verifies whether name resolution is fully consistent:

```
#! /usr/bin/perl
$uname=`uname -n`;
chomp($uname);
($name, $aliases, $addrtype, $length, @addrs) = gethostbyname $uname;
print "uname -n yields: $uname\n";
print "gethostbyname yields: $name\n";
```

If `gethostbyname` does not match `uname`, stale locks are likely. For example, this result reveals a potential problem:

```
uname -n yields: dbserver5
gethostbyname yields: dbserver5.mydomain.org
```

The solution is usually found by changing the order in which hosts appear in `/etc/hosts`. For example, assume that the hosts file includes this entry:

```
10.156.110.201  dbserver5.mydomain.org dbserver5 loghost
```

To resolve this issue, change the order in which the fully qualified domain name and the short host name appear:

```
10.156.110.201  dbserver5 dbserver5.mydomain.org loghost
```

`gethostbyname()` now returns the short `dbserver5` host name, which matches the output of `uname`. Locks are thus cleared automatically after a server crash.

WAFL alignment verification for Oracle databases

Correct WAFL alignment is critical for good performance. Although ONTAP manages blocks in 4KB units, this fact does not mean that ONTAP performs all operations in 4KB units. In fact, ONTAP supports block operations of different sizes, but the underlying accounting is managed by WAFL in 4KB units.

The term “alignment” refers to how Oracle I/O corresponds to these 4KB units. Optimum performance requires an Oracle 8KB block to reside on two 4KB WAFL physical blocks on a drive. If a block is offset by 2KB, this block resides on half of one 4KB block, a separate full 4KB block, and then half of a third 4KB block. This arrangement causes performance degradation.

Alignment is not a concern with NAS file systems. Oracle datafiles are aligned to the start of the file based on the size of the Oracle block. Therefore, block sizes of 8KB, 16KB, and 32KB are always aligned. All block operations are offset from the start of the file in units of 4 kilobytes.

LUNs, in contrast, generally contain some kind of driver header or file system metadata at their start that creates an offset. Alignment is rarely a problem in modern OSs because these OSs are designed for physical drives that might use a native 4KB sector, which also requires I/O to be aligned to 4KB boundaries for optimum performance.

There are, however, some exceptions. A database might have been migrated from an older OS that was not optimized for 4KB I/O, or user error during partition creation might have led to an offset that is not in units of 4KB in size.

The following examples are Linux-specific, but the procedure can be adapted for any OS.

Aligned

The following example shows an alignment check on a single LUN with a single partition.

First, create the partition that uses all partitions available on the drive.

```
[root@host0 iscsi]# fdisk /dev/sdb
Device contains neither a valid DOS partition table, nor Sun, SGI or OSF
disklabel
Building a new DOS disklabel with disk identifier 0xb97f94c1.
Changes will remain in memory only, until you decide to write them.
After that, of course, the previous content won't be recoverable.
The device presents a logical sector size that is smaller than
the physical sector size. Aligning to a physical sector (or optimal
I/O) size boundary is recommended, or performance may be impacted.
Command (m for help): n
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
First cylinder (1-10240, default 1):
Using default value 1
Last cylinder, +cylinders or +size{K,M,G} (1-10240, default 10240):
Using default value 10240
Command (m for help): w
The partition table has been altered!
Calling ioctl() to re-read partition table.
Syncing disks.
[root@host0 iscsi]#
```

The alignment can be checked mathematically with the following command:

```
[root@host0 iscsi]# fdisk -u -l /dev/sdb
Disk /dev/sdb: 10.7 GB, 10737418240 bytes
64 heads, 32 sectors/track, 10240 cylinders, total 20971520 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 4096 bytes
I/O size (minimum/optimal): 4096 bytes / 65536 bytes
Disk identifier: 0xb97f94c1
   Device Boot      Start         End      Blocks   Id  System
/dev/sdb1              32    20971519    10485744   83  Linux
```

The output shows that the units are 512 bytes, and the start of the partition is 32 units. This is a total of 32 x 512 = 16,834 bytes, which is a whole multiple of 4KB WAFL blocks. This partition is correctly aligned.

To verify correct alignment, complete the following steps:

1. Identify the universally unique identifier (UUID) of the LUN.

```
FAS8040SAP::> lun show -v /vol/jfs_luns/lun0
              Vserver Name: jfs
                  LUN UUID: ed95d953-1560-4f74-9006-85b352f58fcd
                    Mapped: mapped`                `
```

2. Enter the node shell on the ONTAP controller.

```
FAS8040SAP::> node run -node FAS8040SAP-02
Type 'exit' or 'Ctrl-D' to return to the CLI
FAS8040SAP-02> set advanced
set not found.  Type '?' for a list of commands
FAS8040SAP-02> priv set advanced
Warning: These advanced commands are potentially dangerous; use
         them only when directed to do so by NetApp
         personnel.
```

3. Start statistical collections on the target UUID identified in the first step.

```
FAS8040SAP-02*> stats start lun:ed95d953-1560-4f74-9006-85b352f58fcd
Stats identifier name is 'Ind0xffffff08b9536188'
FAS8040SAP-02*>
```

4. Perform some I/O. It is important to use the `iflag` argument to make sure that I/O is synchronous and not buffered.

> Be very careful with this command. Reversing the `if` and `of` arguments destroys data.

```
[root@host0 iscsi]# dd if=/dev/sdb1 of=/dev/null iflag=dsync count=1000
bs=4096
1000+0 records in
1000+0 records out
4096000 bytes (4.1 MB) copied, 0.0186706 s, 219 MB/s
```

5. Stop the stats and view the alignment histogram. All I/O should be in the `.0` bucket, which indicates I/O that is aligned to a 4KB block boundary.

```
FAS8040SAP-02*> stats stop
StatisticsID: Ind0xffffff08b9536188
lun:ed95d953-1560-4f74-9006-85b352f58fcd:instance_uuid:ed95d953-1560-
4f74-9006-85b352f58fcd
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.0:186%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.1:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.2:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.3:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.4:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.5:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.6:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.7:0%
```

Misaligned

The following example shows misaligned I/O:

1. Create a partition that does not align to a 4KB boundary. This is not default behavior on modern OSs.

```
[root@host0 iscsi]# fdisk -u /dev/sdb
Command (m for help): n
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
First sector (32-20971519, default 32): 33
Last sector, +sectors or +size{K,M,G} (33-20971519, default 20971519):
Using default value 20971519
Command (m for help): w
The partition table has been altered!
Calling ioctl() to re-read partition table.
Syncing disks.
```

2. The partition has been created with a 33-sector offset instead of the default 32. Repeat the procedure outlined in Aligned. The histogram appears as follows:

```
FAS8040SAP-02*> stats stop
StatisticsID: Ind0xffffff0468242e78
lun:ed95d953-1560-4f74-9006-85b352f58fcd:instance_uuid:ed95d953-1560-
4f74-9006-85b352f58fcd
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.0:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.1:136%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.2:4%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.3:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.4:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.5:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.6:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.7:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_partial_blocks:31%
```

The misalignment is clear. The I/O mostly falls into the* *`.1` bucket, which matches the expected offset. When the partition was created, it was moved 512 bytes further into the device than the optimized default, which means that the histogram is offset by 512 bytes.

Additionally, the `read_partial_blocks` statistic is nonzero, which means I/O was performed that did not fill up an entire 4KB block.

Redo logging

The procedures explained here are applicable to datafiles. Oracle redo logs and archive logs have different I/O patterns. For example, redo logging is a circular overwrite of a single file. If the default 512-byte block size is used, the write statistics look something like this:

```
FAS8040SAP-02*> stats stop
StatisticsID: Ind0xffffff0468242e78
lun:ed95d953-1560-4f74-9006-85b352f58fcd:instance_uuid:ed95d953-1560-4f74-
9006-85b352f58fcd
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.0:12%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.1:8%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.2:4%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.3:10%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.4:13%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.5:6%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.6:8%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.7:10%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_partial_blocks:85%
```

The I/O would be distributed across all histogram buckets, but this is not a performance concern. Extremely high redo-logging rates might, however, benefit from the use of a 4KB block size. In this case, it is desirable to make sure that the redo-logging LUNs are properly aligned. However, this is not as critical to good performance as datafile alignment.

PostgreSQL

PostgreSQL databases on ONTAP

PostgreSQL comes with variants that include PostgreSQL, PostgreSQL Plus, and EDB Postgres Advanced Server (EPAS). PostgreSQL is typically deployed as the back-end database for multitier applications. It is supported by common middleware packages (such as PHP, Java, Python, Tcl/Tk, ODBC, and JDBC) and has historically been a popular choice for open-source database management systems. ONTAP is a excellent choice for running PostgreSQL databases do to its reliability, high performing and efficent data management capabilities.

> This documentation on ONTAP and the PostgreSQL database replaces the previously published *TR-4770: PostgreSQL database on ONTAP best practices.*

As data grows exponentially, data management becomes more complex for enterprises. This complexity increases licensing, operational, support, and maintenance costs. To reduce the overall TCO, consider switching from commercial to open-source databases with reliable, high-performing back-end storage.

ONTAP is an ideal platform because ONTAP is literally designed for databases. Numerous features such as random IO latency optimizations to advanced quality of service (QoS) to basic FlexClone functionality were created specifically to address the needs of database workloads.

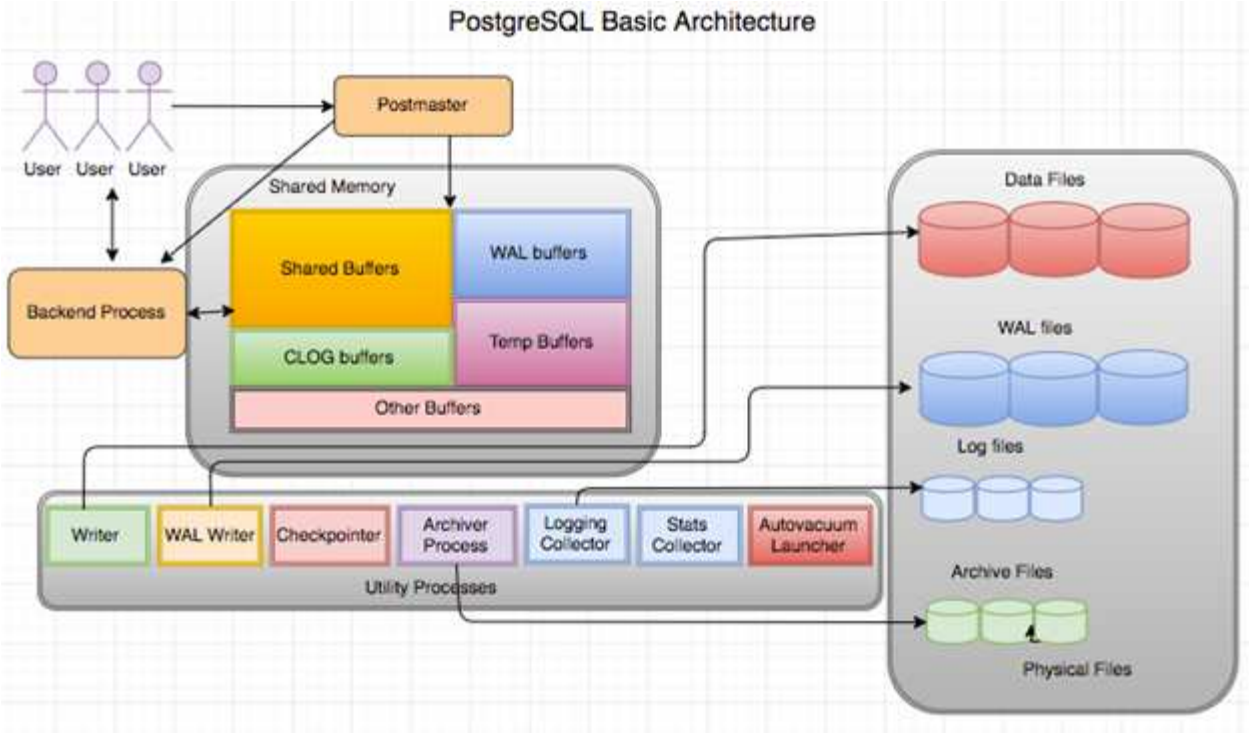
Additional features such as nondisruptive upgrades, (including storage replacement) ensure that your critical databases remain available. You can also have instant disaster recovery for large environments through MetroCluster, or select databases using SnapMirror active sync.

Most importantly, ONTAP delivers unmatched performance with the ability to size the solution for your unique needs. Our high-end systems can deliver over 1M IOPS with latencies measured in microseconds, but if you only need 100K IOPS you can rightsize your storage solution with a smaller controller that still runs the exact same storage operating system.

Database configuration

PostgreSQL architecture

PostgreSQL is an RDBMS based on client and server architecture. A PostgreSQL instance is known as a database cluster, which is a collection of databases as opposed to a collection of servers.



There are three major elements in a PostgreSQL database: the postmaster, the front end (client), and the back end. The client sends requests to the postmaster with information such as IP protocol and which database to connect to. The postmaster authenticates the connection and passes it to the back-end process for further communication. The back-end process executes the query and sends results directly to the front end (client).

A PostgreSQL instance is based on a multiprocess model instead of a multithreaded model. It spawns multiple processes for different jobs, and each process has its own functionality. The major processes include the client process, the WAL writer process, the background writer process, and the checkpointer process:

- When a client (foreground) process sends read or write requests to the PostgreSQL instance, it doesn't read or write data directly to the disk. It first buffers the data in shared buffers and write-ahead logging (WAL) buffers.
- A WAL writer process manipulates the content of the shared buffers and WAL buffers to write into the WAL logs. WAL logs are typically transaction logs of PostgreSQL and are sequentially written. Therefore, to improve the response time from the database, PostgreSQL first writes into the transaction logs and acknowledges the client.
- To put the database in a consistent state, the background writer process checks the shared buffer periodically for dirty pages. It then flushes the data onto the data files that are stored on NetApp volumes or LUNs.
- The checkpointer process also runs periodically (less frequently than the background process) and prevents any modification to the buffers. It signals to the WAL writer process to write and flush the checkpoint record to the end of WAL logs that are stored on the NetApp disk. It also signals the background writer process to write and flush all dirty pages to the disk.

PostgreSQL initialization parameters

You create a new database cluster by using the `initdb` program. An `initdb` script creates the data files, system tables, and template databases (template0 and template1) that define the cluster.

The template database represents a stock database. It contains definitions for system tables, standard views, functions, and data types. `pgdata` acts as an argument to the `initdb` script that specifies the location of the database cluster.

All the database objects in PostgreSQL are internally managed by respective OIDs. Tables and indexes are also managed by individual OIDs. The relationships between database objects and their respective OIDs are stored in appropriate system catalog tables, depending on the type of object. For example, OIDs of databases and heap tables are stored in `pg_database` and `pg_class, respectively. You can determine the OIDs by issuing queries on the PostgreSQL client.

Every database has its own individual tables and index files that are restricted to 1GB. Each table has two associated files, suffixed respectively with `_fsm` and `_vm`. They are referred to as the free space map and the visibility map. These files store the information about free space capacity and have visibility on each page in the table file. Indexes only have individual free space maps and don't have visibility maps.

The `pg_xlog/pg_wal` directory contains the write-ahead logs. Write-ahead logs are used to improve database reliability and performance. Whenever you update a row in a table, PostgreSQL first writes the change to the write-ahead log, and later writes the modifications to the actual data pages to a disk. The `pg_xlog` directory usually contains several files, but initdb creates only the first one. Extra files are added as needed. Each xlog file is 16MB long.

PostgreSQL database configuration with ONTAP

There are several PostgreSQL tuning configurations that can improve performance.

The most commonly used parameters are as follows:

- `max_connections = <num>`: The maximum number of database connections to have at one time. Use this parameter to restrict swapping to disk and killing the performance. Depending on your application requirement, you can also tune this parameter for the connection pool settings.
- `shared_buffers = <num>`: The simplest method for improving the performance of your database server. The default is low for most modern hardware. It is set during deployment to approximately 25% of available RAM on the system. This parameter setting varies depending on how it works with particular database instances; you might have to increase and decrease the values by trial and error. However, setting it high is likely to degrade performance.
- `effective_cache_size = <num>`: This value tells PostgreSQL's optimizer how much memory PostgreSQL has available for caching data and helps in determining whether to use an index. A larger value increases the likelihood of using an index. This parameter should be set to the amount of memory allocated to `shared_buffers` plus the amount of OS cache available. Often this value is more than 50% of the total system memory.
- `work_mem = <num>`: This parameter controls the amount of memory to be used in sort operations and hash tables. If you do heavy sorting in your application, you might need to increase the amount of memory, but be cautious. It isn't a system wide parameter, but a per-operation one. If a complex query has several sort operations in it, it uses multiple work_mem units of memory, and multiple back ends could be doing this simultaneously. This query can often lead your database server to swap if the value is too large. This option was previously called sort_mem in older versions of PostgreSQL.
- `fsync = <boolean> (on or off)`: This parameter determines whether all your WAL pages should be synchronized to disk by using fsync() before a transaction is committed. Turning it off can sometimes improve write performance and turning it on increases protection from the risk of corruption when the system crashes.
- `checkpoint_timeout`: The checkpoint process flushes committed data to disk. This involves a lot of

read/write operations on disk. The value is set in seconds and lower values decrease crash recovery time and increasing values can reduce the load on system resources by reducing the checkpoint calls. Depending on application criticality, usage, database availability, set the value of checkpoint_timeout.

- `commit_delay = <num>` and `commit_siblings = <num>`: These options are used together to help improve performance by writing out multiple transactions that are committing at once. If there are several commit_siblings objects active at the instant your transaction is committing, the server waits for commit_delay microseconds to try to commit multiple transactions at once.
- `max_worker_processes / max_parallel_workers`: Configure the optimal number of workers for processes. Max_parallel_workers correspond to the number of CPUs available. Depending on application design, queries might require a lesser number of workers for parallel operations. It is better to keep the value for both parameters the same but adjust the value after testing.
- `random_page_cost = <num>`: This value controls the way PostgreSQL views non-sequential disk reads. A higher value means PostgreSQL is more likely to use a sequential scan instead of an index scan, indicating that your server has fast disks Modify this setting after evaluating other options like plan-based optimization, vacuuming, indexing to altering queries or schema.
- `effective_io_concurrency = <num>`: This parameter sets the number of concurrent disk I/O operations that PostgreSQL attempts to execute simultaneously. Raising this value increases the number of I/O operations that any individual PostgreSQL session attempts to initiate in parallel. The allowed range is 1 to 1,000, or zero to disable issuance of asynchronous I/O requests. Currently, this setting only affects bitmap heap scans. Solid-state drives (SSDs) and other memory-based storage (NVMe) can often process many concurrent requests, so the best value can be in the hundreds.

See the PostgreSQL documentation for a complete list of PostgreSQL configuration parameters.

TOAST

TOAST stands for The Oversized-Attribute Storage Technique. PostgreSQL uses a fixed page size (commonly 8KB) and does not allow tuples to span multiple pages. Therefore, it is not possible to store large field values directly. When you attempt to store a row that exceeds this size, TOAST breaks up the data of large columns into smaller “pieces” and stores them in a TOAST table.

The large values of toasted attributes are pulled out (if selected at all) only at the time the result set is sent to the client. The table itself is much smaller and can fit more rows into the shared buffer cache than it could without any out-of-line storage (TOAST).

VACUUM

In normal PostgreSQL operation, tuples that are deleted or made obsolete by an update are not physically removed from their table; they remain present until VACUUM is run. Therefore, you must run VACUUM periodically, especially on frequently updated tables. The space it occupies must then be reclaimed for reuse by new rows, to avoid disk space outage. However, it does not return the space to the operating system.

The free space inside a page is not fragmented. VACUUM rewrites the entire block, efficiently packing the remaining rows and leaving a single contiguous block of free space in a page.

In contrast, VACUUM FULL actively compacts tables by writing a completely new version of the table file with no dead space. This action minimizes the size of the table but can take a long time. It also requires extra disk space for the new copy of the table until the operation completes. The goal of routine VACUUM is to avoid VACUUM FULL activity. This process not only keeps tables at their minimum size, but also maintains steady-state usage of disk space.

PostgreSQL tablespaces

Two tablespaces are automatically created when the database cluster is initialized.

The `pg_global` tablespace is used for shared system catalogs. The `pg_default` tablespace is the default tablespace of the template1 and template0 databases. If the partition or volume on which the cluster was initialized runs out of space and cannot be extended, a tablespace can be created on a different partition and used until the system can be reconfigured.

An index that is heavily used can be placed on a fast, highly available disk, like a solid-state device. Also, a table storing archived data that is rarely used or not performance critical can be stored on a less expensive, slower disk system like SAS or SATA drives.

Tablespaces are a part of the database cluster and cannot be treated as an autonomous collection of data files. They depend on metadata contained in the main data directory, and therefore cannot be attached to a different database cluster or backed up individually. Similarly, if you lose a tablespace (through file deletion, disk failure, and so on), the database cluster might become unreadable or unable to start. Placing a tablespace on a temporary file system like a RAM disk risks the reliability of the entire cluster.

After it is created, a tablespace can be used from any database if the requesting user has sufficient privileges. PostgreSQL uses symbolic links to simplify the implementation of tablespaces. PostgreSQL adds a row to the `pg_tablespace` table (a clusterwide table) and assigns a new object identifier (OID) to that row. Finally, the server uses the OID to create a symbolic link between your cluster and the given directory. The directory `$PGDATA/pg_tblspc` contains symbolic links that point to each of the non-built-in tablespaces defined in the cluster.

Storage configuration

PostgreSQL databases with NFS Filesystems

PostgreSQL databases can be hosted on NFSv3 or NFSv4 filesystems. The best option depends on factors outside the database.

For example, NFSv4 locking behavior may be preferable in certain clustered environments. (See here for additional details)

Database functionality should otherwise be close to identical, including performance. The only requirement is the use of the `hard` mount option. This is required to ensure soft timeouts do not produce unrecoverable IO errors.

If NFSv4 is chosen as a protocol, NetApp recommends using NFSv4.1. There are some functional enhancements to the NFSv4 protocol in NFSv4.1 that improve resiliency over NFSv4.0.

Use the following mount options for general database workloads:

```
rw,hard,nointr,bg,vers=[3|4],proto=tcp,rsize=65536,wsize=65536
```

If heavy sequential IO is expected, the NFS transfer sizes can be increased as described in the following section.

NFS Transfer Sizes

By default, ONTAP limits NFS I/O sizes to 64K.

Random I/O with an most applications and databases uses a much smaller block size which is well below the 64K maximum. Large-block I/O is usually parallelized, so the 64K maximum is also not a limitation to obtaining maximum bandwidth.

There are some workloads where the 64K maximum does create a limitation. In particular, single-threaded operations such as backup or recovery operation or a database full table scan run faster and more efficiently if the database can perform fewer but larger I/Os. The optimum I/O handling size for ONTAP is 256K.

The maximum transfer size for a given ONTAP SVM can be changed as follows:

```
Cluster01::> set advanced
Warning: These advanced commands are potentially dangerous; use them only
when directed to do so by NetApp personnel.
Do you want to continue? {y|n}: y
Cluster01::*> nfs server modify -vserver vserver1 -tcp-max-xfer-size
262144
Cluster01::*>
```

Caution
Never decrease the maximum allowable transfer size on ONTAP below the value of rsize/wsize of currently mounted NFS file systems. This can create hangs or even data corruption with some operating systems. For example, if NFS clients are currently set at an rsize/wsize of 65536, then the ONTAP maximum transfer size could be adjusted between 65536 and 1048576 with no effect because the clients themselves are limited. Reducing the maximum transfer size below 65536 can damage availability or data.

Once the transfer size is increased at the ONTAP level, the following mount options would be used:

```
rw,hard,nointr,bg,vers=[3|4],proto=tcp,rsize=262144,wsize=262144
```

NFSv3 TCP Slot Tables

If NFSv3 is used with Linux, it is critical to properly set the TCP slot tables.

TCP slot tables are the NFSv3 equivalent of host bus adapter (HBA) queue depth. These tables control the number of NFS operations that can be outstanding at any one time. The default value is usually 16, which is far too low for optimum performance. The opposite problem occurs on newer Linux kernels, which can automatically increase the TCP slot table limit to a level that saturates the NFS server with requests.

For optimum performance and to prevent performance problems, adjust the kernel parameters that control the TCP slot tables.

Run the `sysctl -a | grep tcp.*.slot_table` command, and observe the following parameters:

```
# sysctl -a | grep tcp.*.slot_table
sunrpc.tcp_max_slot_table_entries = 128
sunrpc.tcp_slot_table_entries = 128
```

All Linux systems should include `sunrpc.tcp_slot_table_entries`, but only some include `sunrpc.tcp_max_slot_table_entries`. They should both be set to 128.

Caution
Failure to set these parameters may have significant effects on performance. In some cases, performance is limited because the linux OS is not issuing sufficient I/O. In other cases, I/O latencies increases as the linux OS attempts to issue more I/O than can be serviced.

PostgreSQL with SAN Filesystems

PostgreSQL databases with SAN are generally hosted on xfs filesystems, but others can be used if supported by the OS vendor

While a single LUN can generally support up to 100K IOPS, IO-intensive databases generally require the use of LVM with striping.

LVM Striping

Before the era of flash drives, striping was used to help overcome the performance limitations of spinning drives. For example, if an OS needs to perform a 1MB read operation, reading that 1MB of data from a single drive would require a lot of drive head seeking and reading as the 1MB is slowly transferred. If that 1MB of data was striped across 8 LUNs, the OS could issue eight 128K read operations in parallel and reduce the time required to complete the 1MB transfer.

Striping with spinning drives was more difficult because the I/O pattern had to be known in advance. If the striping wasn't correctly tuned for the true I/O patterns, striped configurations could damage performance. With Oracle databases, and especially with all-flash configurations, striping is much easier to configure and has been proven to dramatically improve performance.

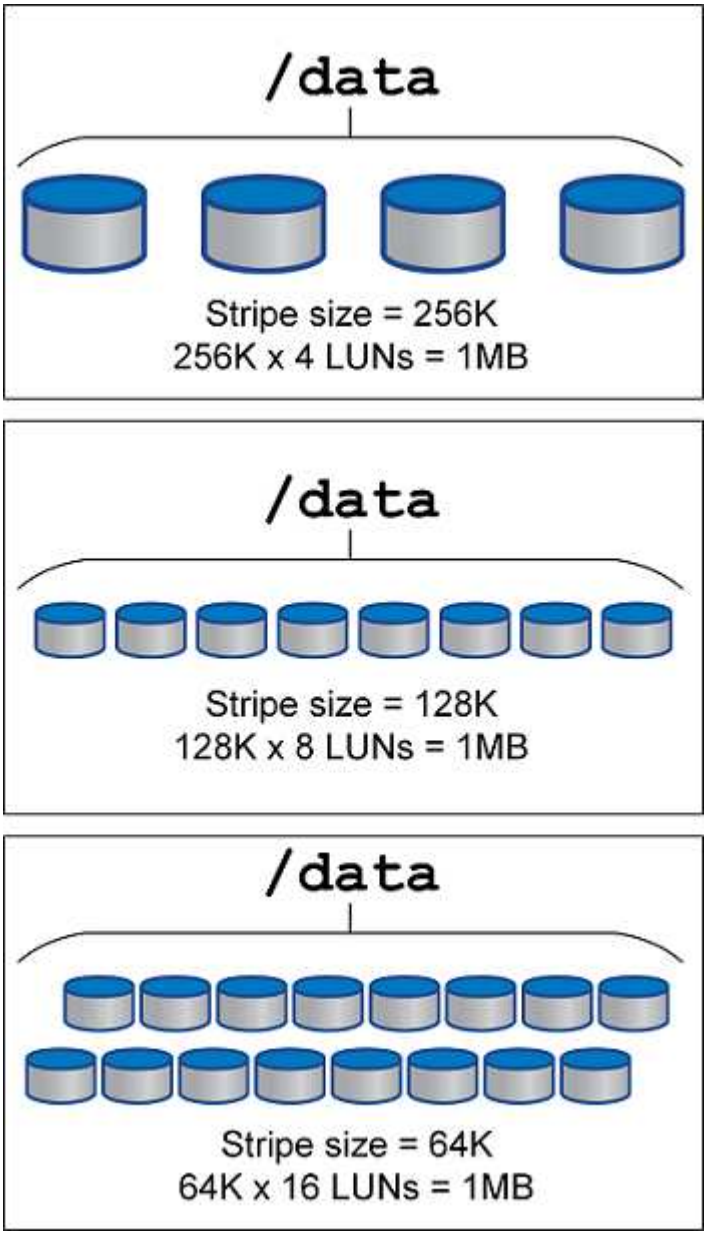
Logical volume managers such as Oracle ASM stripe by default, but native OS LVM do not. Some of them bond multiple LUNs together as a concatenated device, which results in datafiles that exist on one and only one LUN device. This causes hot spots. Other LVM implementations default to distributed extents. This is similar to striping, but it's coarser. The LUNs in the volume group are sliced into large pieces, called extents and typically measured in many megabytes, and the logical volumes are then distributed across those extents. The result is random I/O against a file should be well distributed across LUNs, but sequential I/O operations are not as efficient as they could be.

Performance-intensive application I/O is nearly always either (a) in units of the basic block size or (b) one megabyte.

The primary goal of a striped configuration is to ensure that single-file I/O can be performed as a single unit, and multiblock I/Os, which should be 1MB in size, can be parallelized evenly across all LUNs in the striped volume. This means that the stripe size must not be smaller than the database block size, and the stripe size multiplied by the number of LUNs should be 1MB.

The following figure shows three possible options for stripe size and width tuning. The number of LUNs is selected to meet performance requirements as described above, but in all cases the total data within a single

stripe is 1MB.



Data protection

PostgreSQL data protection

One of the major aspects of storage design is enabling protection for PostgreSQL volumes. Customers can protect their PostgreSQL databases either by using the dump approach or by using file system backups. This section explains the different approaches of backing up individual databases or the entire cluster.

There are three approaches to backing up PostgreSQL data:

- SQL Server dump
- File-system-level backup

- Continuous archiving

The idea behind the SQL Server dump method is to generate a file with SQL Server commands that, when returned to the server, can re-create the database as it was at the time of the dump. PostgreSQL provides the utility programs `pg_dump` and `pg_dump_all` for creating individual and cluster-level backup. These dumps are logical and do not contain enough information to be used by WAL replay.

An alternative backup strategy is to use file-system-level backup, in which administrators directly copy the files that PostgreSQL uses to store the data in the database. This method is done in offline mode: the database or cluster must be shut down. Another alternative is to use `pg_basebackup` to run hot streaming backup of the PostgreSQL database.

PostgreSQL databases and storage snapshots

Snapshot-based backups with PostgreSQL requires configuration of snapshots for datafiles, WAL files, and archived WAL files to provide full or point-in-time recovery.

For PostgreSQL databases, the average backup time with snapshots is in the range of a few seconds to a few minutes. This backup speed is 60 to 100 times faster than `pg_basebackup` and other file-system-based backup approaches.

Snapshots on NetApp storage can be both crash-consistent and application-consistent. A crash-consistent snapshot is created on storage without quiescing the database, whereas an application-consistent snapshot is created while the database is in backup mode. NetApp also ensures that subsequent snapshots are incremental-forever backups to promote storage savings and network efficiency.

Because snapshots are rapid and do not affect system performance, you can schedule multiple snapshots daily instead of creating a single daily backup as with other streaming backup technology. When a restore and recovery operation is necessary, the system downtime is reduced by two key features:

- NetApp SnapRestore data recovery technology means that the restore operation is executed in seconds.
- Aggressive recovery point objectives (RPOs) mean that fewer database logs must be applied and forward recovery is also accelerated.

For backing up PostgreSQL, you must ensure that the data volumes are protected simultaneously with (consistency-group) WAL and the archived logs. While you are using Snapshot technology to copy WAL files, make sure that you run `pg_stop` to flush all the WAL entries that must be archived. If you flush the WAL entries during the restore, then you only need to stop the database, unmount, or delete the existing data directory and perform a SnapRestore operation on storage. After the restore is done, you can mount the system and bring it back to its current state. For point-in-time recovery, you can also restore WAL and archive logs; then PostgreSQL decides the most consistent point and recovers it automatically.

Consistency groups are a feature in ONTAP and are recommended when there are multiple volumes mounted to a single instance or a database with multiple tablespaces. A consistency group snapshot ensures all volumes are grouped together and protected. A consistency group can be managed efficiently from ONTAP System Manager and you can even clone it to create an instance copy of a database for testing or development purposes.

For more information on Consistency groups, see the NetApp Consistency groups overview.

PostgreSQL data protection software

NetApp SnapCenter plugin for PostgreSQL database, combined with Snapshot and

NetApp FlexClone technologies, offer you benefits such as:

- Rapid backup and restore.
- Space-efficient clones.
- The ability to build a speedy and effective disaster recovery system.

> You might prefer to choose NetApp's premium backup partners such as Veeam Software and Commvault under the following circumstances:
>
> - Managing workloads across a heterogenous environment
> - Storing backups to either cloud or tape for long-term retention
> - Support for a wide range of OS versions and types

SnapCenter plugin for PostgreSQL is community supported plugin and the setup and documentation is available on NetApp Automation store. Through SnapCenter, user can backup database, clone and restore data remotely.

VMware

VMware vSphere with ONTAP

VMware vSphere with ONTAP

ONTAP has been a leading storage solution for VMware vSphere environments for almost two decades and continues to add innovative capabilities to simplify management while reducing costs. This document introduces the ONTAP solution for vSphere, including the latest product information and best practices, to streamline deployment, reduce risk, and simplify management.

> This documentation replaces previously published technical reports *TR-4597: VMware vSphere for ONTAP*

Best practices supplement other documents such as guides and compatibility lists. They are developed based on lab testing and extensive field experience by NetApp engineers and customers. They might not be the only supported practices that work in every environment, but they are generally the simplest solutions that meet the needs of most customers.

This document is focused on capabilities in recent releases of ONTAP (9.x) running on vSphere 7.0 or later. See the NetApp Interoperability Matrix Tool and VMware Compatibility Guide for details related to specific releases.

Why ONTAP for vSphere?

There are many reasons why tens of thousands of customers have selected ONTAP as their storage solution for vSphere, such as a unified storage system supporting both SAN and NAS protocols, robust data protection capabilities using space-efficient snapshots and a wealth of tools to help you manage application data. Using a storage system separate from the hypervisor allows you to offload many functions and maximize your investment in vSphere host systems. This approach not only makes sure your host resources are focused on application workloads, but it also avoids random performance effects on applications from storage operations.

Using ONTAP together with vSphere is a great combination that lets you reduce host hardware and VMware software expenses. You can also protect your data at lower cost with consistent high performance. Because virtualized workloads are mobile, you can explore different approaches using Storage vMotion to move VMs across VMFS, NFS, or vVols datastores, all on the same storage system.

Here are key factors customers value today:

- **Unified storage.** Systems running ONTAP software are unified in several significant ways. Originally this approach referred to both NAS and SAN protocols, and ONTAP continues to be a leading platform for SAN along with its original strength in NAS. In the vSphere world, this approach could also mean a unified system for virtual desktop infrastructure (VDI) together with virtual server infrastructure (VSI). Systems running ONTAP software are typically less expensive for VSI than traditional enterprise arrays and yet have advanced storage efficiency capabilities to handle VDI in the same system. ONTAP also unifies a variety of storage media, from SSDs to SATA, and can extend that easily into the cloud. There's no need to buy one flash array for performance, a SATA array for archives, and separate systems for the cloud. ONTAP ties them all together.
- **Virtual volumes and storage policy based management.** NetApp was an early design partner with VMware in the development of vSphere Virtual Volumes (vVols), providing architectural input and early

support for vVols and VMware vSphere APIs for Storage Awareness (VASA). Not only did this approach bring granular VM storage management to VMFS, it also supported automation of storage provisioning through storage policy based management. This approach allows storage architects to design storage pools with different capabilities that can be easily consumed by VM administrators. ONTAP leads the storage industry in vVol scale, supporting hundreds of thousands of vVols in a single cluster, whereas enterprise array and smaller flash array vendors support as few as several thousand vVols per array. NetApp is also driving the evolution of granular VM management with upcoming capabilities in support of vVols 3.0.

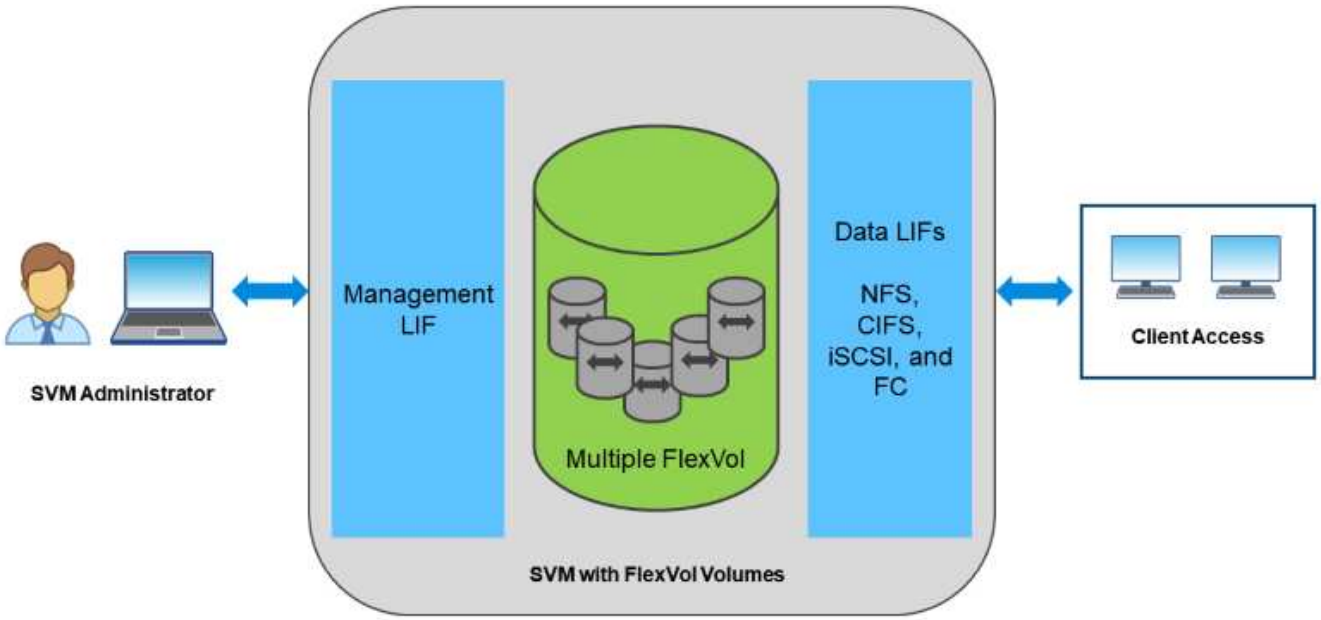
- **Storage efficiency.** Although NetApp was the first to deliver deduplication for production workloads, this innovation wasn't the first or last one in this area. It started with snapshots, a space-efficient data protection mechanism with no performance effect, along with FlexClone technology to instantly make read/write copies of VMs for production and backup use. NetApp went on to deliver inline capabilities, including deduplication, compression, and zero-block deduplication, to squeeze out the most storage from expensive SSDs. Most recently, ONTAP added the ability to pack smaller I/O operations and files into a disk block using compaction. The combination of these capabilities has resulted in customers seeing savings of up to 5:1 for VSI and up to 30:1 for VDI.
- **Hybrid cloud.** Whether used for on-premises private cloud, public cloud infrastructure, or a hybrid cloud that combines the best of both, ONTAP solutions help you build your data fabric to streamline and optimize data management. Start with high-performance all-flash systems, then couple them with either disk or cloud storage systems for data protection and cloud compute. Choose from Azure, AWS, IBM, or Google clouds to optimize costs and avoid lock-in. Leverage advanced support for OpenStack and container technologies as needed. NetApp also offers cloud-based backup (SnapMirror Cloud, Cloud Backup Service, and Cloud Sync) and storage tiering and archiving tools (FabricPool) for ONTAP to help reduce operating expenses and leverage the broad reach of the cloud.
- **And more.** Take advantage of the extreme performance of NetApp AFF A-Series arrays to accelerate your virtualized infrastructure while managing costs. Enjoy completely nondisruptive operations, from maintenance to upgrades to complete replacement of your storage system, using scale-out ONTAP clusters. Protect data at rest with NetApp encryption capabilities at no additional cost. Make sure performance meets business service levels through fine-grained quality of service capabilities. They are all part of the broad range of capabilities that come with ONTAP, the industry's leading enterprise data management software.

Unified Storage

NetApp ONTAP unifies storage through a simplified, software-defined approach for secure and efficient management, improved performance, and seamless scalability. This approach enhances data protection and enables effective use of cloud resources.

Originally this unified approach referred to supporting both NAS and SAN protocols on one storage system, and ONTAP continues to be a leading platform for SAN along with its original strength in NAS. ONTAP now also provides S3 object protocol support. Though S3 isn't used for datastores, you can use it for in-guest applications. You can learn more about the S3 protocol support in ONTAP in the S3 configuration overview.

A storage virtual machine (SVM) is the unit of secure multi-tenancy in ONTAP. It is a logical construct allowing client access to systems running ONTAP software. SVMs can serve data concurrently through multiple data access protocols via logical interfaces (LIFs). SVMs provide file-level data access through NAS protocols, such as CIFS and NFS, and block-level data access through SAN protocols, such as iSCSI, FC/FCoE, and NVMe. SVMs can serve data to SAN and NAS clients independently at the same time, as well as with S3.



In the vSphere world, this approach could also mean a unified system for virtual desktop infrastructure (VDI) together with virtual server infrastructure (VSI). Systems running ONTAP software are typically less expensive for VSI than traditional enterprise arrays and yet have advanced storage efficiency capabilities to handle VDI in the same system. ONTAP also unifies a variety of storage media, from SSDs to SATA, and can extend that easily into the cloud. There's no need to buy one flash array for performance, a SATA array for archives, and separate systems for the cloud. ONTAP ties them all together.

NOTE: For more information on SVMs, unified storage and client access, see Storage Virtualization in the ONTAP 9 Documentation center.

Virtualization tools for ONTAP

NetApp offers several standalone software tools that can be used together with ONTAP and vSphere to manage your virtualized environment.

The following tools are included with the ONTAP license at no additional cost. See Figure 1 for a depiction of how these tools work together in your vSphere environment.

ONTAP tools for VMware vSphere

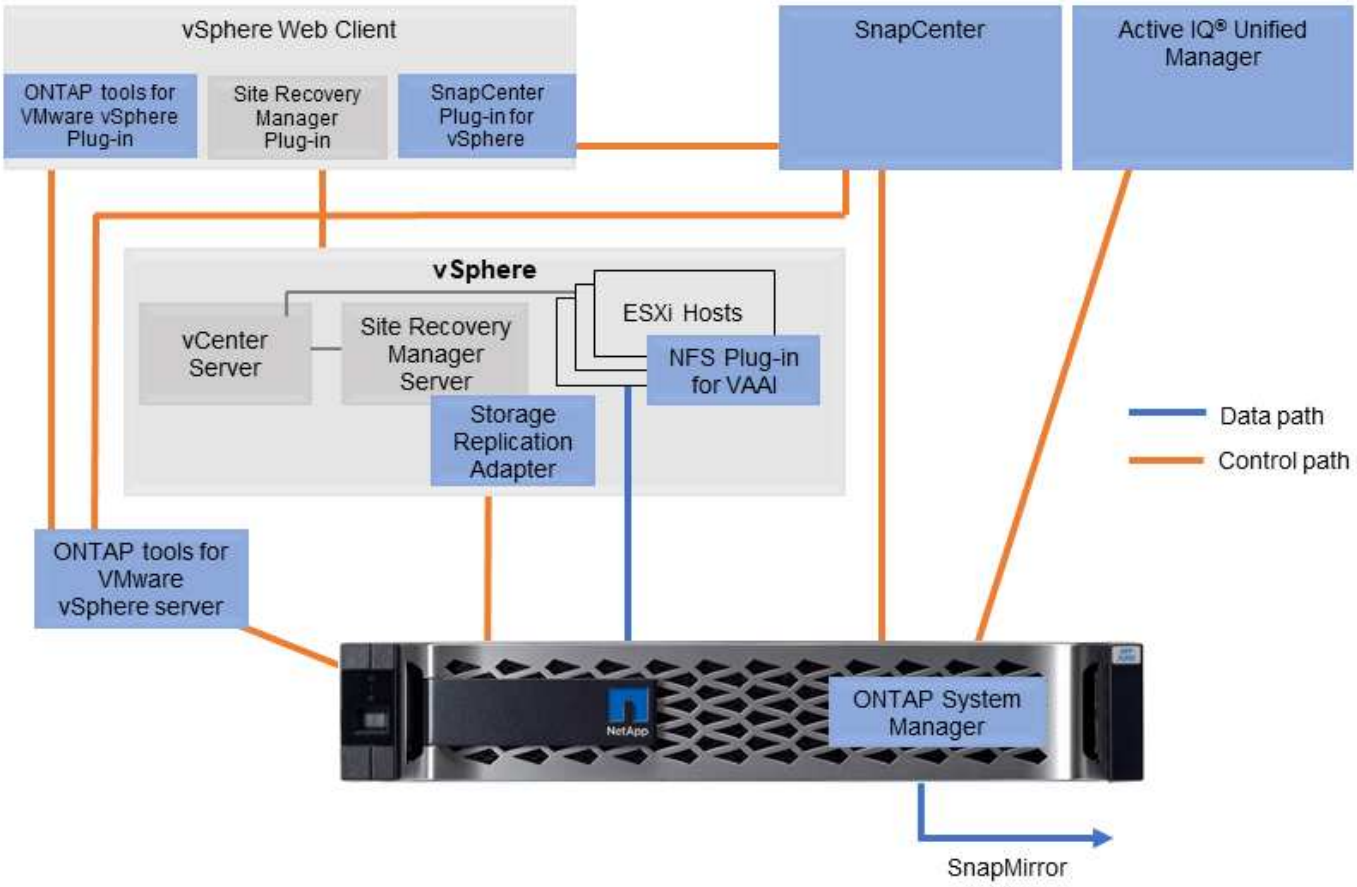
ONTAP tools for VMware vSphere is a set of tools for using ONTAP storage together with vSphere. The vCenter plug-in, formerly known as the Virtual Storage Console (VSC), simplifies storage management and efficiency features, enhances availability, and reduces storage costs and operational overhead, whether you are using SAN or NAS. It uses best practices for provisioning datastores and optimizes ESXi host settings for NFS and block storage environments. For all these benefits, NetApp recommends using these ONTAP tools as a best practice when using vSphere with systems running ONTAP software. It includes a server appliance, user interface extensions for vCenter, VASA Provider, and Storage Replication Adapter. Nearly everything in ONTAP tools can be automated by using simple REST APIs, consumable by most modern automation tools.

- **vCenter UI extensions.** The ONTAP tools UI extensions simplify the job of operations teams and vCenter

admins by embeding easy-to-use, context-sensitive menus for managing hosts and storage, informational portlets, and native alerting capabilities directly in the vCenter UI for streamlined workflows.

- **VASA Provider for ONTAP.** The VASA Provider for ONTAP supports the VMware vStorage APIs for Storage Awareness (VASA) framework. It is supplied as part of ONTAP tools for VMware vSphere as a single virtual appliance for ease of deployment. VASA Provider connects vCenter Server with ONTAP to aid in provisioning and monitoring VM storage. It enables VMware Virtual Volumes (vVols) support, management of storage capability profiles and individual VM vVols performance, and alarms for monitoring capacity and compliance with the profiles.
- **Storage Replication Adapter.** The SRA is used together with VMware Site Recovery Manager (SRM) to manage data replication between production and disaster recovery sites and test the DR replicas nondisruptively. It helps automate the tasks of discovery, recovery, and reprotection. It includes both an SRA server appliance and SRA adapters for the Windows SRM server and SRM appliance.

The following figure depicts ONTAP tools for vSphere.



NFS Plug-In for VMware VAAI

The NetApp NFS Plug-In for VMware VAAI is a plug-in for ESXi hosts that allows them to use VAAI features with NFS datastores on ONTAP. It supports copy offload for clone operations, space reservation for thick virtual disk files, and snapshot offload. Offloading copy operations to storage is not necessarily faster to complete, but it does reduce network bandwidth requirements and offloads host resources such as CPU cycles, buffers, and queues. You can use ONTAP tools for VMware vSphere to install the plug-in on ESXi hosts or, where supported, vSphere Lifecycle Manager (vLCM).

Virtual Volumes (vVols) and Storage Policy Based Management (SPBM)

NetApp was an early design partner with VMware in the development of vSphere Virtual Volumes (vVols), providing architectural input and early support for vVols and VMware vSphere APIs for Storage Awareness (VASA). Not only did this approach bring VM granular storage management to VMFS, it also supported automation of storage provisioning through Storage Policy based Management (SPBM).

SPBM provides a framework that serves as an abstraction layer between the storage services available to your virtualization environment and the provisioned storage elements via policies. This approach allows storage architects to design storage pools with different capabilities that can be easily consumed by VM administrators. Administrators can then match virtual machine workload requirements against the provisioned storage pools, allowing for granular control of various settings on a per-VM or virtual disk level.

ONTAP leads the storage industry in vVols scale, supporting hundreds of thousands of vVols in a single cluster, whereas enterprise array and smaller flash array vendors support as few as several thousand vVols per array. NetApp is also driving the evolution of VM granular management with upcoming capabilities in support of vVols 3.0.

> For more information on VMware vSphere Virtual Volumes, SPBM, and ONTAP, see TR-4400: VMware vSphere Virtual Volumes with ONTAP.

Datastores and protocols

vSphere datastore and protocol features overview

Seven protocols are used to connect VMware vSphere to datastores on a system running ONTAP software:

- FCP
- FCoE
- NVMe/FC
- NVMe/TCP
- iSCSI
- NFS v3
- NFS v4.1

FCP, FCoE, NVMe/FC, NVMe/TCP, and iSCSI are block protocols that use the vSphere Virtual Machine File System (VMFS) to store VMs inside ONTAP LUNs or NVMe namespaces that are contained in an ONTAP FlexVol volume. Note that, starting from vSphere 7.0, VMware no longer supports software FCoE in production environments. NFS is a file protocol that places VMs into datastores (which are simply ONTAP volumes) without the need for VMFS. SMB (CIFS), iSCSI, NVMe/TCP, or NFS can also be used directly from a guest OS to ONTAP.

The following tables present vSphere-supported traditional datastore features with ONTAP. This information does not apply to vVols datastores, but it does generally applies to vSphere 6.x and later releases using supported ONTAP releases. You can also consult VMware configuration maximums for specific vSphere releases to confirm specific limits.

Capability/Feature	FC/FCoE	iSCSI	NVMe-oF	NFS
Format	VMFS or raw device mapping (RDM)	VMFS or RDM	VMFS	N/A
Maximum number of datastores or LUNs	1024 LUNs per host	1024 LUNs per server	256 Namespeces per server	256 mounts Default NFS. MaxVolumes is 8. Use ONTAP tools for VMware vSphere to increase to 256.
Maximum datastore size	64TB	64TB	64TB	100TB FlexVol volume or greater with FlexGroup volume
Maximum datastore file size	62TB	62TB	62TB	62TB with ONTAP 9.12.1P2 and later
Optimal queue depth per LUN or file system	64-256	64-256	Autonegotiated	Refer to NFS.MaxQueueDepth in Recommended ESXi host and other ONTAP settings.

The following table lists supported VMware storage-related functionalities.

Capacity/Feature	FC/FCoE	iSCSI	NVMe-oF	NFS
vMotion	Yes	Yes	Yes	Yes
Storage vMotion	Yes	Yes	Yes	Yes
VMware HA	Yes	Yes	Yes	Yes
Storage Distributed Resource Scheduler (SDRS)	Yes	Yes	Yes	Yes
VMware vStorage APIs for Data Protection (VADP)–enabled backup software	Yes	Yes	Yes	Yes
Microsoft Cluster Service (MSCS) or failover clustering within a VM	Yes	Yes*	Yes*	Not supported
Fault Tolerance	Yes	Yes	Yes	Yes
Site Recovery Manager	Yes	Yes	No**	V3 only**

Capacity/Feature	FC/FCoE	iSCSI	NVMe-oF	NFS
Thin-provisioned VMs (virtual disks)	Yes	Yes	Yes	Yes This setting is the default for all VMs on NFS when not using VAAI.
VMware native multipathing	Yes	Yes	Yes, using the new High Performance Plugin (HPP)	NFS v4.1 session trunking requires ONTAP 9.14.1 and later

The following table lists supported ONTAP storage management features.

Capability/Feature	FC/FCoE	iSCSI	NVMe-oF	NFS
Data deduplication	Savings in the array	Savings in the array	Savings in the array	Savings in the datastore
Thin provisioning	Datastore or RDM	Datastore or RDM	Datastore	Datastore
Resize datastore	Grow only	Grow only	Grow only	Grow, autogrow, and shrink
SnapCenter plug-ins for Windows, Linux applications (in guest)	Yes	Yes	No	Yes
Monitoring and host configuration using ONTAP tools for VMware vSphere	Yes	Yes	No	Yes
Provisioning using ONTAP tools for VMware vSphere	Yes	Yes	No	Yes

The following table lists supported backup features.

Capability/Feature	FC/FCoE	iSCSI	NVMe-oF	NFS
ONTAP Snapshots	Yes	Yes	Yes	Yes
SRM supported by replicated backups	Yes	Yes	No**	V3 only**
Volume SnapMirror	Yes	Yes	Yes	Yes
VMDK image access	VADP-enabled backup software	VADP-enabled backup software	VADP-enabled backup software	VADP-enabled backup software, vSphere Client, and vSphere Web Client datastore browser

Capability/Feature	FC/FCoE	iSCSI	NVMe-oF	NFS
VMDK file-level access	VADP-enabled backup software, Windows only	VADP-enabled backup software, Windows only	VADP-enabled backup software, Windows only	VADP-enabled backup software and third-party applications
NDMP granularity	Datastore	Datastore	Datastore	Datastore or VM

*NetApp recommends using in-guest iSCSI for Microsoft clusters rather than multiwriter-enabled VMDKs in a VMFS datastore. This approach is fully supported by Microsoft and VMware, offers great flexibility with ONTAP (SnapMirror to ONTAP systems on-premises or in the cloud), is easy to configure and automate, and can be protected with SnapCenter. vSphere 7 adds a new clustered VMDK option. This is different from multiwriter-enabled VMDKs, which requires a datastore presented via the FC protocol that has clustered VMDK support enabled. Other restrictions apply. See VMware's Setup for Windows Server Failover Clustering documentation for configuration guidelines.

**Datastores using NVMe-oF and NFS v4.1 require vSphere replication. Array-based replication is not supported by SRM.

Selecting a storage protocol

Systems running ONTAP software support all major storage protocols, so customers can choose what is best for their environment, depending on existing and planned networking infrastructure and staff skills. NetApp testing has generally shown little difference between protocols running at similar line speeds, so it is best to focus on your network infrastructure and staff capabilities over raw protocol performance.

The following factors might be useful in considering a choice of protocol:

- **Current customer environment.** Although IT teams are generally skilled at managing Ethernet IP infrastructure, not all are skilled at managing an FC SAN fabric. However, using a general-purpose IP network that's not designed for storage traffic might not work well. Consider the networking infrastructure you have in place, any planned improvements, and the skills and availability of staff to manage them.
- **Ease of setup.** Beyond initial configuration of the FC fabric (additional switches and cabling, zoning, and the interoperability verification of HBA and firmware), block protocols also require creation and mapping of LUNs and discovery and formatting by the guest OS. After the NFS volumes are created and exported, they are mounted by the ESXi host and ready to use. NFS has no special hardware qualification or firmware to manage.
- **Ease of management.** With SAN protocols, if more space is needed, several steps are necessary, including growing a LUN, rescanning to discover the new size, and then growing the file system). Although growing a LUN is possible, reducing the size of a LUN is not, and recovering unused space can require additional effort. NFS allows easy sizing up or down, and this resizing can be automated by the storage system. SAN offers space reclamation through guest OS TRIM/UNMAP commands, allowing space from deleted files to be returned to the array. This type of space reclamation is more difficult with NFS datastores.
- **Storage space transparency.** Storage utilization is typically easier to see in NFS environments because thin provisioning returns savings immediately. Likewise, deduplication and cloning savings are immediately available for other VMs in the same datastore or for other storage system volumes. VM density is also typically greater in an NFS datastore, which can improve deduplication savings as well as reduce management costs by having fewer datastores to manage.

Datastore layout

ONTAP storage systems offer great flexibility in creating datastores for VMs and virtual disks. Although many ONTAP best practices are applied when using the VSC to provision datastores for vSphere (listed in the section Recommended ESXi host and other ONTAP settings), here are some additional guidelines to consider:

- Deploying vSphere with ONTAP NFS datastores results in a high-performing, easy-to-manage implementation that provides VM-to-datastore ratios that cannot be obtained with block-based storage protocols. This architecture can result in a tenfold increase in datastore density with a correlating reduction in the number of datastores. Although a larger datastore can benefit storage efficiency and provide operational benefits, consider using at least four datastores (FlexVol volumes) to store your VMs on a single ONTAP controller to get maximum performance from the hardware resources. This approach also allows you to establish datastores with different recovery policies. Some can be backed up or replicated more frequently than others based on business needs. Multiple datastores are not required with FlexGroup volumes for performance because they scale by design.
- NetApp recommends the use of FlexVol volumes for most NFS datastores. Starting with ONTAP 9.8 FlexGroup volumes are supported for use as datastores as well, and are generally recommended for certain use cases. Other ONTAP storage containers such as qtrees are not generally recommended because these are not currently supported by either ONTAP tools for VMware vSphere or the NetApp SnapCenter plugin for VMware vSphere. That being said, deploying datastores as multiple qtrees in a single volume might be useful for highly automated environments that can benefit from datastore-level quotas or VM file clones.
- A good size for a FlexVol volume datastore is around 4TB to 8TB. This size is a good balance point for performance, ease of management, and data protection. Start small (say, 4TB) and grow the datastore as needed (up to the maximum 100TB). Smaller datastores are faster to recover from backup or after a disaster and can be moved quickly across the cluster. Consider the use of ONTAP autosize to automatically grow and shrink the volume as used space changes. The ONTAP tools for VMware vSphere Datastore Provisioning Wizard use autosize by default for new datastores. Additional customization of the grow and shrink thresholds and maximum and minimum size can be done with System Manager or the command line.
- Alternately, VMFS datastores can be configured with LUNs that are accessed by FC, iSCSI, or FCoE. VMFS allows traditional LUNs to be accessed simultaneously by every ESX server in a cluster. VMFS datastores can be up to 64TB in size and consist of up to 32 2TB LUNs (VMFS 3) or a single 64TB LUN (VMFS 5). The ONTAP maximum LUN size is 16TB on most systems, and 128TB on All-SAN-Array systems. Therefore, a maximum size VMFS 5 datastore on most ONTAP systems can be created by using four 16TB LUNs. While there can be a performance benefit for high-I/O workloads with multiple LUNs (with high-end FAS or AFF systems), this benefit is offset by added management complexity to create, manage, and protect the datastore LUNs and increased availability risk. NetApp generally recommends using a single, large LUN for each datastore and only span if there is a special need to go beyond a 16TB datastore. As with NFS, consider using multiple datastores (volumes) to maximize performance on a single ONTAP controller.
- Older guest operating systems (OSs) needed alignment with the storage system for best performance and storage efficiency. However, modern vendor-supported OSs from Microsoft and Linux distributors such as Red Hat no longer require adjustments to align the file system partition with the blocks of the underlying storage system in a virtual environment. If you are using an old OS that might require alignment, search the NetApp Support Knowledgebase for articles using “VM alignment” or request a copy of TR-3747 from a NetApp sales or partner contact.
- Avoid the use of defragmentation utilities within the guest OS, as this offers no performance benefit and affects storage efficiency and snapshot space usage. Also consider turning off search indexing in the guest OS for virtual desktops.
- ONTAP has led the industry with innovative storage efficiency features, allowing you to get the most out of your usable disk space. AFF systems take this efficiency further with default inline deduplication and

compression. Data is deduplicated across all volumes in an aggregate, so you no longer need to group similar operating systems and similar applications within a single datastore to maximize savings.

- In some cases, you might not even need a datastore. For the best performance and manageability, avoid using a datastore for high-I/O applications such as databases and some applications. Instead, consider guest-owned file systems such as NFS or iSCSI file systems managed by the guest or with RDMs. For specific application guidance, see NetApp technical reports for your application. For example, Oracle Databases on ONTAP has a section about virtualization with helpful details.
- First Class Disks (or Improved Virtual Disks) allow for vCenter-managed disks independent of a VM with vSphere 6.5 and later. While primarily managed by API, they can be useful with vVols, especially when managed by OpenStack or Kubernetes tools. They are supported by ONTAP as well as ONTAP tools for VMware vSphere.

Datastore and VM migration

When migrating VMs from an existing datastore on another storage system to ONTAP, here are some practices to keep in mind:

- Use Storage vMotion to move the bulk of your virtual machines to ONTAP. Not only is this approach nondisruptive to running VMs, it also allows ONTAP storage efficiency features such as inline deduplication and compression to process the data as it migrates. Consider using vCenter capabilities to select multiple VMs from the inventory list and then schedule the migration (use Ctrl key while clicking Actions) at an appropriate time.
- While you could carefully plan a migration to appropriate destination datastores, it is often simpler to migrate in bulk and then organize later as needed. You might want to use this approach to guide your migration to different datastores if you have specific data protection needs, such as different Snapshot schedules.
- Most VMs and their storage may be migrated while running (hot), but migrating attached (not in datastore) storage such as ISOs, LUNs, or NFS volumes from another storage system might require cold migration.
- Virtual machines that need more careful migration include databases and applications that use attached storage. In general, consider the use of the application's tools to manage migration. For Oracle, consider using Oracle tools such as RMAN or ASM to migrate the database files. See TR-4534 for more information. Likewise, for SQL Server, consider using either SQL Server Management Studio or NetApp tools such as SnapManager for SQL Server or SnapCenter.

ONTAP tools for VMware vSphere

The most important best practice when using vSphere with systems running ONTAP software is to install and use the ONTAP tools for VMware vSphere plug-in (formerly known as Virtual Storage Console). This vCenter plug-in simplifies storage management, enhances availability, and reduces storage costs and operational overhead, whether using SAN or NAS. It uses best practices for provisioning datastores and optimizes ESXi host settings for multipath and HBA timeouts (these are described in Appendix B). Because it's a vCenter plug-in, it's available to all vSphere web clients that connect to the vCenter server.

The plug-in also helps you use other ONTAP tools in vSphere environments. It allows you to install the NFS Plug-In for VMware VAAI, which enables copy offload to ONTAP for VM cloning operations, space reservation for thick virtual disk files, and ONTAP snapshot offload.

The plug-in is also the management interface for many functions of the VASA Provider for ONTAP, supporting storage policy-based management with vVols. After ONTAP tools for VMware vSphere is registered, use it to create storage capability profiles, map them to storage, and make sure of datastore compliance with the profiles over time. The VASA Provider also provides an interface to create and manage vVol datastores.

In general, NetApp recommends using the ONTAP tools for VMware vSphere interface within vCenter to

provision traditional and vVols datastores to make sure best practices are followed.

General Networking

Configuring network settings when using vSphere with systems running ONTAP software is straightforward and similar to other network configuration. Here are some things to consider:

- Separate storage network traffic from other networks. A separate network can be achieved by using a dedicated VLAN or separate switches for storage. If the storage network shares physical paths such as uplinks, you might need QoS or additional uplink ports to make sure of sufficient bandwidth. Don't connect hosts directly to storage; use switches to have redundant paths and allow VMware HA to work without intervention. See Direct connect networking for additional information.
- Jumbo frames can be used if desired and supported by your network, especially when using iSCSI. If they are used, make sure they are configured identically on all network devices, VLANs, and so on in the path between storage and the ESXi host. Otherwise, you might see performance or connection problems. The MTU must also be set identically on the ESXi virtual switch, the VMkernel port, and also on the physical ports or interface groups of each ONTAP node.
- NetApp only recommends disabling network flow control on the cluster network ports within an ONTAP cluster. NetApp makes no other recommendations for best practices for the remaining network ports used for data traffic. You should enable or disable as necessary. See TR-4182 for more background on flow control.
- When ESXi and ONTAP storage arrays are connected to Ethernet storage networks, NetApp recommends configuring the Ethernet ports to which these systems connect as Rapid Spanning Tree Protocol (RSTP) edge ports or by using the Cisco PortFast feature. NetApp recommends enabling the Spanning-Tree PortFast trunk feature in environments that use the Cisco PortFast feature and that have 802.1Q VLAN trunking enabled to either the ESXi server or the ONTAP storage arrays.
- NetApp recommends the following best practices for link aggregation:
  - Use switches that support link aggregation of ports on two separate switch chassis using a multi-chassis link aggregation group approach such as Cisco's Virtual PortChannel (vPC).
  - Disable LACP for switch ports connected to ESXi unless you are using dvSwitches 5.1 or later with LACP configured.
  - Use LACP to create link aggregates for ONTAP storage systems with dynamic multimode interface groups with port or IP hash. Refer to Network Management for further guidance.
  - Use an IP hash teaming policy on ESXi when using static link aggregation (e.g., EtherChannel) and standard vSwitches, or LACP-based link aggregation with vSphere Distributed Switches. If link aggregation is not used, then use "Route based on the originating virtual port ID" instead.

The following table provides a summary of network configuration items and indicates where the settings are applied.

Item	ESXi	Switch	Node	SVM
IP address	VMkernel	No**	No**	Yes
Link aggregation	Virtual switch	Yes	Yes	No*
VLAN	VMkernel and VM port groups	Yes	Yes	No*
Flow control	NIC	Yes	Yes	No*
Spanning tree	No	Yes	No	No

Item	ESXi	Switch	Node	SVM
MTU (for jumbo frames)	Virtual switch and VMkernel port (9000)	Yes (set to max)	Yes (9000)	No*
Failover groups	No	No	Yes (create)	Yes (select)

*SVM LIFs connect to ports, interface groups, or VLAN interfaces that have VLAN, MTU, and other settings. However, the settings are not managed at the SVM level.

**These devices have IP addresses of their own for management, but these addresses are not used in the context of ESXi storage networking.

SAN (FC, FCoE, NVMe/FC, iSCSI), RDM

NetApp ONTAP provides enterprise-class block storage for VMware vSphere using iSCSI, Fibre Channel Protocol(FCP, or FC for short), and NVMe over Fabrics (NVMe-oF). The following are best practices for implementing block protocols for VM storage with vSphere and ONTAP.

In vSphere, there are three ways to use block storage LUNs:

- With VMFS datastores
- With raw device mapping (RDM)
- As a LUN accessed and controlled by a software initiator from a VM guest OS

VMFS is a high-performance clustered file system that provides datastores that are shared storage pools. VMFS datastores can be configured with LUNs accessed using FC, iSCSI, FCoE, or with NVMe namespaces accessed using the NVMe/FC or NVMe/TCP protocols. VMFS allows storage to be accessed simultaneously by every ESX server in a cluster. The maximum LUN size is generally 128TB beginning with ONTAP 9.12.1P2 (and earlier with ASA systems); therefore, a maximum-size VMFS 5 or 6 datastore of 64TB can be created by using a single LUN.

vSphere includes built-in support for multiple paths to storage devices, referred to as native multipathing (NMP). NMP can detect the type of storage for supported storage systems and automatically configures the NMP stack to support the capabilities of the storage system in use.

Both NMP and ONTAP support Asymmetric Logical Unit Access (ALUA) to negotiate optimized and nonoptimized paths. In ONTAP, an ALUA-optimized path follows a direct data path, using a target port on the node that hosts the LUN being accessed. ALUA is turned on by default in both vSphere and ONTAP. The NMP recognizes the ONTAP cluster as ALUA, and it uses the ALUA storage array type plug-in (`VMW_SATP_ALUA`) and selects the round robin path selection plug-in (`VMW_PSP_RR`).

ESXi 6 supports up to 256 LUNs and up to 1,024 total paths to LUNs. ESXi does not see any LUNs or paths beyond these limits. Assuming the maximum number of LUNs, the path limit allows four paths per LUN. In a larger ONTAP cluster, it is possible to reach the path limit before the LUN limit. To address this limitation, ONTAP supports selective LUN map (SLM) in release 8.3 and later.

SLM limits the nodes that advertise paths to a given LUN. It is a NetApp best practice to have at least one LIF per node per SVM and to use SLM to limit the paths advertised to the node hosting the LUN and its HA partner. Although other paths exist, they aren’t advertised by default. It is possible to modify the paths advertised with the add and remove reporting node arguments within SLM. Note that LUNs created in releases

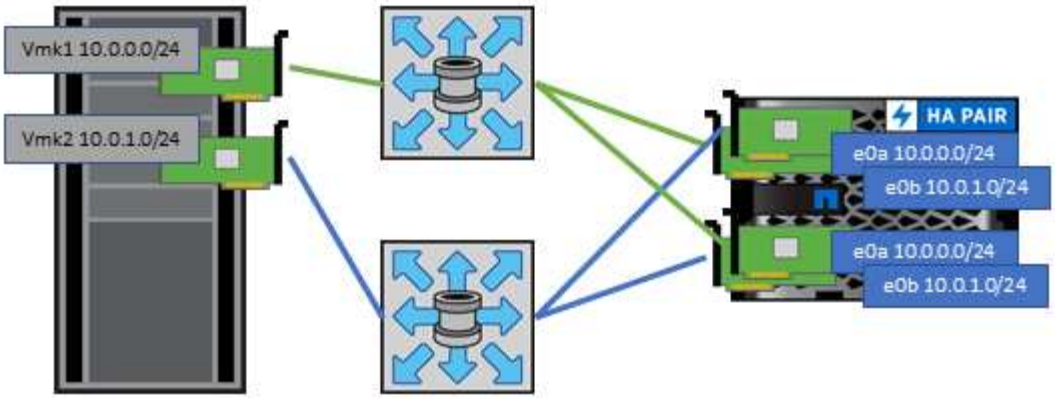
prior to 8.3 advertise all paths and need to be modified to only advertise the paths to the hosting HA pair. For more information about SLM, review section 5.9 of TR-4080. The previous method of portsets can also be used to further reduce the available paths for a LUN. Portsets help by reducing the number of visible paths through which initiators in an igroup can see LUNs.

- SLM is enabled by default. Unless you are using portsets, no additional configuration is required.
- For LUNs created prior to Data ONTAP 8.3, manually apply SLM by running the `lun mapping remove-reporting-nodes` command to remove the LUN reporting nodes and restrict LUN access to the LUN-owning node and its HA partner.

Block protocols (iSCSI, FC, and FCoE) access LUNs by using LUN IDs and serial numbers, along with unique names. FC and FCoE use worldwide names (WWNNs and WWPNs), and iSCSI uses iSCSI qualified names (IQNs). The path to LUNs inside the storage is meaningless to the block protocols and is not presented anywhere in the protocol. Therefore, a volume that contains only LUNs does not need to be internally mounted at all, and a junction path is not needed for volumes that contain LUNs used in datastores. The NVMe subsystem in ONTAP works similarly.

Other best practices to consider:

- Make sure that a logical interface (LIF) is created for each SVM on each node in the ONTAP cluster for maximum availability and mobility. ONTAP SAN best practice is to use two physical ports and LIFs per node, one for each fabric. ALUA is used to parse paths and identify active optimized (direct) paths versus active nonoptimized paths. ALUA is used for FC, FCoE, and iSCSI.
- For iSCSI networks, use multiple VMkernel network interfaces on different network subnets with NIC teaming when multiple virtual switches are present. You can also use multiple physical NICs connected to multiple physical switches to provide HA and increased throughput. The following figure provides an example of multipath connectivity. In ONTAP, configure either a single-mode interface group for failover with two or more links that are connected to two or more switches, or use LACP or other link-aggregation technology with multimode interface groups to provide HA and the benefits of link aggregation.
- If the Challenge-Handshake Authentication Protocol (CHAP) is used in ESXi for target authentication, it must also be configured in ONTAP using the CLI (`vserver iscsi security create`) or with System Manager (edit Initiator Security under Storage > SVMs > SVM Settings > Protocols > iSCSI).
- Use ONTAP tools for VMware vSphere to create and manage LUNs and igroups. The plug-in automatically determines the WWPNs of servers and creates appropriate igroups. It also configures LUNs according to best practices and maps them to the correct igroups.
- Use RDMs with care because they can be more difficult to manage, and they also use paths, which are limited as described earlier. ONTAP LUNs support both physical and virtual compatibility mode RDMs.
- For more on using NVMe/FC with vSphere 7.0, see this ONTAP NVMe/FC Host Configuration guide and TR-4684.The following figure depicts multipath connectivity from a vSphere host to an ONTAP LUN.



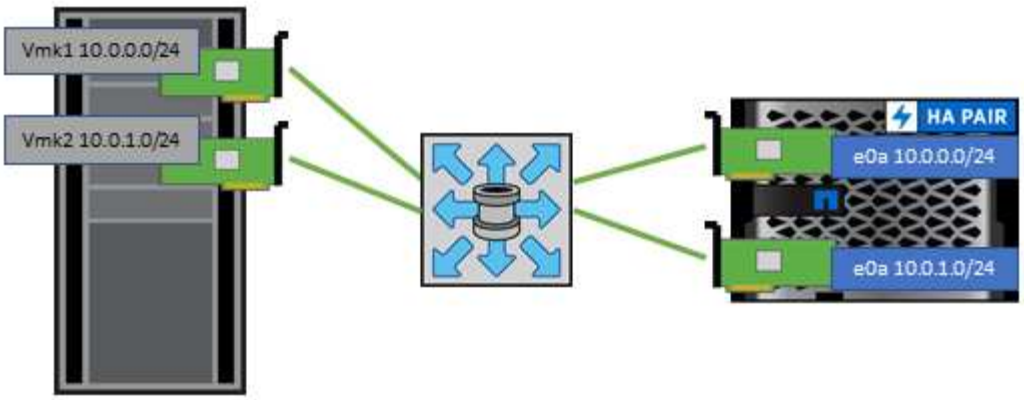
NFS

NetApp ONTAP is, among many other things, an enterprise-class scale-out NAS array. ONTAP empowers VMware vSphere with concurrent access to NFS-connected datastores from many ESXi hosts, far exceeding the limits imposed on VMFS file systems. Using NFS with vSphere provides some ease of use and storage efficiency visibility benefits, as mentioned in the datastores section.

The following best practices are recommended when using ONTAP NFS with vSphere:

- Use a single logical interface (LIF) for each SVM on each node in the ONTAP cluster. Past recommendations of a LIF per datastore are no longer necessary. While direct access (LIF and datastore on same node) is best, don't worry about indirect access because the performance effect is generally minimal (microseconds).
- VMware has supported NFSv3 since VMware Infrastructure 3. vSphere 6.0 added support for NFSv4.1, which enables some advanced capabilities such as Kerberos security. Where NFSv3 uses client-side locking, NFSv4.1 uses server-side locking. Although an ONTAP volume can be exported through both protocols, ESXi can only mount through one protocol. This single protocol mount does not preclude other ESXi hosts from mounting the same datastore through a different version. Make sure to specify the protocol version to use when mounting so that all hosts use the same version and, therefore, the same locking style. Do not mix NFS versions across hosts. If possible, use host profiles to check compliancy.
    - Because there is no automatic datastore conversion between NFSv3 and NFSv4.1, create a new NFSv4.1 datastore and use Storage vMotion to migrate VMs to the new datastore.
    - Please refer to the NFS v4.1 Interoperability table notes in the NetApp Interoperability Matrix tool for specific ESXi patch levels required for support.
    - VMware supports nconnect with NFSv3 beginning in vSphere 8.0U2. More information on nconnect can be found at the NFSv3 nConnect feature with NetApp and VMware
- NFS export policies are used to control access by vSphere hosts. You can use one policy with multiple volumes (datastores). With NFSv3, ESXi uses the sys (UNIX) security style and requires the root mount option to execute VMs. In ONTAP, this option is referred to as superuser, and when the superuser option is used, it is not necessary to specify the anonymous user ID. Note that export policy rules with different values for `-anon` and `-allow-suid` can cause SVM discovery problems with the ONTAP tools. Here's a sample policy:
    - Access Protocol: nfs (which includes both nfs3 and nfs4)
    - Client Match Spec: 192.168.42.21

    - RO Access Rule: sys
    - RW Access Rule: sys
    - Anonymous UID
    - Superuser: sys
- If the NetApp NFS Plug-In for VMware VAAI is used, the protocol should be set as `nfs` instead of `nfs3` when the export policy rule is created or modified. The VAAI copy offload feature requires the NFSv4 protocol to function, even if the data protocol is NFSv3. Specifying the protocol as `nfs` includes both the NFSv3 and NFSv4 versions.
- NFS datastore volumes are junctioned from the root volume of the SVM; therefore, ESXi must also have access to the root volume to navigate and mount datastore volumes. The export policy for the root volume, and for any other volumes in which the datastore volume's junction is nested, must include a rule or rules for the ESXi servers granting them read-only access. Here's a sample policy for the root volume, also using the VAAI plug-in:
    - Access Protocol: nfs (which includes both nfs3 and nfs4)
    - Client Match Spec: 192.168.42.21
    - RO Access Rule: sys
    - RW Access Rule: never (best security for root volume)
    - Anonymous UID
    - Superuser: sys (also required for root volume with VAAI)
- Use ONTAP tools for VMware vSphere (the most important best practice):
    - Use ONTAP tools for VMware vSphere to provision datastores because it simplifies management of export policies automatically.
    - When creating datastores for VMware clusters with the plug-in, select the cluster rather than a single ESX server. This choice triggers it to automatically mount the datastore to all hosts in the cluster.
    - Use the plug- in mount function to apply existing datastores to new servers.
    - When not using ONTAP tools for VMware vSphere, use a single export policy for all servers or for each cluster of servers where additional access control is needed.
- Although ONTAP offers a flexible volume namespace structure to arrange volumes in a tree using junctions, this approach has no value for vSphere. It creates a directory for each VM at the root of the datastore, regardless of the namespace hierarchy of the storage. Thus, the best practice is to simply mount the junction path for volumes for vSphere at the root volume of the SVM, which is how ONTAP tools for VMware vSphere provisions datastores. Not having nested junction paths also means that no volume is dependent on any volume other than the root volume and that taking a volume offline or destroying it, even intentionally, does not affect the path to other volumes.
- A block size of 4K is fine for NTFS partitions on NFS datastores. The following figure depicts connectivity from a vSphere host to an ONTAP NFS datastore.



The following table lists NFS versions and supported features.

vSphere Features	NFSv3	NFSv4.1
vMotion and Storage vMotion	Yes	Yes
High availability	Yes	Yes
Fault tolerance	Yes	Yes
DRS	Yes	Yes
Host profiles	Yes	Yes
Storage DRS	Yes	No
Storage I/O control	Yes	No
SRM	Yes	No
Virtual volumes	Yes	No
Hardware acceleration (VAAI)	Yes	Yes
Kerberos authentication	No	Yes (enhanced with vSphere 6.5 and later to support AES, krb5i)
Multipathing support	No	Yes

FlexGroup volumes

Use ONTAP and FlexGroup volumes with VMware vSphere for simple and scalable datastores that leverage the full power of an entire ONTAP cluster.

ONTAP 9.8, along with the ONTAP tools for VMware vSphere 9.8 and SnapCenter plugin for VMware 4.4 releases added support for FlexGroup volume-backed datastores in vSphere. FlexGroup volumes simplify the creation of large datastores and automatically create the necessary distributed constituent volumes across the ONTAP cluster to get the maximum performance from an ONTAP system.

Learn more about FlexGroup volumes in FlexCache and FlexGroup volume technical reports.

Use FlexGroup volumes with vSphere if you require a single, scalable vSphere datastore with the power of a full ONTAP cluster, or if you have very large cloning workloads that can benefit from the new FlexGroup cloning mechanism.

Copy offload

In addition to extensive system testing with vSphere workloads, ONTAP 9.8 added a new copy offload mechanism for FlexGroup datastores. This new system uses an improved copy engine to replicate files between constituents in the background while allowing access to both source and destination. This local cache is then used to rapidly instantiate VM clones on demand.

To enable FlexGroup optimized copy offload, refer to How to Configure ONTAP FlexGroups to allow VAAI copy offload

You may find that if you use VAAI cloning, but do not clone enough to keep the cache warm, your clones may be no faster than a host-based copy. If that is the case you may tune the cache timeout to better suit your needs.

Consider the following scenario:

- You've created a new FlexGroup with 8 constituents
- The cache timeout for the new FlexGroup is set to 160 minutes

In this scenario, the first 8 clones to complete will be full copies, not local file clones. Any additional cloning of that VM before the 160-second timeout expires will use the file clone engine inside of each constituent in a round-robin fashion to create nearly immediate copies evenly distributed across the constituent volumes.

Every new clone job a volume receives resets the timeout. If a constituent volume in the example FlexGroup does not receive a clone request before the timeout, the cache for that particular VM will be cleared and the volume will need to be populated again. Also, if the source of the original clone changes (e.g., you've updated the template) then the local cache on each constituent will be invalidated to prevent any conflict. As previously stated, the cache is tunable and can be set to match the needs of your environment.

For more information on using FlexGroups with VAAI, refer to this KB article: VAAI: How does caching work with FlexGroup volumes?

In environments where you are not able to take full advantage of the FlexGroup cache, but still require rapid cross-volume cloning, consider using vVols. Cross-volume cloning with vVols is much faster than using traditional datastores, and does not rely on a cache.

QoS settings

Configuring QoS at the FlexGroup level using ONTAP System Manager or the cluster shell is supported, however it does not provide VM awareness or vCenter integration.

QoS (max/min IOPS) can be set on individual VMs or on all VMs in a datastore at that time in the vCenter UI or via REST APIs by using ONTAP tools. Setting QoS on all VMs replaces any separate per-VM settings. Settings do not extend to new or migrated VMs in the future; either set QoS on the new VMs or re-apply QoS to all VMs in the datastore.

Note that VMware vSphere treats all IO for an NFS datastore as a single queue per host, and QoS throttling on one VM can impact performance for other VMs in the same datastore. This is in contrast with vVols which can maintain their QoS policy settings if they migrate to another datastore and do not impact IO of other VMs when throttled.

Metrics

ONTAP 9.8 also added new file-based performance metrics (IOPS, throughput, and latency) for FlexGroup files, and these metrics can be viewed in the ONTAP tools for VMware vSphere dashboard and VM reports.

The ONTAP tools for VMware vSphere plug-in also allows you to set Quality of Service (QoS) rules using a combination of maximum and/or minimum IOPS. These can be set across all VMs in a datastore or individually for specific VMs.

Best practices

- Use ONTAP tools to create FlexGroup datastores to ensure your FlexGroup is created optimally and export policies are configured to match your vSphere environment. However, after creating the FlexGroup volume with ONTAP tools, you will find that all nodes in your vSphere cluster are using a single IP address to mount the datastore. This could result in a bottleneck on the network port. To avoid this problem, unmount the datastore, and then remount it using the standard vSphere datastore wizard using a round-robin DNS name that load balancing across LIFs on the SVM. After remounting, ONTAP tools will again be able to manage the datastore. If ONTAP tools isn't available, use the FlexGroup defaults and create your export policy following the guidelines in datastores and protocols - NFS.
- When sizing a FlexGroup datastore, keep in mind that the FlexGroup consists of multiple smaller FlexVol volumes that create a larger namespace. As such, size the datastore to be at least 8x (assuming the default 8 constituents) the size of your largest VMDK file plus 10-20% unused headroom to allow for flexibility in rebalancing. For example, if you have a 6TB VMDK in your environment, size the FlexGroup datastore no smaller than 52.8TB (6x8+10%).
- VMware and NetApp support NFSv4.1 session trunking beginning with ONTAP 9.14.1. Refer to the NetApp NFS 4.1 interoperability matrix notes for specific version details. NFSv3 does not support multiple physical paths to a volume but does support nconnect beginning in vSphere 8.0U2. More information on nconnect can be found at the NFSv3 nConnect feature with NetApp and VMware.
- Use the NFS Plug-In for VMware VAAI for copy offload. Note that while cloning is enhanced within a FlexGroup datastore, as mentioned previously, ONTAP does not provide significant performance advantages versus ESXi host copy when copying VMs between FlexVol and/or FlexGroup volumes. Therefore consider your cloning workloads when deciding to use VAAI or FlexGroups. Modifying the number of constituent volumes is one way to optimize for FlexGroup-based cloning. As is tuning the cache timeout previously mentioned.
- Use ONTAP tools for VMware vSphere 9.8 or later to monitor the performance of FlexGroup VMs using ONTAP metrics (dashboard and VM reports), and to manage QoS on individual VMs. These metrics are not currently available through ONTAP commands or APIs.
- SnapCenter Plug-In for VMware vSphere release 4.4 and later supports backup and recovery of VMs in a FlexGroup datastore on the primary storage system. SCV 4.6 adds SnapMirror support for FlexGroup-based datastores. Using array-based snapshots and replication is the most efficient way to protect your data.

Network configuration

Configuring network settings when using vSphere with systems running ONTAP software is straightforward and similar to other network configuration.

Here are some things to consider:

- Separate storage network traffic from other networks. A separate network can be achieved by using a dedicated VLAN or separate switches for storage. If the storage network shares physical paths such as uplinks, you might need QoS or additional uplink ports to make sure of sufficient bandwidth. Don't connect hosts directly to storage; use switches to have redundant paths and allow VMware HA to work without intervention. See Direct connect networking for additional information.
- Jumbo frames can be used if desired and supported by your network, especially when using iSCSI. If they are used, make sure they are configured identically on all network devices, VLANs, and so on in the path

between storage and the ESXi host. Otherwise, you might see performance or connection problems. The MTU must also be set identically on the ESXi virtual switch, the VMkernel port, and also on the physical ports or interface groups of each ONTAP node.

- NetApp only recommends disabling network flow control on the cluster network ports within an ONTAP cluster. NetApp makes no other recommendations for best practices for the remaining network ports used for data traffic. You should enable or disable it as necessary. See TR-4182 for more background on flow control.
- When ESXi and ONTAP storage arrays are connected to Ethernet storage networks, NetApp recommends configuring the Ethernet ports to which these systems connect as Rapid Spanning Tree Protocol (RSTP) edge ports or by using the Cisco PortFast feature. NetApp recommends enabling the Spanning-Tree PortFast trunk feature in environments that use the Cisco PortFast feature and that have 802.1Q VLAN trunking enabled to either the ESXi server or the ONTAP storage arrays.
- NetApp recommends the following best practices for link aggregation:
  - Use switches that support link aggregation of ports on two separate switch chassis using a multi-chassis link aggregation group approach such as Cisco's Virtual PortChannel (vPC).
  - Disable LACP for switch ports connected to ESXi unless you are using dvSwitches 5.1 or later with LACP configured.
  - Use LACP to create link aggregates for ONTAP storage systems with dynamic multimode interface groups with IP hash.
  - Use an IP hash teaming policy on ESXi.

The following table provides a summary of network configuration items and indicates where the settings are applied.

Item	ESXi	Switch	Node	SVM
IP address	VMkernel	No**	No**	Yes
Link aggregation	Virtual switch	Yes	Yes	No*
VLAN	VMkernel and VM port groups	Yes	Yes	No*
Flow control	NIC	Yes	Yes	No*
Spanning tree	No	Yes	No	No
MTU (for jumbo frames)	Virtual switch and VMkernel port (9000)	Yes (set to max)	Yes (9000)	No*
Failover groups	No	No	Yes (create)	Yes (select)

*SVM LIFs connect to ports, interface groups, or VLAN interfaces that have VLAN, MTU, and other settings. However, the settings are not managed at the SVM level.

**These devices have IP addresses of their own for management, but these addresses are not used in the context of ESXi storage networking.

SAN (FC, FCoE, NVMe/FC, iSCSI), RDM

In vSphere, there are three ways to use block storage LUNs:

- With VMFS datastores

- With raw device mapping (RDM)
- As a LUN accessed and controlled by a software initiator from a VM guest OS

VMFS is a high-performance clustered file system that provides datastores that are shared storage pools. VMFS datastores can be configured with LUNs that are accessed using FC, iSCSI, FCoE, or NVMe namespaces accessed by the NVMe/FC protocol. VMFS allows traditional LUNs to be accessed simultaneously by every ESX server in a cluster. The ONTAP maximum LUN size is generally 16TB; therefore, a maximum-size VMFS 5 datastore of 64TB (see the first table in this section) is created by using four 16TB LUNs (All SAN Array systems support the maximum VMFS LUN size of 64TB). Because the ONTAP LUN architecture does not have small individual queue depths, VMFS datastores in ONTAP can scale to a greater degree than with traditional array architectures in a relatively simple manner.

vSphere includes built-in support for multiple paths to storage devices, referred to as native multipathing (NMP). NMP can detect the type of storage for supported storage systems and automatically configures the NMP stack to support the capabilities of the storage system in use.

Both NMP and ONTAP support Asymmetric Logical Unit Access (ALUA) to negotiate optimized and nonoptimized paths. In ONTAP, an ALUA-optimized path follows a direct data path, using a target port on the node that hosts the LUN being accessed. ALUA is turned on by default in both vSphere and ONTAP. The NMP recognizes the ONTAP cluster as ALUA, and it uses the ALUA storage array type plug-in (`VMW_SATP_ALUA`) and selects the round-robin path selection plug-in (`VMW_PSP_RR`).

ESXi 6 supports up to 256 LUNs and up to 1,024 total paths to LUNs. Any LUNs or paths beyond these limits are not seen by ESXi. Assuming the maximum number of LUNs, the path limit allows four paths per LUN. In a larger ONTAP cluster, it is possible to reach the path limit before the LUN limit. To address this limitation, ONTAP supports selective LUN map (SLM) in release 8.3 and later.

SLM limits the nodes that advertise paths to a given LUN. It is a NetApp best practice to have at least one LIF per node per SVM and to use SLM to limit the paths advertised to the node hosting the LUN and its HA partner. Although other paths exist, they aren't advertised by default. It is possible to modify the paths advertised with the add and remove reporting node arguments within SLM. Note that LUNs created in releases before 8.3 advertise all paths and need to be modified to only advertise the paths to the hosting HA pair. For more information about SLM, review section 5.9 of TR-4080. The previous method of portsets can also be used to further reduce the available paths for a LUN. Portsets help by reducing the number of visible paths through which initiators in an igroup can see LUNs.

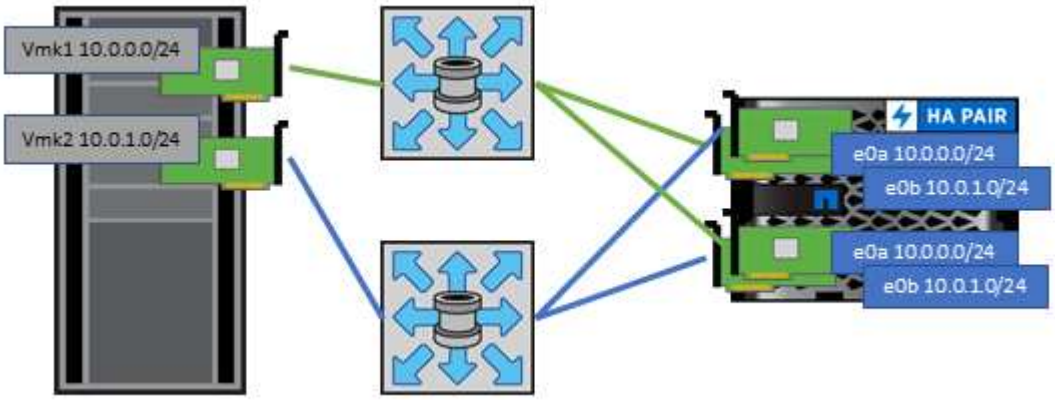
- SLM is enabled by default. Unless you are using portsets, no additional configuration is required.
- For LUNs created before Data ONTAP 8.3, manually apply SLM by running the `lun mapping remove-reporting-nodes` command to remove the LUN reporting nodes and restrict LUN access to the LUN-owning node and its HA partner.

Block protocols (iSCSI, FC, and FCoE) access LUNs by using LUN IDs and serial numbers, along with unique names. FC and FCoE use worldwide names (WWNNs and WWPNs), and iSCSI uses iSCSI qualified names (IQNs). The path to LUNs inside the storage is meaningless to the block protocols and is not presented anywhere in the protocol. Therefore, a volume that contains only LUNs does not need to be internally mounted at all, and a junction path is not needed for volumes that contain LUNs used in datastores. The NVMe subsystem in ONTAP works similarly.

Other best practices to consider:

- Make sure that a logical interface (LIF) is created for each SVM on each node in the ONTAP cluster for maximum availability and mobility. ONTAP SAN best practice is to use two physical ports and LIFs per node, one for each fabric. ALUA is used to parse paths and identify active optimized (direct) paths versus active nonoptimized paths. ALUA is used for FC, FCoE, and iSCSI.

- For iSCSI networks, use multiple VMkernel network interfaces on different network subnets with NIC teaming when multiple virtual switches are present. You can also use multiple physical NICs connected to multiple physical switches to provide HA and increased throughput. The following figure provides an example of multipath connectivity. In ONTAP, use a single-mode interface group with multiple links to different switches or LACP with multimode interface groups for high availability and link aggregation benefits.
- If the Challenge-Handshake Authentication Protocol (CHAP) is used in ESXi for target authentication, it must also be configured in ONTAP using the CLI (`vserver iscsi security create`) or with System Manager (edit Initiator Security under Storage > SVMs > SVM Settings > Protocols > iSCSI).
- Use ONTAP tools for VMware vSphere to create and manage LUNs and igroups. The plug-in automatically determines the WWPNs of servers and creates appropriate igroups. It also configures LUNs according to best practices and maps them to the correct igroups.
- Use RDMs with care because they can be more difficult to manage, and they also use paths, which are limited as described earlier. ONTAP LUNs support both physical and virtual compatibility mode RDMs.
- For more on using NVMe/FC with vSphere 7.0, see this ONTAP NVMe/FC Host Configuration guide and TR-4684. The following figure depicts multipath connectivity from a vSphere host to an ONTAP LUN.



NFS

vSphere allows customers to use enterprise-class NFS arrays to provide concurrent access to datastores to all the nodes in an ESXi cluster. As mentioned in the datastore section, there are some ease of use and storage efficiency visibility benefits when using NFS with vSphere.

The following best practices are recommended when using ONTAP NFS with vSphere:

- Use a single logical interface (LIF) for each SVM on each node in the ONTAP cluster. Past recommendations of a LIF per datastore are no longer necessary. While direct access (LIF and datastore on the same node) is best, don't worry about indirect access because the performance effect is generally minimal (microseconds).
- All versions of VMware vSphere that are currently supported can use both NFS v3 and v4.1. Official support for nconnect was added to vSphere 8.0 update 2 for NFS v3. For NFS v4.1, vSphere continues to support session trunking, Kerberos authentication, and Kerberos authentication with integrity. It's important to note that session trunking requires ONTAP 9.14.1 or a later version. You can learn more about the nconnect feature and how it improves performance at NFSv3 nConnect feature with NetApp and VMware.

It's worth noting that NFSv3 and NFSv4.1 use different locking mechanisms. NFSv3 uses client-side locking, while NFSv4.1 uses server-side locking. Although an ONTAP volume can be exported through both protocols, ESXi can only mount a datastore through one protocol. However, this doesn't mean that other ESXi hosts

cannot mount the same datastore through a different version. To avoid any issues, it's essential to specify the protocol version to use when mounting, ensuring that all hosts use the same version and, therefore, the same locking style. It's critical to avoid mixing NFS versions across hosts. If possible, use host profiles to check compliance.

Because there is no automatic datastore conversion between NFSv3 and NFSv4.1, create a new NFSv4.1 datastore and use Storage vMotion to migrate VMs to the new datastore.

Please refer to the NFS v4.1 Interoperability table notes in the NetApp Interoperability Matrix tool for specific ESXi patch levels required for support.

* NFS export policies are used to control access by vSphere hosts. You can use one policy with multiple volumes (datastores). With NFSv3, ESXi uses the sys (UNIX) security style and requires the root mount option to execute VMs. In ONTAP, this option is referred to as superuser, and when the superuser option is used, it is not necessary to specify the anonymous user ID. Note that export policy rules with different values for `-anon` and `-allow-suid` can cause SVM discovery problems with the ONTAP tools. Here's a sample policy:

Access Protocol: nfs3

Client Match Spec: 192.168.42.21

RO Access Rule: sys

RW Access Rule: sys

Anonymous UID

Superuser: sys

* If the NetApp NFS Plug-In for VMware VAAI is used, the protocol should be set as `nfs` when the export policy rule is created or modified. The NFSv4 protocol is required for VAAI copy offload to work, and specifying the protocol as `nfs` automatically includes both the NFSv3 and the NFSv4 versions.

* NFS datastore volumes are junctioned from the root volume of the SVM; therefore, ESXi must also have access to the root volume to navigate and mount datastore volumes. The export policy for the root volume, and for any other volumes in which the datastore volume's junction is nested, must include a rule or rules for the ESXi servers granting them read-only access. Here's a sample policy for the root volume, also using the VAAI plug-in:

Access Protocol: nfs (which includes both nfs3 and nfs4)

Client Match Spec: 192.168.42.21

RO Access Rule: sys

RW Access Rule: never (best security for root volume)

Anonymous UID

Superuser: sys (also required for root volume with VAAI)

* Use ONTAP tools for VMware vSphere (the most important best practice):

Use ONTAP tools for VMware vSphere to provision datastores because it simplifies the management of export policies automatically.

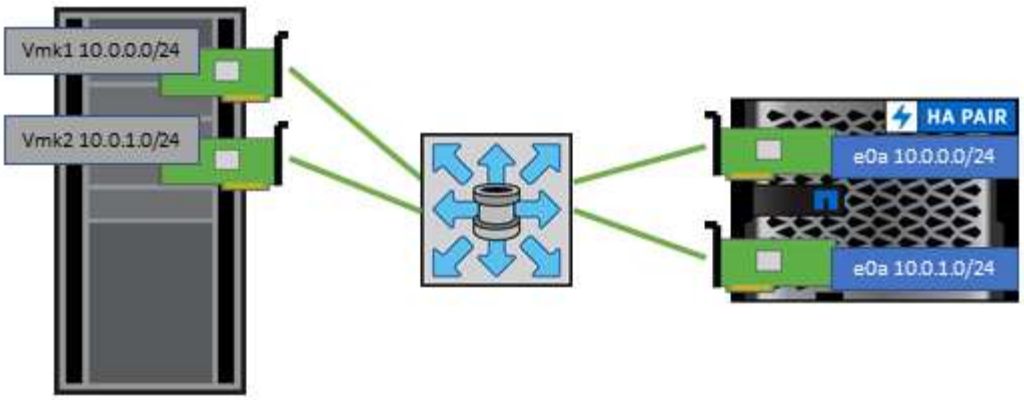
When creating datastores for VMware clusters with the plug-in, select the cluster rather than a single ESX server. This choice triggers it to automatically mount the datastore to all hosts in the cluster.

Use the plug-in mount function to apply existing datastores to new servers.

When not using ONTAP tools for VMware vSphere, use a single export policy for all servers or for each cluster of servers where additional access control is needed.

* Although ONTAP offers a flexible volume namespace structure to arrange volumes in a tree using junctions, this approach has no value for vSphere. It creates a directory for each VM at the root of the datastore, regardless of the namespace hierarchy of the storage. Thus, the best practice is to simply mount the junction path for volumes for vSphere at the root volume of the SVM, which is how ONTAP tools for VMware vSphere provisions datastores. Not having nested junction paths also means that no volume is dependent on any volume other than the root volume and that taking a volume offline or destroying it, even intentionally, does not affect the path to other volumes.

* A block size of 4K is fine for NTFS partitions on NFS datastores. The following figure depicts connectivity from a vSphere host to an ONTAP NFS datastore.



The following table lists NFS versions and supported features.

vSphere Features	NFSv3	NFSv4.1
vMotion and Storage vMotion	Yes	Yes
High availability	Yes	Yes
Fault tolerance	Yes	Yes
DRS	Yes	Yes
Host profiles	Yes	Yes
Storage DRS	Yes	No
Storage I/O control	Yes	No
SRM	Yes	No
Virtual volumes	Yes	No
Hardware acceleration (VAAI)	Yes	Yes
Kerberos authentication	No	Yes (enhanced with vSphere 6.5 and later to support AES, krb5i)
Multipathing support	No	Yes (ONTAP 9.14.1)

Direct connect networking

Storage administrators sometimes prefer to simplify their infrastructures by removing network switches from the configuration. This can be supported in some scenarios.

iSCSI and NVMe/TCP

A host using iSCSI or NVMe/TCP can be directly connected to a storage system and operate normally. The reason is pathing. Direct connections to two different storage controllers result in two independent paths for data flow. The loss of path, port, or controller does not prevent the other path from being used.

NFS

Direct-connected NFS storage can be used, but with a significant limitation - failover will not work without a significant scripting effort, which would be the responsibility of the customer.

The reason nondisruptive failover is complicated with direct-connected NFS storage is the routing that occurs

on the local OS. For example, assume a host has an IP address of 192.168.1.1/24 and is directly connected to an ONTAP controller with an IP address of 192.168.1.50/24. During failover, that 192.168.1.50 address can fail over to the other controller, and it will be available to the host, but how does the host detect its presence? The original 192.168.1.1 address still exists on the host NIC that no longer connects to an operational system. Traffic destined for 192.168.1.50 would continue to be sent to an inoperable network port.

The second OS NIC could be configured as 19 2.168.1.2 and would be capable of communicating with the failed over 192.168.1.50 address, but the local routing tables would have a default of using one **and only one** address to communicate with the 192.168.1.0/24 subnet. A sysadmin could create a scripting framework that would detect a failed network connection and alter the local routing tables or bring interfaces up and down. The exact procedure would depend on the OS in use.

In practice, NetApp customers do have direct-connected NFS, but normally only for workloads where IO pauses during failovers are acceptable. When hard mounts are used, there should not be any IO errors during such pauses. The IO should hang until services are restored, either by a failback or manual intervention to move IP addresses between NICs on the host.

FC Direct Connect

It is not possible to directly connect a host to an ONTAP storage system using the FC protocol. The reason is the use of NPIV. The WWN that identifies an ONTAP FC port to the FC network uses a type of virtualization called NPIV. Any device connected to an ONTAP system must be able to recognize an NPIV WWN. There are no current HBA vendors who offer an HBA that can be installed in a host that would be able to support an NPIV target.

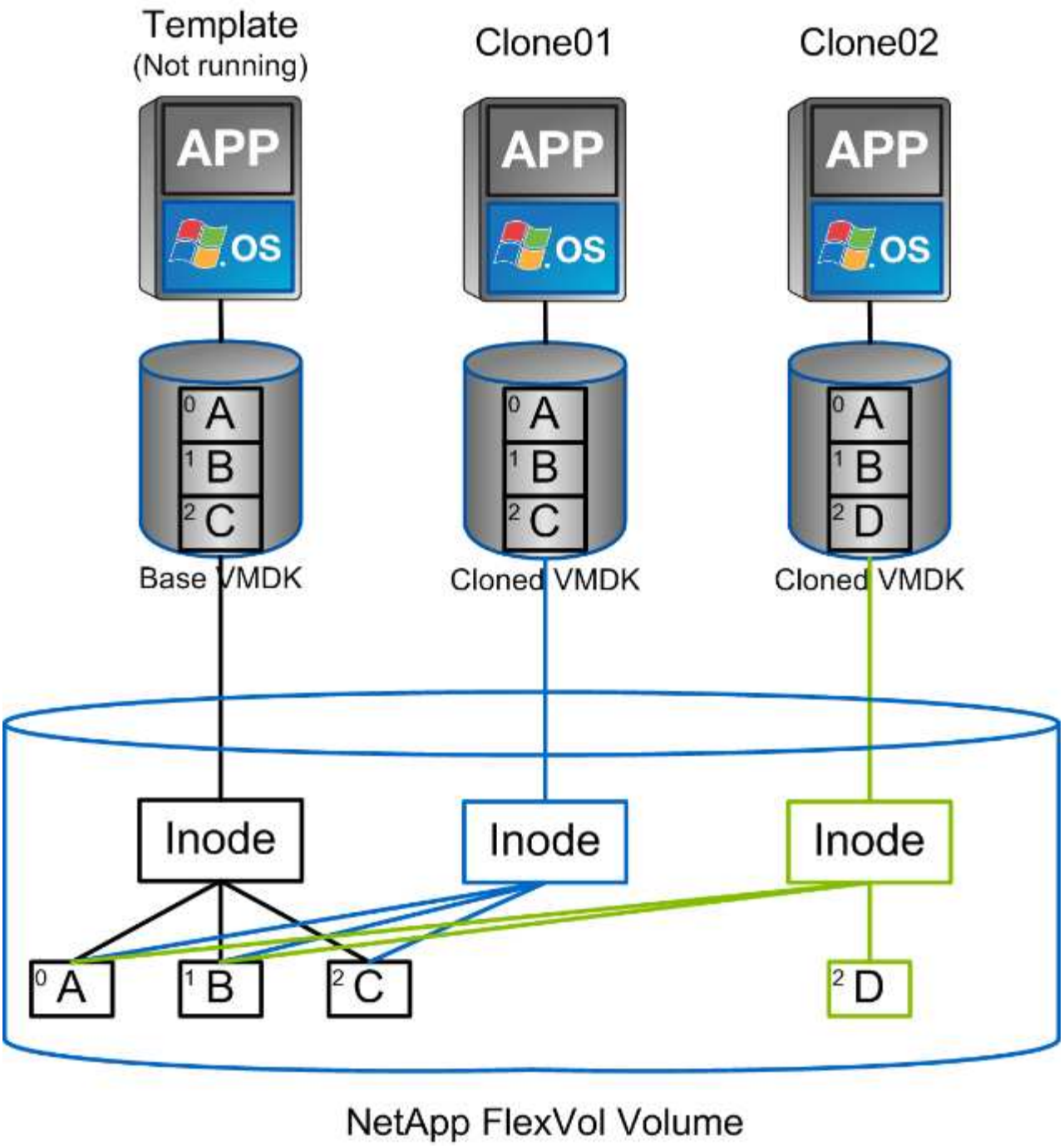
VM and datastore cloning

Cloning a storage object allows you to quickly create copies for further use, such as provisioning additional VMs, backup/recovery operations, and so on.

In vSphere, you can clone a VM, virtual disk, vVol, or datastore. After being cloned, the object can be further customized, often through an automated process. vSphere supports both full copy clones, as well as linked clones, where it tracks changes separately from the original object.

Linked clones are great for saving space, but they increase the amount of I/O that vSphere handles for the VM, affecting performance of that VM and perhaps the host overall. That’s why NetApp customers often use storage system-based clones to get the best of both worlds: efficient use of storage and increased performance.

The following figure depicts ONTAP cloning.



Cloning can be offloaded to systems running ONTAP software through several mechanisms, typically at the VM, vVol, or datastore level. These include the following:

- vVols using the NetApp vSphere APIs for Storage Awareness (VASA) Provider. ONTAP clones are used to support vVol snapshots managed by vCenter that are space-efficient with minimal I/O effect to create and delete them. VMs can also be cloned using vCenter, and these are also offloaded to ONTAP, whether within a single datastore/volume or between datastores/volumes.
- vSphere cloning and migration using vSphere APIs – Array Integration (VAAI). VM cloning operations can be offloaded to ONTAP in both SAN and NAS environments (NetApp supplies an ESXi plug-in to enable VAAI for NFS). vSphere only offloads operations on cold (powered off) VMs in a NAS datastore, whereas operations on hot VMs (cloning and storage vMotion) are also offloaded for SAN. ONTAP uses the most efficient approach based on source, destination, and installed product licenses. This capability is also used by VMware Horizon View.

- SRA (used with VMware Site Recovery Manager). Here, clones are used to test recovery of the DR replica nondisruptively.
- Backup and recovery using NetApp tools such as SnapCenter. VM clones are used to verify backup operations as well as to mount a VM backup so that individual files can be copied.

ONTAP offloaded cloning can be invoked by VMware, NetApp, and third-party tools. Clones that are offloaded to ONTAP have several advantages. They are space-efficient in most cases, needing storage only for changes to the object; there is no additional performance effect to read and write them, and in some cases performance is improved by sharing blocks in high-speed caches. They also offload CPU cycles and network I/O from the ESXi server. Copy offload within a traditional datastore using a FlexVol volume can be fast and efficient with FlexClone licensed, but copies between FlexVol volumes might be slower. If you maintain VM templates as a source of clones, consider placing them within the datastore volume (use folders or content libraries to organize them) for fast, space efficient clones.

You can also clone a volume or LUN directly within ONTAP to clone a datastore. With NFS datastores, FlexClone technology can clone an entire volume, and the clone can be exported from ONTAP and mounted by ESXi as another datastore. For VMFS datastores, ONTAP can clone a LUN within a volume or a whole volume, including one or more LUNs within it. A LUN containing a VMFS must be mapped to an ESXi initiator group (igroup) and then resignatured by ESXi to be mounted and used as a regular datastore. For some temporary use cases, a cloned VMFS can be mounted without resignaturing. After a datastore is cloned, VMs inside it can be registered, reconfigured, and customized as if they were individually cloned VMs.

In some cases, additional licensed features can be used to enhance cloning, such as SnapRestore for backup or FlexClone. These licenses are often included in license bundles at no additional cost. A FlexClone license is required for vVol cloning operations as well as to support managed snapshots of a vVol (which are offloaded from the hypervisor to ONTAP). A FlexClone license can also improve certain VAAI-based clones when used within a datastore/volume (creates instant, space-efficient copies instead of block copies). It is also used by the SRA when testing recovery of a DR replica, and SnapCenter for clone operations and to browse backup copies to restore individual files.

Data protection

Backing up your VMs and quickly recovering them are among the great strengths of ONTAP for vSphere, and it is easy to manage this ability inside vCenter with the SnapCenter Plug-In for VMware vSphere.

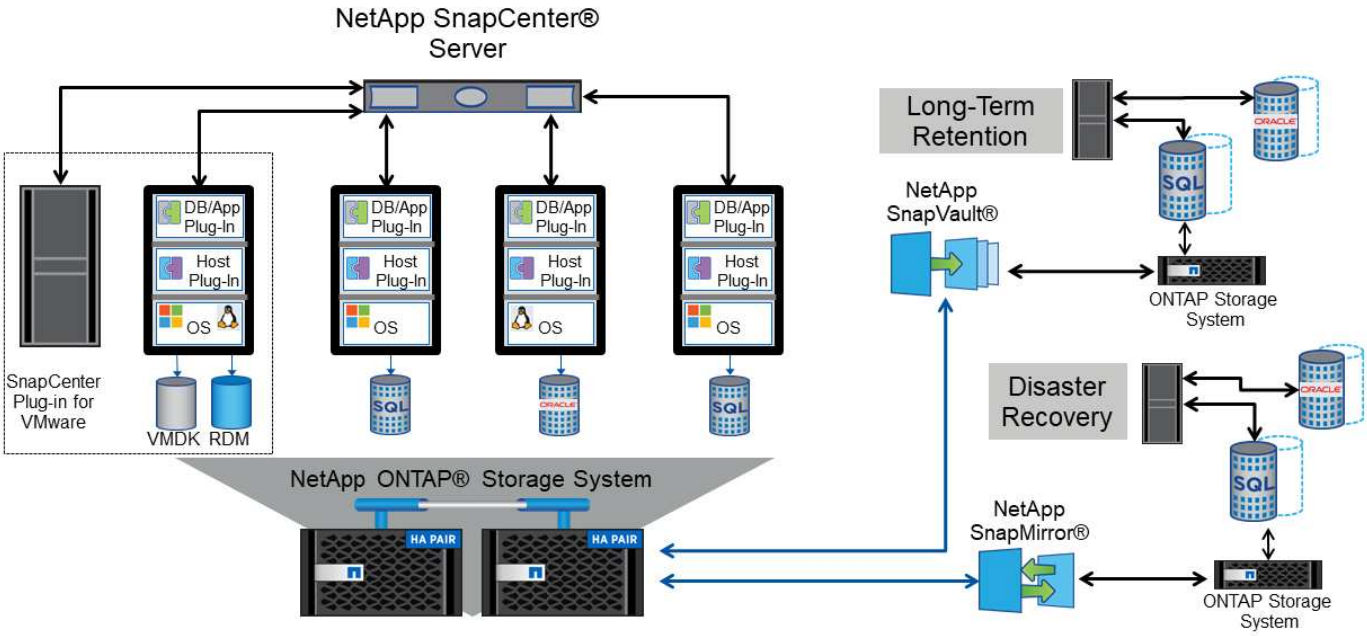
Use snapshots to make quick copies of your VM or datastore without affecting performance, and then send them to a secondary system using SnapMirror for longer-term off-site data protection. This approach minimizes storage space and network bandwidth by only storing changed information.

SnapCenter allows you to create backup policies that can be applied to multiple jobs. These policies can define schedule, retention, replication, and other capabilities. They continue to allow optional selection of VM-consistent snapshots, which leverages the hypervisor's ability to quiesce I/O before taking a VMware snapshot. However, due to the performance effect of VMware snapshots, they are generally not recommended unless you need the guest file system to be quiesced. Instead, use snapshots for general protection, and use application tools such as SnapCenter plug-ins to protect transactional data such as SQL Server or Oracle. These snapshots are different from VMware (consistency) snapshots and are suitable for longer term protection. VMware snapshots are only recommended for short term use due to performance and other effects.

These plug-ins offer extended capabilities to protect the databases in both physical and virtual environments. With vSphere, you can use them to protect SQL Server or Oracle databases where data is stored on RDM LUNs, iSCSI LUNs directly connected to the guest OS, or VMDK files on either VMFS or NFS datastores. The plug-ins allow specification of different types of database backups, supporting online or offline backup, and

protecting database files along with log files. In addition to backup and recovery, the plug-ins also support cloning of databases for development or test purposes.

The following figure depicts an example of SnapCenter deployment.



For enhanced disaster recovery capabilities, consider using the NetApp SRA for ONTAP with VMware Site Recovery Manager. In addition to support for the replication of datastores to a DR site, it also enables nondisruptive testing in the DR environment by cloning the replicated datastores. Recovery from a disaster and reprotecting production after the outage has been resolved are also made easy by automation built into SRA.

Finally, for the highest level of data protection, consider a VMware vSphere Metro Storage Cluster (vMSC) configuration using NetApp MetroCluster. vMSC is a VMware-certified solution that combines synchronous replication with array-based clustering, giving the same benefits of a high-availability cluster but distributed across separate sites to protect against site disaster. NetApp MetroCluster offers cost-effective configurations for synchronous replication with transparent recovery from any single storage component failure as well as single-command recovery in the event of a site disaster. vMSC is described in greater detail in TR-4128.

Quality of service (QoS)

Systems running ONTAP software can use the ONTAP storage QoS feature to limit throughput in MBps and/or I/Os per second (IOPS) for different storage objects such as files, LUNs, volumes, or entire SVMs.

Throughput limits are useful in controlling unknown or test workloads before deployment to make sure they don’t affect other workloads. They can also be used to constrain a bully workload after it is identified. Minimum levels of service based on IOPS are also supported to provide consistent performance for SAN objects in ONTAP 9.2 and for NAS objects in ONTAP 9.3.

With an NFS datastore, a QoS policy can be applied to the entire FlexVol volume or individual VMDK files within it. With VMFS datastores using ONTAP LUNs, the QoS policies can be applied to the FlexVol volume that contains the LUNs or individual LUNs, but not individual VMDK files because ONTAP has no awareness of the VMFS file system. When using vVols, minimum and/or maximum QoS can be set on individual VMs using the storage capability profile and VM storage policy.

The QoS maximum throughput limit on an object can be set in MBps and/or IOPS. If both are used, the first limit reached is enforced by ONTAP. A workload can contain multiple objects, and a QoS policy can be applied to one or more workloads. When a policy is applied to multiple workloads, the workloads share the total limit of the policy. Nested objects are not supported (for example, files within a volume cannot each have their own policy). QoS minimums can only be set in IOPS.

The following tools are currently available for managing ONTAP QoS policies and applying them to objects:

- ONTAP CLI
- ONTAP System Manager
- OnCommand Workflow Automation
- Active IQ Unified Manager
- NetApp PowerShell Toolkit for ONTAP
- ONTAP tools for VMware vSphere VASA Provider

To assign a QoS policy to a VMDK on NFS, note the following guidelines:

- The policy must be applied to the `vmname- flat.vmdk` that contains the actual virtual disk image, not the `vmname.vmdk` (virtual disk descriptor file) or `vmname.vmx` (VM descriptor file).
- Do not apply policies to other VM files such as virtual swap files (`vmname.vswp`).
- When using the vSphere web client to find file paths (Datastore > Files), be aware that it combines the information of the `- flat.vmdk` and `. vmdk` and simply shows one file with the name of the `. vmdk` but the size of the `- flat.vmdk`. Add `-flat` into the file name to get the correct path.

To assign a QoS policy to a LUN, including VMFS and RDM, the ONTAP SVM (displayed as Vserver), LUN path, and serial number can be obtained from the Storage Systems menu on the ONTAP tools for VMware vSphere home page. Select the storage system (SVM), and then Related Objects > SAN. Use this approach when specifying QoS using one of the ONTAP tools.

Maximum and minimum QoS can be easily assigned to a vVol-based VM with ONTAP tools for VMware vSphere or Virtual Storage Console 7.1 and later. When creating the storage capability profile for the vVol container, specify a max and/or min IOPS value under the performance capability and then reference this SCP with the VM's storage policy. Use this policy when creating the VM or apply the policy to an existing VM.

FlexGroup datastores offer enhanced QoS capabilities when using ONTAP tools for VMware vSphere 9.8 and later. You can easily set QoS on all VMs in a datastore or on specific VMs. See the FlexGroup section of this report for more information.

ONTAP QoS and VMware SIOC

ONTAP QoS and VMware vSphere Storage I/O Control (SIOC) are complementary technologies that vSphere and storage administrators can use together to manage performance of vSphere VMs hosted on systems running ONTAP software. Each tool has its own strengths, as shown in the following table. Because of the different scopes of VMware vCenter and ONTAP, some objects can be seen and managed by one system and not the other.

Property	ONTAP QoS	VMware SIOC
When active	Policy is always active	Active when contention exists (datastore latency over threshold)

Property	ONTAP QoS	VMware SIOC
Type of units	IOPS, MBps	IOPS, shares
vCenter or application scope	Multiple vCenter environments, other hypervisors and applications	Single vCenter server
Set QoS on VM?	VMDK on NFS only	VMDK on NFS or VMFS
Set QoS on LUN (RDM)?	Yes	No
Set QoS on LUN (VMFS)?	Yes	No
Set QoS on volume (NFS datastore)?	Yes	No
Set QoS on SVM (tenant)?	Yes	No
Policy based approach?	Yes; can be shared by all workloads in the policy or applied in full to each workload in the policy.	Yes, with vSphere 6.5 and later.
License required	Included with ONTAP	Enterprise Plus

VMware Storage Distributed Resource Scheduler

VMware Storage Distributed Resource Scheduler (SDRS) is a vSphere feature that places VMs on storage based on the current I/O latency and space usage. It then moves the VM or VMDKs nondisruptively between the datastores in a datastore cluster (also referred to as a pod), selecting the best datastore in which to place the VM or VMDKs in the datastore cluster. A datastore cluster is a collection of similar datastores that are aggregated into a single unit of consumption from the vSphere administrator's perspective.

When using SDRS with ONTAP tools for VMware vSphere, you must first create a datastore with the plug-in, use vCenter to create the datastore cluster, and then add the datastore to it. After the datastore cluster is created, additional datastores can be added to the datastore cluster directly from the provisioning wizard on the Details page.

Other ONTAP best practices for SDRS include the following:

- All datastores in the cluster should use the same type of storage (such as SAS, SATA, or SSD), be either all VMFS or NFS datastores, and have the same replication and protection settings.
- Consider using SDRS in default (manual) mode. This approach allows you to review the recommendations and decide whether to apply them or not. Be aware of these effects of VMDK migrations:
  - When SDRS moves VMDKs between datastores, any space savings from ONTAP cloning or deduplication are lost. You can rerun deduplication to regain these savings.
  - After SDRS moves VMDKs, NetApp recommends recreating the snapshots at the source datastore because space is otherwise locked by the VM that was moved.
  - Moving VMDKs between datastores on the same aggregate has little benefit, and SDRS does not have visibility into other workloads that might share the aggregate.

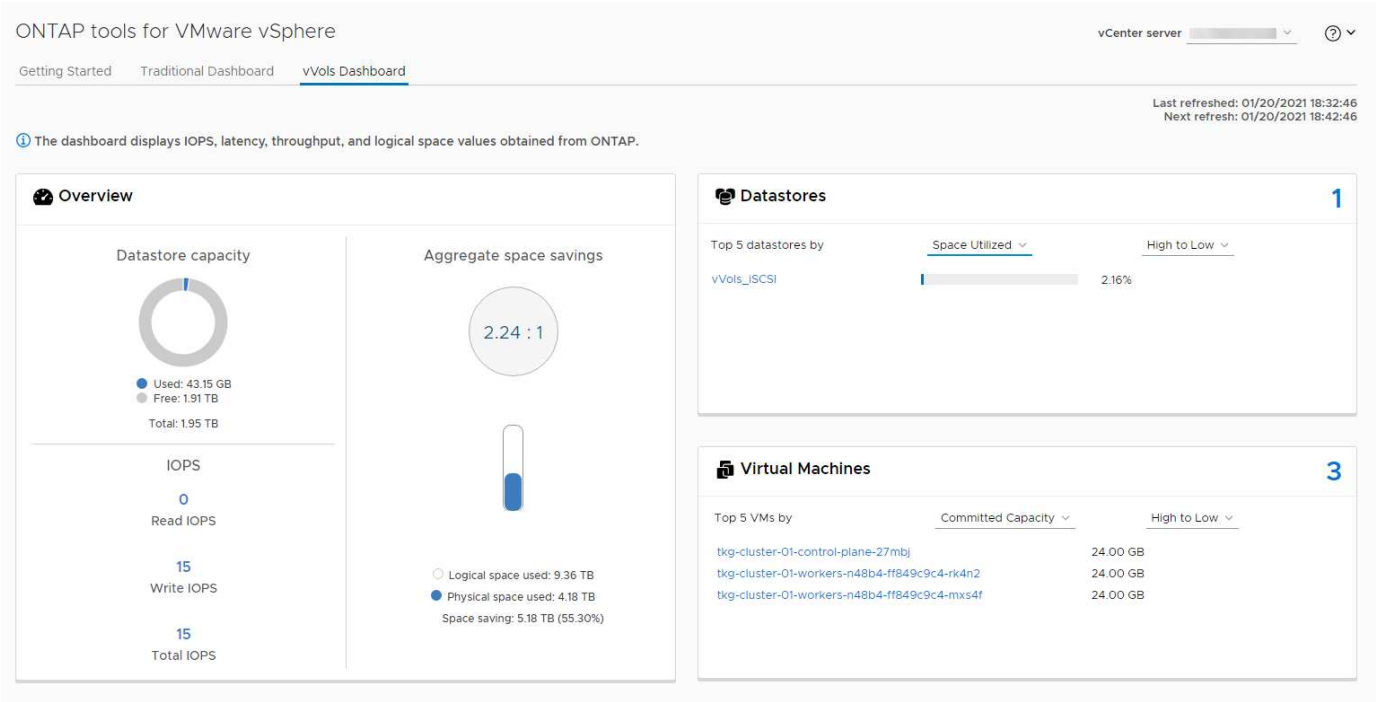
Storage policy based management and vVols

VMware vSphere APIs for Storage Awareness (VASA) make it easy for a storage administrator to configure datastores with well-defined capabilities and let the VM administrator use those whenever needed to provision VMs without having to interact with each other. It's worth taking a look at this approach to see how it can streamline your virtualization storage operations and avoid a lot of trivial work.

Prior to VASA, VM administrators could define VM storage policies, but they had to work with the storage administrator to identify appropriate datastores, often by using documentation or naming conventions. With VASA, the storage administrator can define a range of storage capabilities, including performance, tiering, encryption, and replication. A set of capabilities for a volume or a set of volumes is called a storage capability profile (SCP).

The SCP supports minimum and/or maximum QoS for a VM's data vVols. Minimum QoS is supported only on AFF systems. ONTAP tools for VMware vSphere includes a dashboard that displays VM granular performance and logical capacity for vVols on ONTAP systems.

The following figure depicts ONTAP tools for VMware vSphere 9.8 vVols dashboard.



After the storage capability profile is defined, it can be used to provision VMs using the storage policy that identifies its requirements. The mapping between the VM storage policy and the datastore storage capability profile allows vCenter to display a list of compatible datastores for selection. This approach is known as storage policy based management.

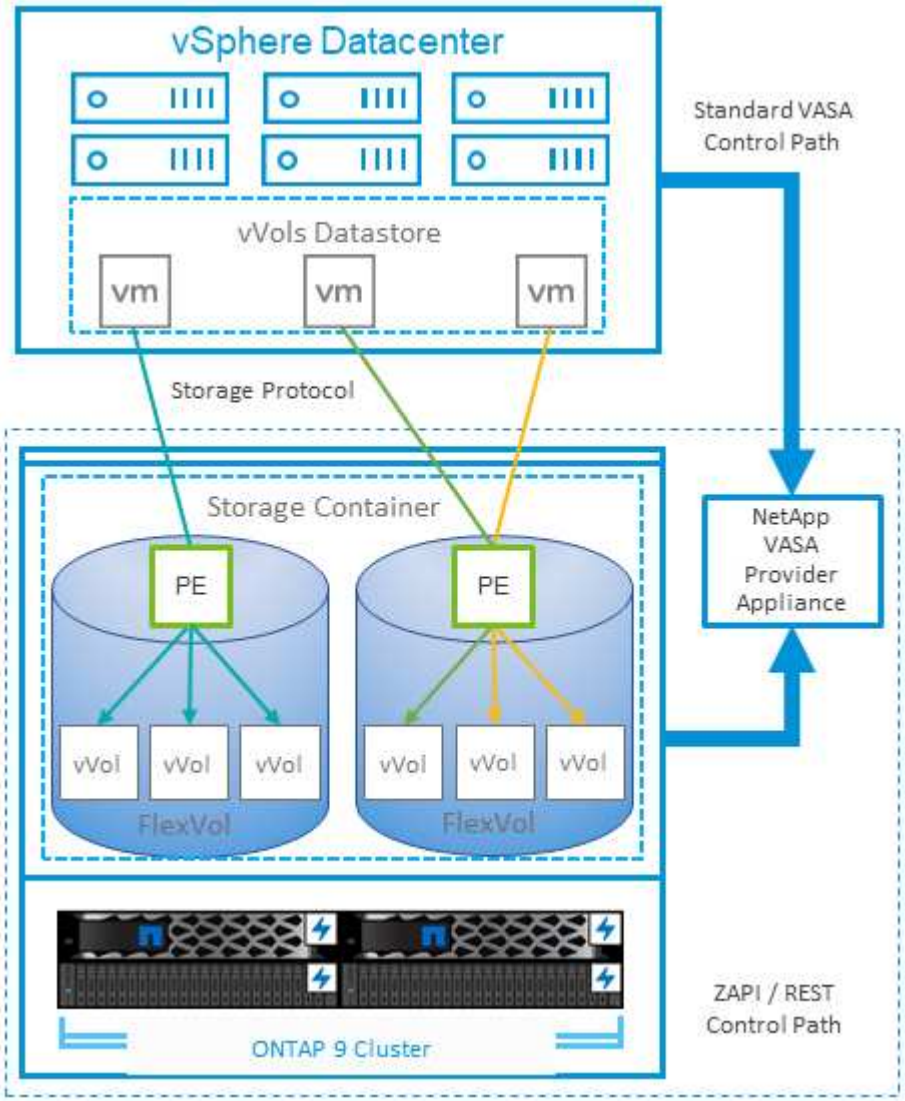
VASA provides the technology to query storage and return a set of storage capabilities to vCenter. VASA vendor providers supply the translation between the storage system APIs and constructs and the VMware APIs that are understood by vCenter. NetApp's VASA Provider for ONTAP is offered as part of the ONTAP tools for VMware vSphere appliance VM, and the vCenter plug-in provides the interface to provision and manage vVol datastores, as well as the ability to define storage capability profiles (SCPs).

ONTAP supports both VMFS and NFS vVol datastores. Using vVols with SAN datastores brings some of the benefits of NFS such as VM-level granularity. Here are some best practices to consider, and you can find additional information in TR-4400:

- A vVol datastore can consist of multiple FlexVol volumes on multiple cluster nodes. The simplest approach is a single datastore, even when the volumes have different capabilities. SPBM makes sure that a compatible volume is used for the VM. However, the volumes must all be part of a single ONTAP SVM and accessed using a single protocol. One LIF per node for each protocol is sufficient. Avoid using multiple ONTAP releases within a single vVol datastore because the storage capabilities might vary across releases.

- Use the ONTAP tools for VMware vSphere plug-in to create and manage vVol datastores. In addition to managing the datastore and its profile, it automatically creates a protocol endpoint to access the vVols if needed. If LUNs are used, note that LUN PEs are mapped using LUN IDs 300 and higher. Verify that the ESXi host advanced system setting `Disk.MaxLUN` allows a LUN ID number that is higher than 300 (the default is 1,024). Do this step by selecting the ESXi host in vCenter, then the Configure tab, and find `Disk.MaxLUN` in the list of Advanced System Settings.
- Do not install or migrate VASA Provider, vCenter Server (appliance or Windows based), or ONTAP tools for VMware vSphere itself onto a vVols datastore, because they are then mutually dependent, limiting your ability to manage them in the event of a power outage or other data center disruption.
- Back up the VASA Provider VM regularly. At a minimum, create hourly snapshots of the traditional datastore that contains VASA Provider. For more about protecting and recovering the VASA Provider, see this KB article.

The following figure shows vVols components.



Cloud migration and backup

Another ONTAP strength is broad support for the hybrid cloud, merging systems in your on-premises private cloud with public cloud capabilities. Here are some NetApp cloud solutions that can be used in conjunction with vSphere:

- **Cloud Volumes.** NetApp Cloud Volumes Service for Amazon Web Services or Google Cloud Platform and Azure NetApp Files for ANF provide high-performance, multi-protocol managed storage services in the leading public cloud environments. They can be used directly by VMware Cloud VM guests.
- **Cloud Volumes ONTAP.** NetApp Cloud Volumes ONTAP data management software delivers control, protection, flexibility, and efficiency to your data on your choice of cloud. Cloud Volumes ONTAP is cloud-native data management software built on ONTAP storage. Use together with Cloud Manager to deploy and manage Cloud Volumes ONTAP instances together with your on-premises ONTAP systems. Take advantage of advanced NAS and iSCSI SAN capabilities together with unified data management, including snapshots and SnapMirror replication.
- **Cloud Services.** Use Cloud Backup Service or SnapMirror Cloud to protect data from on-premises systems using public cloud storage. Cloud Sync helps migrate and keep your data in sync across NAS, object stores, and Cloud Volumes Service storage.
- **FabricPool.** FabricPool offers quick and easy tiering for ONTAP data. Cold blocks can be migrated to an object store in either public clouds or a private StorageGRID object store and are automatically recalled when the ONTAP data is accessed again. Or use the object tier as a third level of protection for data that is already managed by SnapVault. This approach can allow you to store more snapshots of your VMs on primary and/or secondary ONTAP storage systems.
- **ONTAP Select.** Use NetApp software-defined storage to extend your private cloud across the Internet to remote facilities and offices, where you can use ONTAP Select to support block and file services as well as the same vSphere data management capabilities you have in your enterprise data center.

When designing your VM-based applications, consider future cloud mobility. For example, rather than placing application and data files together use a separate LUN or NFS export for the data. This allows you to migrate the VM and data separately to cloud services.

Encryption for vSphere data

Today, there are increasing demands to protect data at rest through encryption. Although the initial focus was on financial and healthcare information, there is growing interest in protecting all information, whether it's stored in files, databases, or other data types.

Systems running ONTAP software make it easy to protect any data with at-rest encryption. NetApp Storage Encryption (NSE) uses self-encrypting disk drives with ONTAP to protect SAN and NAS data. NetApp also offers NetApp Volume Encryption and NetApp Aggregate Encryption as a simple, software-based approach to encrypt volumes on any disk drives. This software encryption doesn't require special disk drives or external key managers and is available to ONTAP customers at no additional cost. You can upgrade and start using it without any disruption to your clients or applications, and they are validated to the FIPS 140-2 level 1 standard, including the onboard key manager.

There are several approaches for protecting the data of virtualized applications running on VMware vSphere. One approach is to protect the data with software inside the VM at the guest OS level. Newer hypervisors such as vSphere 6.5 now support encryption at the VM level as another alternative. However, NetApp software encryption is simple and easy and has these benefits:

- **No effect on the virtual server CPU.** Some virtual server environments need every available CPU cycle for their applications, yet tests have shown up to 5x CPU resources are needed with hypervisor-level encryption. Even if the encryption software supports Intel's AES-NI instruction set to offload encryption workload (as NetApp software encryption does), this approach might not be feasible due to the requirement for new CPUs that are not compatible with older servers.
- **Onboard key manager included.** NetApp software encryption includes an onboard key manager at no additional cost, which makes it easy to get started without high-availability key management servers that are complex to purchase and use.

- **No effect on storage efficiency.** Storage efficiency techniques such as deduplication and compression are widely used today and are key to using flash disk media cost-effectively. However, encrypted data cannot typically be deduplicated or compressed. NetApp hardware and storage encryption operate at a lower level and allow full use of industry-leading NetApp storage efficiency features, unlike other approaches.
- **Easy datastore granular encryption.** With NetApp Volume Encryption, each volume gets its own AES 256-bit key. If you need to change it, you can do so with a single command. This approach is great if you have multiple tenants or need to prove independent encryption for different departments or apps. This encryption is managed at the datastore level, which is a lot easier than managing individual VMs.

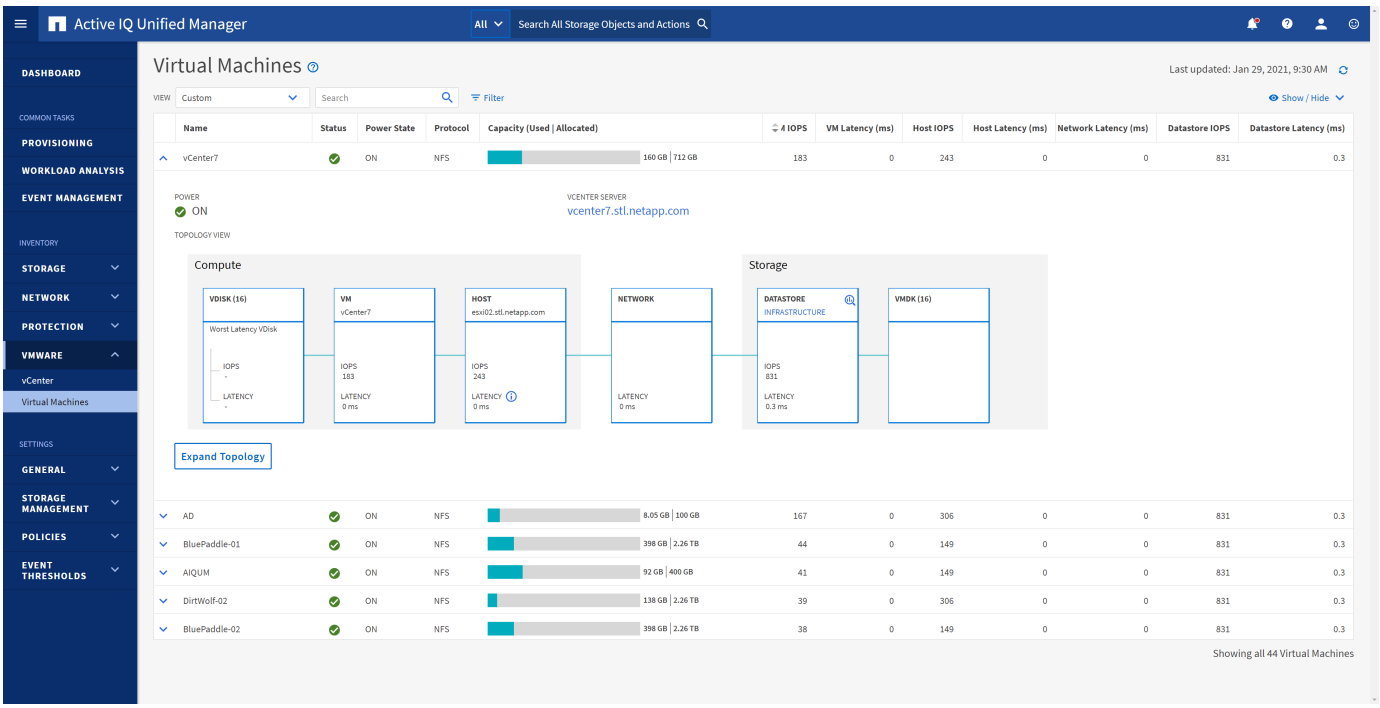
It's simple to get started with software encryption. After the license is installed, simply configure the onboard key manager by specifying a passphrase and then either create a new volume or do a storage-side volume move to enable encryption. NetApp is working to add more integrated support for encryption capabilities in future releases of its VMware tools.

Active IQ Unified Manager

Active IQ Unified Manager provides visibility into the VMs in your virtual infrastructure and enables monitoring and troubleshooting storage and performance issues in your virtual environment.

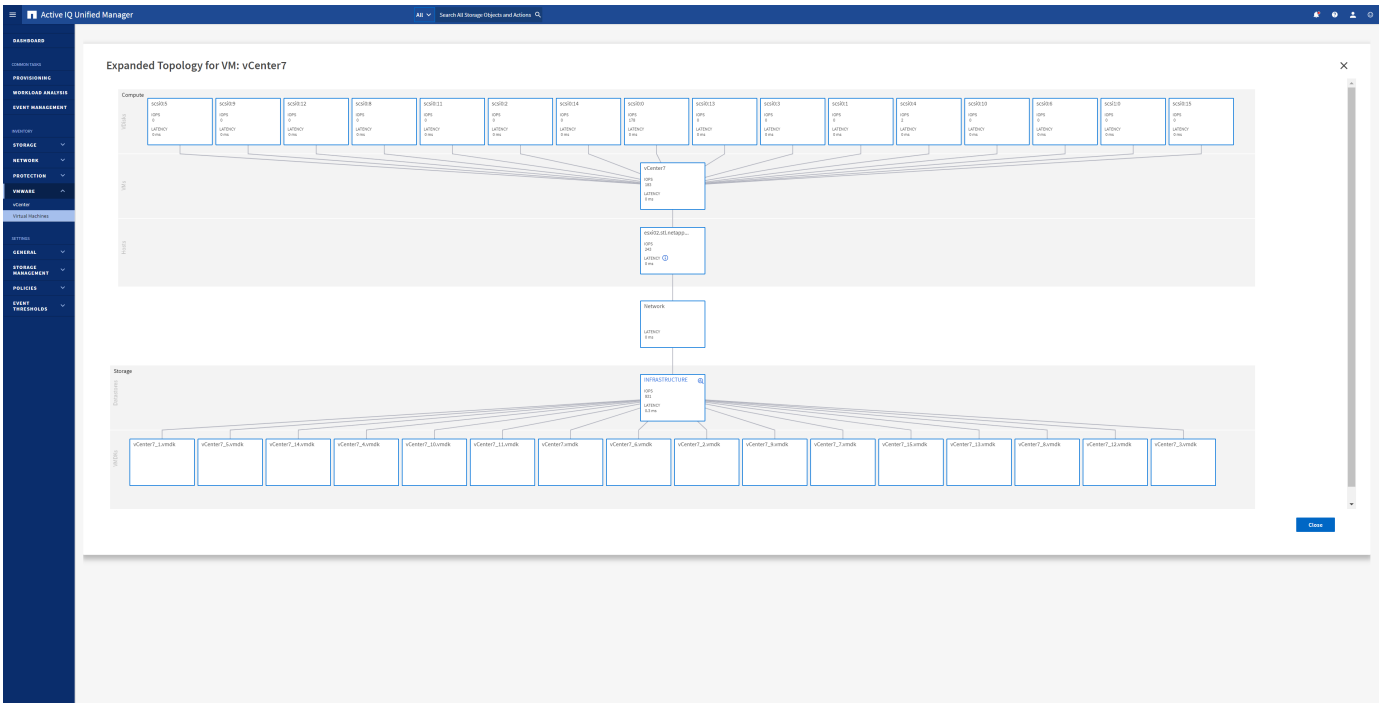
A typical virtual infrastructure deployment on ONTAP has various components that are spread across compute, network, and storage layers. Any performance lag in a VM application might occur due to a combination of latencies faced by the various components at the respective layers.

The following screenshot shows the Active IQ Unified Manager Virtual Machines view.



Unified Manager presents the underlying sub-system of a virtual environment in a topological view for determining whether a latency issue has occurred in the compute node, network, or storage. The view also highlights the specific object that causes the performance lag for taking remedial steps and addressing the underlying issue.

The following screenshot shows the AIQUM expanded topology.



Storage policy based management and vVols

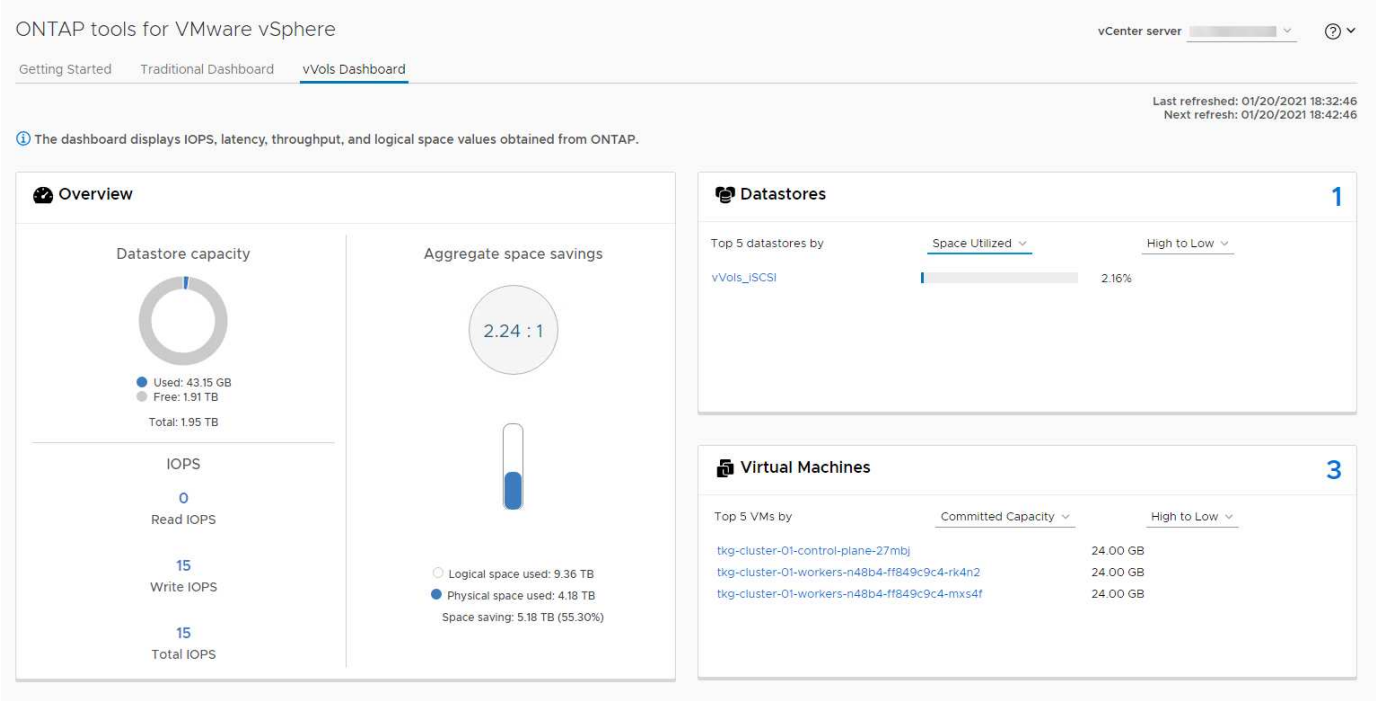
VMware vSphere APIs for Storage Awareness (VASA) make it easy for a storage administrator to configure datastores with well-defined capabilities and let the VM administrator use those whenever needed to provision VMs without having to interact with each other.

It's worth taking a look at this approach to see how it can streamline your virtualization storage operations and avoid a lot of trivial work.

Prior to VASA, VM administrators could define VM storage policies, but they had to work with the storage administrator to identify appropriate datastores, often by using documentation or naming conventions. With VASA, the storage administrator can define a range of storage capabilities, including performance, tiering, encryption, and replication. A set of capabilities for a volume or a set of volumes is called a storage capability profile (SCP).

The SCP supports minimum and/or maximum QoS for a VM's data vVols. Minimum QoS is supported only on AFF systems. ONTAP tools for VMware vSphere includes a dashboard that displays VM granular performance and logical capacity for vVols on ONTAP systems.

The following figure depicts ONTAP tools for VMware vSphere 9.8 vVols dashboard.



After the storage capability profile is defined, it can be used to provision VMs using the storage policy that identifies its requirements. The mapping between the VM storage policy and the datastore storage capability profile allows vCenter to display a list of compatible datastores for selection. This approach is known as storage policy based management.

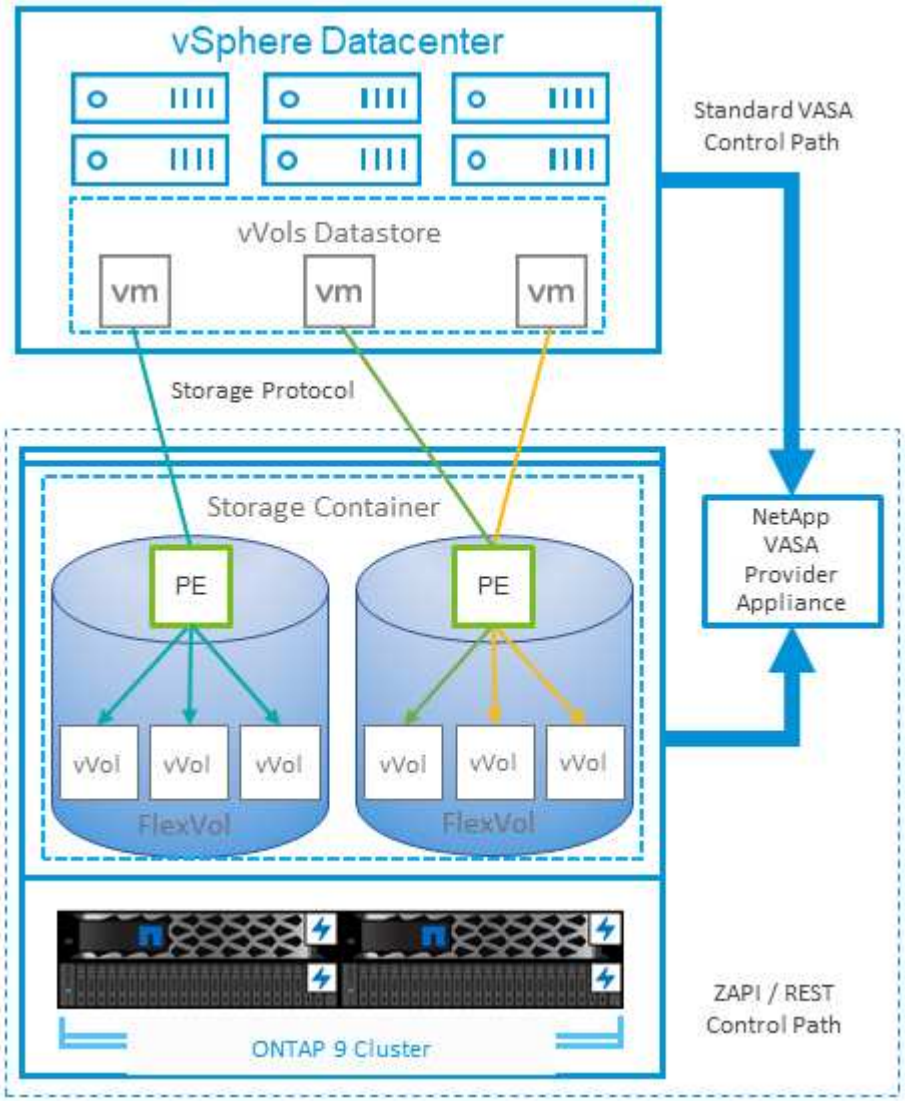
VASA provides the technology to query storage and return a set of storage capabilities to vCenter. VASA vendor providers supply the translation between the storage system APIs and constructs and the VMware APIs that are understood by vCenter. NetApp's VASA Provider for ONTAP is offered as part of the ONTAP tools for VMware vSphere appliance VM, and the vCenter plug-in provides the interface to provision and manage vVol datastores, as well as the ability to define storage capability profiles (SCPs).

ONTAP supports both VMFS and NFS vVol datastores. Using vVols with SAN datastores brings some of the benefits of NFS such as VM-level granularity. Here are some best practices to consider, and you can find additional information in TR-4400:

- A vVol datastore can consist of multiple FlexVol volumes on multiple cluster nodes. The simplest approach is a single datastore, even when the volumes have different capabilities. SPBM makes sure that a compatible volume is used for the VM. However, the volumes must all be part of a single ONTAP SVM and accessed using a single protocol. One LIF per node for each protocol is sufficient. Avoid using multiple ONTAP releases within a single vVol datastore because the storage capabilities might vary across releases.
- Use the ONTAP tools for VMware vSphere plug-in to create and manage vVol datastores. In addition to managing the datastore and its profile, it automatically creates a protocol endpoint to access the vVols if needed. If LUNs are used, note that LUN PEs are mapped using LUN IDs 300 and higher. Verify that the ESXi host advanced system setting `Disk.MaxLUN` allows a LUN ID number that is higher than 300 (the default is 1,024). Do this step by selecting the ESXi host in vCenter, then the Configure tab, and find `Disk.MaxLUN` in the list of Advanced System Settings.
- Do not install or migrate VASA Provider, vCenter Server (appliance or Windows based), or ONTAP tools for VMware vSphere itself onto a vVols datastore, because they are then mutually dependent, limiting your ability to manage them in the event of a power outage or other data center disruption.
- Back up the VASA Provider VM regularly. At a minimum, create hourly snapshots of the traditional datastore that contains VASA Provider. For more about protecting and recovering the VASA Provider, see

this KB article.

The following figure shows vVols components.



VMware Storage Distributed Resource Scheduler

VMware Storage Distributed Resource Scheduler (SDRS) is a vSphere feature that places VMs on storage based on the current I/O latency and space usage.

It then moves the VM or VMDKs nondisruptively between the datastores in a datastore cluster (also referred to as a pod), selecting the best datastore in which to place the VM or VMDKs in the datastore cluster. A datastore cluster is a collection of similar datastores that are aggregated into a single unit of consumption from the vSphere administrator's perspective.

When using SDRS with ONTAP tools for VMware vSphere, you must first create a datastore with the plug-in, use vCenter to create the datastore cluster, and then add the datastore to it. After the datastore cluster is created, additional datastores can be added to the datastore cluster directly from the provisioning wizard on the Details page.

Other ONTAP best practices for SDRS include the following:

- All datastores in the cluster should use the same type of storage (such as SAS, SATA, or SSD), be either all VMFS or NFS datastores, and have the same replication and protection settings.
- Consider using SDRS in default (manual) mode. This approach allows you to review the recommendations and decide whether to apply them or not. Be aware of these effects of VMDK migrations:
  - When SDRS moves VMDKs between datastores, any space savings from ONTAP cloning or deduplication are lost. You can rerun deduplication to regain these savings.
  - After SDRS moves VMDKs, NetApp recommends recreating the snapshots at the source datastore because space is otherwise locked by the VM that was moved.
  - Moving VMDKs between datastores on the same aggregate has little benefit, and SDRS does not have visibility into other workloads that might share the aggregate.

Recommended ESXi host and other ONTAP settings

NetApp has developed a set of optimal ESXi host settings for both NFS and block protocols. Specific guidance is also provided for multipathing and HBA timeout settings for proper behavior with ONTAP based on NetApp and VMware internal testing.

These values are easily set using ONTAP tools for VMware vSphere: From the Summary dashboard, click Edit Settings in the Host Systems portlet or right-click the host in vCenter, then navigate to ONTAP tools > Set Recommended Values.

Here are the currently recommended host settings with the 9.8-9.13 releases.

Host Setting	NetApp Recommended Value	Reboot Required
ESXi Advanced Configuration		
VMFS3.HardwareAcceleratedLocking	Keep default (1)	No
VMFS3.EnableBlockDelete	Keep default (0), but can be changed if needed. For more information, see VMware KB 2007427	No
VMFS3.EnableVMFS6Unmap	Keep default (1) For more information, see VMware vSphere APIs: Array Integration (VAAI)	No
NFS Settings		
Net.TcpipHeapSize	vSphere 6.0 or later, set to 32. All other NFS configurations, set to 30	Yes
Net.TcpipHeapMax	Set to 512MB for most vSphere 6.X releases. Set to 1024MB for 6.5U3, 6.7U3, and 7.0 or later.	Yes
NFS.MaxVolumes	vSphere 6.0 or later, set to 256 All other NFS configurations set to 64.	No
NFS41.MaxVolumes	vSphere 6.0 or later, set to 256.	No
NFS.MaxQueueDepth[1]	vSphere 6.0 or later, set to 128	Yes
NFS.HeartbeatMaxFailures	Set to 10 for all NFS configurations	No

NFS.HeartbeatFrequency	Set to 12 for all NFS configurations	No
NFS.HeartbeatTimeout	Set to 5 for all NFS configurations.	No
SunRPC.MaxConnPerIP	vSphere 7.0 or later, set to 128.	No
FC/FCoE Settings		
Path selection policy	Set to RR (round robin) when FC paths with ALUA are used. Set to FIXED for all other configurations. Setting this value to RR helps provide load balancing across all active/optimized paths. The value FIXED is for older, non-ALUA configurations and helps prevent proxy I/O. In other words, it helps keep I/O from going to the other node of a high-availability (HA) pair in an environment that has Data ONTAP operating in 7-Mode	No
Disk.QFullSampleSize	Set to 32 for all configurations. Setting this value helps prevent I/O errors.	No
Disk.QFullThreshold	Set to 8 for all configurations. Setting this value helps prevent I/O errors.	No
Emulex FC HBA timeouts	Use the default value.	No
QLogic FC HBA timeouts	Use the default value.	No
iSCSI Settings		
Path selection policy	Set to RR (round robin) for all iSCSI paths. Setting this value to RR helps provide load balancing across all active/optimized paths.	No
Disk.QFullSampleSize	Set to 32 for all configurations. Setting this value helps prevent I/O errors	No
Disk.QFullThreshold	Set to 8 for all configurations. Setting this value helps prevent I/O errors.	No

1 - NFS advanced configuration option MaxQueueDepth may not work as intended when using VMware vSphere ESXi 7.0.1 and VMware vSphere ESXi 7.0.2. Please reference VMware KB 86331 for more information.

ONTAP tools also specify certain default settings when creating ONTAP FlexVol volumes and LUNs:

ONTAP Tool	Default Setting
Snapshot reserve (-percent-snapshot-space)	0
Fractional reserve (-fractional-reserve)	0
Access time update (-atime-update)	False
Minimum readahead (-min-readahead)	False

Scheduled snapshots	None
Storage efficiency	Enabled
Volume guarantee	None (thin provisioned)
Volume Autosize	grow_shrink
LUN space reservation	Disabled
LUN space allocation	Enabled

Multipath settings for performance

While not currently configured by available ONTAP tools, NetApp suggests these configuration options:

- In high-performance environments or when testing performance with a single LUN datastore, consider changing the load balance setting of the round-robin (VMW_PSP_RR) path selection policy (PSP) from the default IOPS setting of 1000 to a value of 1. See VMware KB 2069356 for more info.
- In vSphere 6.7 Update 1, VMware introduced a new latency load balance mechanism for the Round Robin PSP. The new option considers I/O bandwidth and path latency when selecting the optimal path for I/O. You might benefit from using it in environments with non-equivalent path connectivity, such as cases with more network hops on one path than another, or when using a NetApp All SAN Array system. See Path Selection Plug-Ins and Policies for more information.

Additional documentation

For FCP and iSCSI with vSphere 7, more details can be found at Use VMware vSphere 7.x with ONTAP
For FCP and iSCSI with vSphere 8, more details can be found at Use VMware vSphere 8.x with ONTAP
For NVMe-oF with vSphere 7, more details can be found at For NVMe-oF, more details can be found at NVMe-oF Host Configuration for ESXi 7.x with ONTAP
For NVMe-oF with vSphere 8, more details can be found at For NVMe-oF, more details can be found at NVMe-oF Host Configuration for ESXi 8.x with ONTAP

Virtual Volumes (vVols) with ONTAP

Overview

ONTAP has been a leading storage solution for VMware vSphere environments for over two decades and continues to add innovative capabilities to simplify management while reducing costs.

This document covers ONTAP capabilities for VMware vSphere Virtual Volumes (vVols), including the latest product information and use cases along with best practices and other information to streamline deployment and reduce errors.

> This documentation replaces previously published technical reports *TR-4400: VMware vSphere Virtual Volumes (vVols) with ONTAP*

Best practices supplement other documents such as guides and compatibility lists. They are developed based on lab testing and extensive field experience by NetApp engineers and customers. They might not be the only practices that work or are supported but are generally the simplest solutions that meet the needs of most customers.

This document has been updated to include new vVols features found in vSphere 8.0 update 1 which are supported with the ONTAP tools 9.12 release.

Virtual Volumes (vVols) overview

NetApp began working with VMware to support vSphere APIs for Storage Awareness (VASA) for vSphere 5 in 2012. This early VASA Provider allowed for the definition of storage capabilities in a profile that could be used to filter datastores when provisioning and for checking compliance with the policy afterwards. Over time this evolved to add new capabilities to enable more automation in provisioning, as well as adding Virtual Volumes or vVols, where individual storage objects are used for virtual machine files and virtual disks. These objects could be LUNs, files, and now with vSphere 8 - NVMe namespaces.NetApp worked closely with VMware as a reference partner for vVols released with vSphere 6 in 2015, and again as a design partner for vVols using NVMe over fabrics in vSphere 8. NetApp continues to enhance vVols to take advantage of the latest capabilities in ONTAP.

There are several components to be aware of:

VASA Provider
This is the software component that handles communication between VMware vSphere and the storage system. For ONTAP, the VASA Provider runs in an appliance known as ONTAP tools for VMware vSphere (ONTAP tools for short). ONTAP tools also includes a vCenter plugin, a storage replication adapter (SRA) for VMware Site Recovery Manager, and REST API server for building your own automation. Once ONTAP tools is configured and registered with vCenter, there is little need to directly interact with the ONTAP system anymore, since nearly all of your storage needs can be managed from directly within the vCenter UI, or through REST API automation.
Protocol Endpoint (PE)
The protocol endpoint is a proxy for I/O between the ESXi hosts and the vVols datastore. The ONTAP VASA Provider creates these automatically, either one protocol endpoint LUN (4MB in size) per FlexVol volume of the vVols datastore, or one NFS mount point per NFS interface (LIF) on the storage node hosting a FlexVol volume in the datastore. The ESXi host mounts these protocol endpoints directly rather than individual vVol LUNs and virtual disk files. There is no need to manage the protocol endpoints as they are created, mounted, unmounted, and deleted automatically by the VASA Provider, along with any necessary interface groups or export policies.
Virtual Protocol Endpoint (vPE)
New in vSphere 8, when using NVMe over Fabrics (NVMe-oF) with vVols, the concept of a protocol endpoint is no longer relevant in ONTAP. Instead, a virtual PE is instantiated automatically by the ESXi host for each ANA group as soon as the first VM is powered on. ONTAP automatically creates ANA groups for each FlexVol volume used by the datastore. An additional advantage to using NVMe-oF for vVols is that there are no bind requests required of the VASA Provider. Instead, the ESXi host handles vVol binding functionality internally based on the vPE. This reduces the opportunity for a vVol bind storm to impact service. For more information, see NVMe and Virtual Volumes on VMware.com
Virtual Volume datastore

The Virtual Volume datastore is a logical datastore representation of a vVols container which is created and maintained by a VASA Provider. The container represents a pool of storage capacity provisioned from storage systems managed by the VASA Provider. ONTAP tools supports allocating multiple FlexVol volumes (referred to as backing volumes) to a single vVols datastore, and these vVols datastores can span multiple nodes in an ONTAP cluster, combining flash and hybrid systems with different capabilities. The administrator may create new FlexVol volumes using the provisioning wizard or REST API, or select pre-created FlexVol volumes for backing storage if they are available.
Virtual Volumes (vVols)
vVols are the actual virtual machine files and disks stored in the vVols datastore. Using the term vVol (singular) is referring to a single specific file, LUN, or namespace. ONTAP creates NVMe namespaces, LUNs or files depending on what protocol the datastore uses. There are several distinct types of vVols; most common are Config (metadata files), Data (virtual disk or VMDK), and Swap (created when VM is powered on). vVols protected by VMware VM encryption will be of type Other. VMware VM encryption should not be confused with ONTAP volume or aggregate encryption.

Policy based management

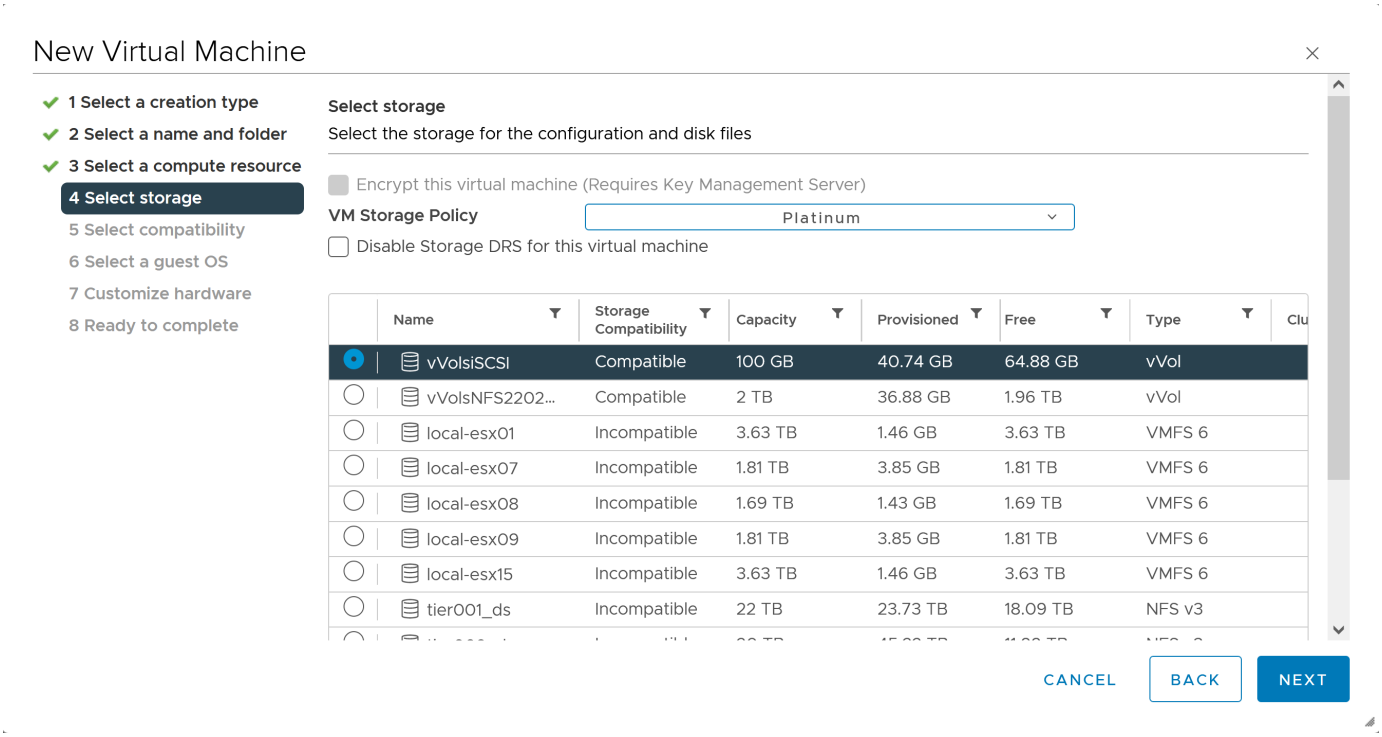
VMware vSphere APIs for Storage Awareness (VASA) make it easy for a VM administrator to use whatever storage capabilities are needed to provision VMs without having to interact with their storage team. Prior to VASA, VM administrators could define VM storage policies, but had to work with their storage administrators to identify appropriate datastores, often by using documentation or naming conventions. With VASA, vCenter administrators with the appropriate permissions can define a range of storage capabilities which vCenter users can then use to provision VMs. The mapping between VM storage policy and datastore storage capability profile allows vCenter to display a list of compatible datastores for selection, as well as enabling other technologies like Aria (formerly known as vRealize) Automation or Tanzu Kubernetes Grid to automatically select storage from an assigned policy. This approach is known as storage policy based management. While storage capability profiles and policies may also be used with traditional datastores, our focus here is on vVols datastores.

There are two elements:

Storage Capability Profile (SCP)
A storage capability profile (SCP) is a form of storage template that allows the vCenter admin to define what storage features they require without actually needing to understand how to manage those features in ONTAP. By taking a template style approach, it allows the admin to easily deliver storage services in a consistent and predictable way. Capabilities described in an SCP include performance, protocol, storage efficiency, and other features. Specific features vary by version. They are created using the ONTAP tools for VMware vSphere menu within the vCenter UI. You can also use REST APIs to create SCPs. They may be manually created by selecting individual capabilities, or automatically generated from existing (traditional) datastores.
VM Storage Policy
VM Storage Policies are created in vCenter under Policies and Profiles. For vVols, create a ruleset using rules from the NetApp vVols storage type provider. ONTAP tools provides a simplified approach by allowing you to simply select an SCP rather than forcing you to specify individual rules.

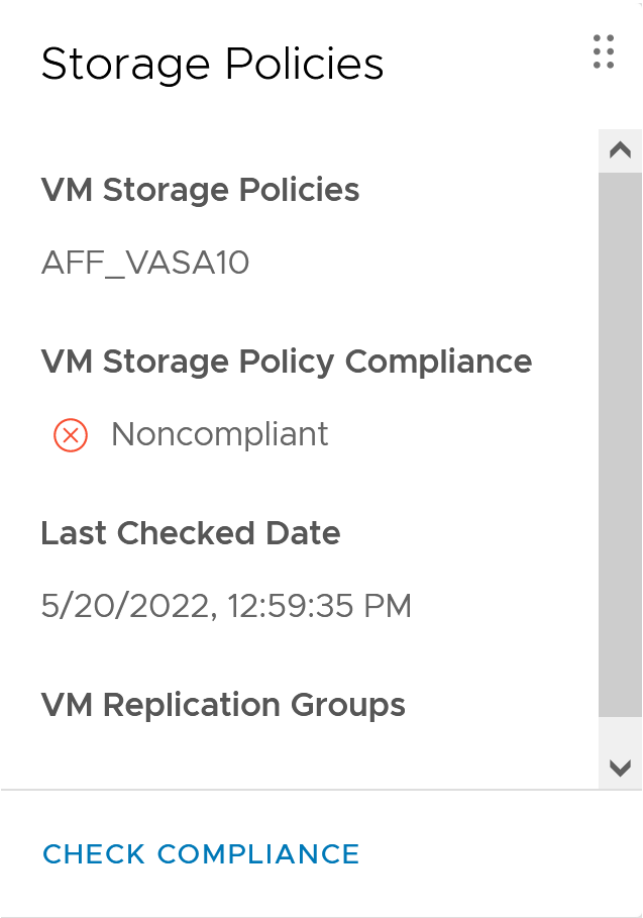
As mentioned above, using policies can help streamline the task of provisioning a volume. Simply select an appropriate policy, and the VASA Provider will show vVols datastores that support that policy and place the vVol into an individual FlexVol volume that is compliant (Figure 1).

Deploy VM using Storage Policy



Once a VM is provisioned, the VASA Provider will continue to check compliance, and alert the VM administrator with an alarm in vCenter when the backing volume is no longer compliant with the policy (Figure 2).

VM Storage Policy Compliance



NetApp vVols suppor

ONTAP has supported the VASA specification since its initial release in 2012. While other NetApp storage systems may support VASA, this document focuses on currently supported releases of ONTAP 9.

ONTAP

In addition to ONTAP 9 on AFF, ASA, and FAS systems, NetApp supports VMware workloads on ONTAP Select, Amazon FSx for NetApp with VMware Cloud on AWS, Azure NetApp Files with Azure VMware Solution, Cloud Volumes Service with Google Cloud VMware Engine, and NetApp Private Storage in Equinix, but specific functionality may vary based on service provider and available network connectivity. Access from vSphere guests to data stored in those configurations as well as Cloud Volumes ONTAP is also available.

At the time of publication, hyperscaler environments are limited to traditional NFS v3 datastores only, therefore, vVols are only available with on-premises ONTAP systems, or cloud connected systems that offer the full functionality of an on-premises systems such as those hosted by NetApp partners and services providers around the world.

For more information about ONTAP, see ONTAP product documentation

For more information about ONTAP and VMware vSphere best practices, see TR-4597

Benefits of using vVols with ONTAP

When VMware introduced vVols support with VASA 2.0 in 2015 they described it as “an integration and management framework delivering a new operational model for external storage (SAN/NAS).” This operational model offers several benefits together with ONTAP storage.

Policy based management

As covered in section 1.2, policy based management allows VMs to be provisioned and subsequently managed using pre-defined policies. This can help IT operations in several ways:

- **Increase velocity.** ONTAP tools eliminates the requirement for the vCenter administrator to open tickets with the storage team for storage provisioning activities. However, ONTAP tools RBAC roles in vCenter and on the ONTAP system still allow for independent teams (such as storage teams), or independent activities by the same team by restricting access to specific functions if desired.
- **Smarter provisioning.** Storage system capabilities can be exposed through the VASA APIs, allowing provisioning workflows to take advantage of advanced capabilities without the VM administrator needing to understand how to manage the storage system.
- **Faster provisioning.** Different storage capabilities can be supported in a single datastore and automatically selected as appropriate for a VM based on the VM policy.
- **Avoid mistakes.** Storage and VM policies are developed in advance and applied as needed without having to customize storage each time a VM is provisioned. Compliance alarms are raised when storage capabilities drift from the defined policies. As previously mentioned, SCPs make the initial provisioning predictable and repeatable, while basing VM storage policies on the SCPs guarantees accurate placement.
- **Better capacity management.** VASA and ONTAP tools make it possible to view storage capacity down to the induvial aggregate level if needed and provide multiple layers of alerting in the event capacity starts to run low.

VM granular management on the modern SAN

SAN storage systems using Fibre Channel and iSCSI were the first to be supported by VMware for ESX, but they have lacked the ability to manage individual VM files and disks from the storage system. Instead, LUNs are provisioned and VMFS manages the individual files. This makes it difficult for the storage system to directly manage individual VM storage performance, cloning, and protection. vVols bring storage granularity that customers using NFS storage already enjoy, with the robust, high performance SAN capabilities of ONTAP.

Now, with vSphere 8 and ONTAP tools for VMware vSphere 9.12 and later, those same granular controls used by vVols for legacy SCSI based protocols are now available in the modern Fibre Channel SAN using NVMe over Fabrics for even greater performance at scale. With vSphere 8.0 update 1, it is now possible to deploy a complete end-to-end NVMe solution using vVols without any I/O translation in the hypervisor storage stack.

Greater storage offload capabilities

While VAAI offers a variety of operations that are offloaded to storage, there are some gaps that are addressed by the VASA Provider. SAN VAAI is not able to offload VMware managed snapshots to the storage system. NFS VAAI can offload VM managed snapshots, but there are limitations placed a VM with storage native snapshots. Since vVols use individual LUNs, namespaces, or files for virtual machine disks, ONTAP can quickly and efficiently clone the files or LUNs to create VM-granular snapshots that no longer require delta files. NFS VAAI also does not support offloading clone operations for hot (powered on) Storage vMotion migrations. The VM must be powered off to allow offload of the migration when using VAAI with traditional NFS datastores. The VASA Provider in ONTAP tools allows for near instant, storage efficient clones for hot and cold migrations, and it also supports near instant copies for cross-volume migrations of vVols. Because of these significant storage efficiency benefits, you may be able to take full advantage of vVols workloads under the Efficiency Guarantee program. Likewise, if cross volume clones using VAAI don't meet your requirements, you will likely be able to solve your business challenge thanks to the improvements in the copy experience with vVols.

Common use cases for vVols

In addition to these benefits, we also see these common use cases for vVol storage:

- **On-Demand provisioning of VMs**
  - Private cloud or service provider IaaS.
  - Leverage automation and orchestration via the Aria (formerly vRealize) suite, OpenStack, etc.
- **First Class Disks (FCDs)**
  - VMware Tanzu Kubernetes Grid [TKG] persistent volumes.
  - Provide Amazon EBS-like services though independent VMDK lifecycle management.
- **On-Demand Provisioning of Temporary VMs**
  - Test/dev labs
  - Training environments

Common benefits with vVols

When used to their full advantage, such as in the above use cases, vVols provide the following specific improvements:

- Clones are quickly created within a single volume, or across multiple volumes in an ONTAP cluster, which is an advantage when compared to traditional VAAI enabled clones. They are also storage efficient. Clones within a volume use ONTAP file clone, which are like FlexClone volumes and only store changes from the source vVol file/LUN/namespace. So long-term VMs for production or other application purposes are created quickly, take minimal space, and can benefit from VM level protection (using NetApp SnapCenter plugin for VMware vSphere, VMware managed snapshots or VADP backup) and performance management (with ONTAP QoS).
- vVols are the ideal storage technology when using TKG with the vSphere CSI, providing discrete storage classes and capacities managed by the vCenter administrator.
- Amazon EBS-like services can be delivered through FCDs because an FCD VMDK, as the name suggests, is a first-class citizen in vSphere and has a lifecycle which can be independently managed separate from VMs that it might be attached to.

Using vVols with ONTAP

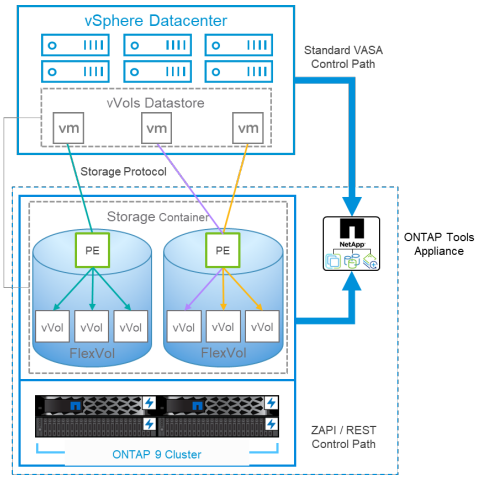
The key to using vVols with ONTAP is the VASA Provider software included as part of the ONTAP tools for VMware vSphere virtual appliance.

ONTAP tools also includes the vCenter UI extensions, REST API server, Storage Replication Adapter for VMware Site Recovery Manager, Monitoring and Host configuration tools, and an array of reports which help you better manage your VMware environment.

Products and Documentation

The ONTAP FlexClone license (included with ONTAP One) and the ONTAP tools appliance are the only additional products required to use vVols with ONTAP. Recent releases of ONTAP tools are supplied as a single unified appliance that runs on ESXi, providing the functionality of what formerly were three different appliances and servers. For vVols, it is important to use the ONTAP tools vCenter UI extensions or REST APIs as general management tools and user interfaces for ONTAP functions with vSphere, together with the VASA Provider which provides specific vVols functionality. The SRA component is included for traditional datastores, but VMware Site Recovery Manager does not use SRA for vVols, instead implementing new services in SRM 8.3 and later which leverage the VASA provider for vVols replication.

ONTAP tools VASA Provider architecture when using iSCSI or FCP



Product Installation

For new installations, deploy the virtual appliance into your vSphere environment. Current releases of ONTAP tools will automatically register themselves with your vCenter and enable the VASA Provider by default. In addition to ESXi host and vCenter Server information, you will also need the IP address configuration details for the appliance. As previously stated, the VASA Provider requires the ONTAP FlexClone license be already installed onto any ONTAP clusters you plan to use for vVols. The appliance has a built-in watchdog to ensure availability, and as a best practice should be configured with VMware High Availability and optionally Fault Tolerance features. See section 4.1 for additional details. Do not install or move the ONTAP tools appliance or vCenter Server appliance (VCSA) to vVols storage as this can prevent the appliances from restarting.

In-place upgrades of ONTAP tools are supported by using the upgrade ISO file available for download on the NetApp Support Site (NSS). Follow the Deployment and Setup Guide instructions to upgrade the appliance.

For sizing your virtual appliance, and understanding the configuration limits, refer to this knowledge base article: Sizing Guide for ONTAP tools for VMware vSphere

Product Documentation

The following documentation is available to help you deploy ONTAP tools.

For the complete documentation repository, visit this link to docs.netapp.com

Get started

- Release notes
- Learn about ONTAP tools for Vmware vSphere
- ONTAP tools Quick start
- Deploy ONTAP tools
- Upgrade ONTAP tools

Use ONTAP tools

- Provision traditional datastores
- Provision vVols datastores

- Configure role-based access control
- Configure remote diagnostics
- Configure high availability

Protect and manage datastores

- Protect traditional datastores with SRM
- Protect vVols based virtual machines with SRM
- Monitor traditional datastores and virtual machines
- Monitor vVols datastores and virtual machines

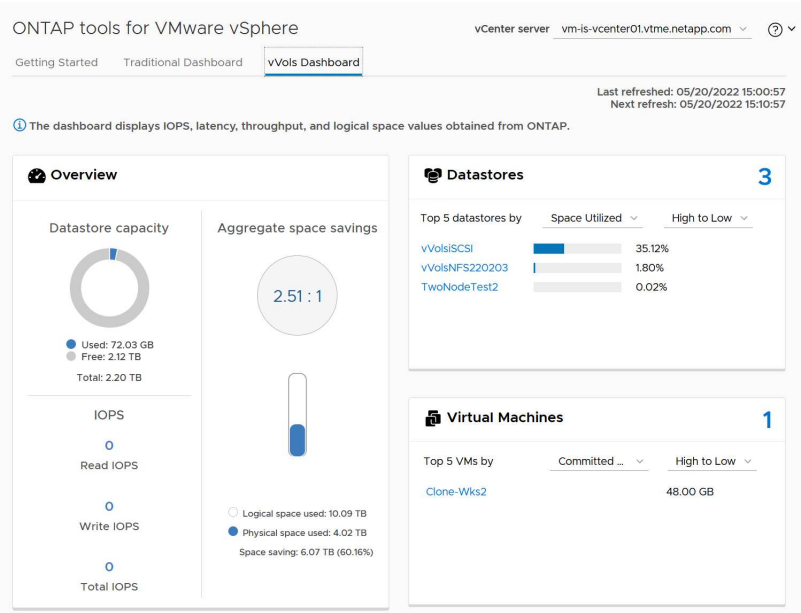
In addition to product documentation, there are Support Knowledgebase articles that may be useful.

- How to perform a VASA Provider Disaster Recovery - Resolution Guide

VASA Provider Dashboard

The VASA Provider includes a dashboard with performance and capacity information for individual vVols VMs. This information comes directly from ONTAP for the vVol files and LUNs, including latency, IOPS, throughput, and uptime for the top 5 VMs, and latency and IOPS for the top 5 datastores. It is enabled by default when using ONTAP 9.7 or later. It can take up to 30 minutes for initial data to be retrieved and displayed in the dashboard.

ONTAP tools vVols dashboard



Best Practices

Using ONTAP vVols with vSphere is simple and follows published vSphere methods (see Working with Virtual Volumes under vSphere Storage in VMware documentation for your version of ESXi). Here are a few additional practices to consider in conjunction with ONTAP.

Limits

In general, ONTAP supports vVols limits as defined by VMware (see published Configuration Maximums). The following table summarizes specific ONTAP limits in size and number of vVols. Always check the NetApp Hardware Universe for updated limits on numbers and sizes of LUNs and files.

ONTAP vVols Limits

Capacity/Feature	SAN (SCSI or NVMe-oF)	NFS
Maximum vVols size	62 TiB*	62 TiB*
Maximum number of vVols per FlexVol volume	1024	2 billion
Maximum number of vVols per ONTAP node	Up to 12,288**	50 billion
Maximum number of vVols per ONTAP pair	Up to 24,576**	50 billion
Maximum number of vVols per ONTAP cluster	Up to 98,304**	No specific cluster limit
Maximum QoS objects (shared policy group and individual vVols service level)	12,000 through ONTAP 9.3; 40,000 with ONTAP 9.4 and later	

- Size limit based on ASA systems or AFF and FAS systems running ONTAP 9.12.1P2 and later.
  - Number of SAN vVols (NVMe namespaces or LUNs) varies based on platform. Always check the NetApp Hardware Universe for updated limits on numbers and sizes of LUNs and files.

Use ONTAP tools for VMware vSphere's UI extensions or REST APIs to provision vVols datastores and Protocol Endpoints.

While it's possible to create vVols datastores with the general vSphere interface, using ONTAP tools will automatically create protocol endpoints as needed, and creates FlexVol volumes using ONTAP best practices and in compliance with your defined storage capability profiles. Simply right click on the host/cluster/datacenter, then select *ONTAP tools* and *Provision datastore*. From there simply choose the desired vVols options in the wizard.

Never store the ONTAP tools appliance or vCenter Server Appliance (VCSA) on a vVols datastore that they are managing.

This can result in a "chicken and egg situation" if you need to reboot the appliances because they won't be able to rebind their own vVols while they are rebooting. You may store them on a vVols datastore managed by a different ONTAP tools and vCenter deployment.

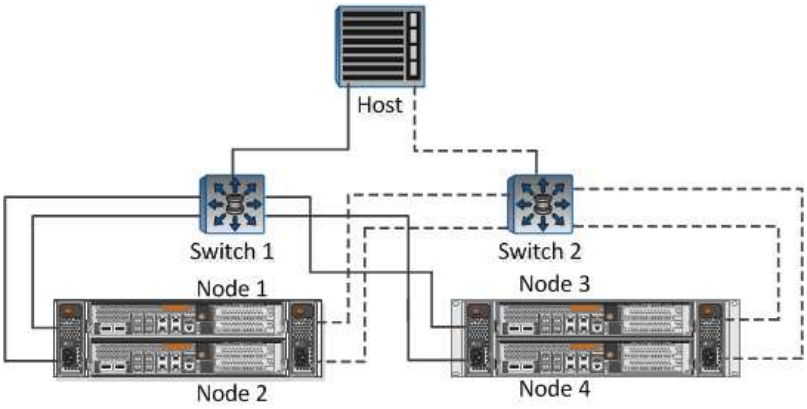
Avoid vVols operations across different ONTAP releases.

Supported storage capabilities such as QoS, personality and more have changed in various releases of the VASA Provider, and some are dependent on ONTAP release. Using different releases in an ONTAP cluster or moving vVols between clusters with different releases can result in unexpected behavior or compliance alarms.

Zone your Fibre Channel fabric before using NVMe/FC or FCP for vVols.

The ONTAP tools VASA provider takes care of managing FCP and iSCSI igroups as well as NVMe subsystems in ONTAP based on discovered initiators of managed ESXi hosts. However, it does not integrate with Fibre Channel switches to manage zoning. Zoning must be done according to best practices before any provisioning can take place. The following is an example of single initiator zoning to four ONTAP systems:

Single initiator zoning:



Refer to the following documents for more best practices:

TR-4080 Best practices for modern SAN ONTAP 9

TR-4684 Implementing and configuring modern SANs with NVMe-oF

Plan your backing FlexVols according to your needs.

It can be desirable to add several backing volumes to your vVols datastore to distribute workload across the ONTAP cluster, to support different policy options, or to increase the number of allowed LUNs or files. However, if maximum storage efficiency is required, then place all your backing volumes on a single aggregate. Or if maximum cloning performance is required, then consider using a single FlexVol volume and keeping your templates or content library in the same volume. The VASA Provider offloads many vVols storage operations to ONTAP, including migration, cloning and snapshots. When this is done within a single FlexVol volume, space efficient file clones are used and are almost instantly available. When this is done across FlexVol volumes, the copies are quickly available and use inline deduplication and compression, but maximum storage efficiency may not be recovered until background jobs run on volumes using background deduplication and compression. Depending on the source and destination, some efficiency may be degraded.

Keep Storage Capability Profiles (SCPs) simple.

Avoid specifying capabilities that aren't required by setting them to Any. This will minimize problems when selecting or creating FlexVol volumes. For example, with VASA Provider 7.1 and earlier, if compression is left at the default SCP setting of No, it will attempt to disable compression, even on an AFF system.

Use the default SCPs as example templates to create your own.

The included SCPs are suitable for most general-purpose uses, but your requirements may be different.

Consider using Max IOPS to control unknown or test VMs.

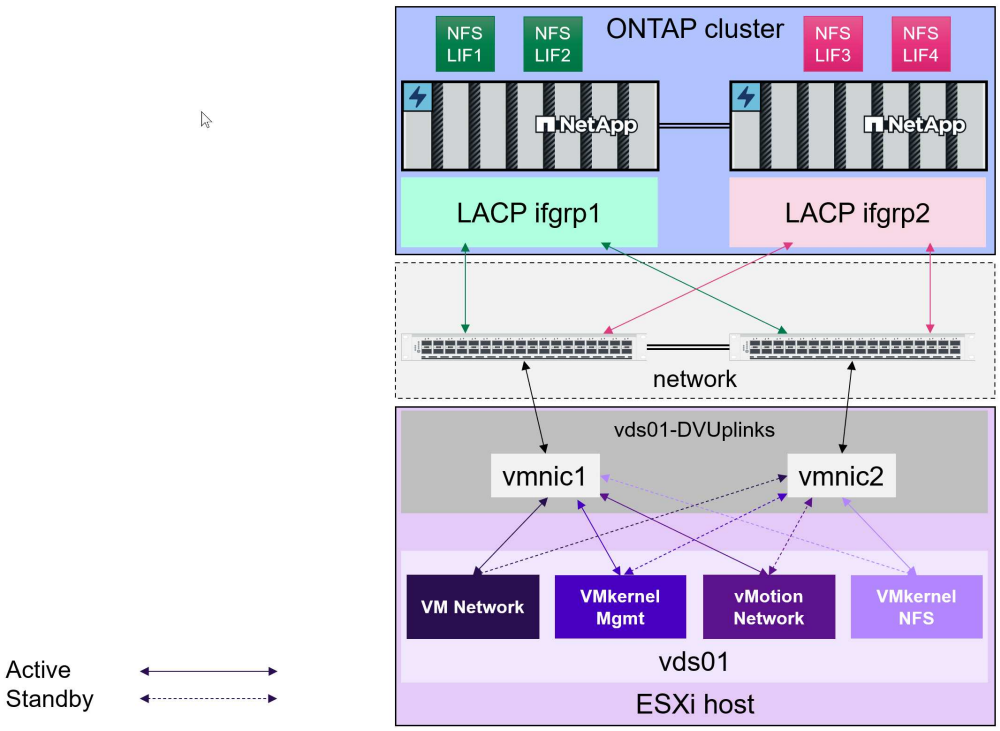
First available in VASA Provider 7.1, Max IOPS can be used to limit IOPS to a specific vVol for an unknown workload to avoid impact on other, more critical workloads. See Table 4 for more on performance management.

Ensure you have sufficient data LIFs.
Create at least two LIFs per node per HA pair. More may be required based on your workload.

Follow all protocol best practices.

Refer to NetApp and VMware's other best practice guides specific to the protocol you've selected. In general, there are not any changes other than those already mentioned.

Example network configuration using vVols over NFS v3



Deploying vVols Storage

There are several steps to creating vVols storage for your VMs.

The first two steps may not be needed for an existing vSphere environment that uses ONTAP for traditional datastores. You may already be using ONTAP tools for managing, automating, and reporting with your VMFS or traditional NFS based storage. These steps are covered in more detail in the following section.

1. Create the Storage Virtual Machine (SVM) and its protocol configuration. You will select NVMe/FC, NFSv3, NFSv4.1, iSCSI, FCP, or a mix of those options. You may use either ONTAP System Manager wizards, or the cluster shell command line.
    - At least one LIF per node for each switch/fabric connection. As a best practice, create two or more per node for FCP, iSCSI, or NVMe-based protocols.
    - Volumes may be created at this time, but it is simpler to let the *Provision Datastore* wizard create them. The only exception to this rule is if you plan to use vVols replication with VMware Site Recovery Manager. This is easier to set up with pre-existing FlexVol volumes with existing SnapMirror relationships. Be mindful to not enable QoS on any volumes to be used for vVols as this is intended to be managed by SPBM and ONTAP tools.
2. Deploy ONTAP tools for VMware vSphere using the OVA downloaded from the NetApp Support Site.
3. Configure ONTAP tools for your environment.
    - Add the ONTAP cluster to ONTAP tools under *Storage Systems*
        - While ONTAP tools and SRA support both cluster-level and SVM-level credentials, the VASA Provider supports only cluster-level credentials for storage systems. This is because many of the APIs used for vVols are only available at the cluster level. Therefore, if you plan to use vVols, you must add your ONTAP clusters using cluster-scoped credentials.
    - If your ONTAP data LIFs are on different subnets from your VMkernel adapters, then you must add the

VMkernel adapter subnets to the selected subnets list in the settings menu of ONTAP tools. By default, ONTAP tools will secure your storage traffic by only allowing local subnet access.

- The ONTAP tools comes with several pre-defined policies that may be used or see Managing VMs with policies for guidance on creating SCPs.

4. Use the *ONTAP tools* menu in vCenter to start the *Provision datastore* wizard.
5. Provide a meaningful name and select the desired protocol. You may provide a description of the datastore as well.
6. Select one or more SCPs to be supported by the vVols datastore. This will filter out any ONTAP systems which are unable to match the profile. From the resulting list, select your desired cluster and SVM.
7. Use the wizard to create new FlexVol volumes for each of the specified SCPs or use existing volumes by selecting the appropriate radio button.
8. Create VM policies for each SCP that will be used in the datastore from the *Policies and Profiles* menu in the vCenter UI.
9. Choose the "NetApp.clustered.Data.ONTAP.VP.vvol" storage rule set. The "NetApp.clustered.Data.ONTAP.VP.VASA10" storage rule set is for SPBM support with non-vVols datastores
10. You will specify the Storage Capability Profile by name when creating a VM Storage Policy. While at this step, you may also configure SnapMirror policy matching by using the replication tab, and tag-based matching using the tags tab. Note that tags must already be created in order to be selectable.
11. Create your VMs, selecting the VM Storage Policy and compatible datastore under Select storage.

Migrating VMs from traditional datastores to vVols

Migration of VMs from traditional datastores to a vVols datastore is as simple as moving VMs between traditional datastores. Simply select the VM(s), then select Migrate from the list of Actions, and select a migration type of *change storage only*. Migration copy operations will be offloaded with vSphere 6.0 and later for SAN VMFS to vVols migrations, but not from NAS VMDKs to vVols.

Managing VMs with policies

To automate storage provisioning with policy based management, we need to:

- Define the capabilities of the storage (ONTAP node and FlexVol volume) with Storage Capability Profiles (SCPs).
- Create VM storage policies that map to the defined SCPs.

NetApp has simplified the capabilities and mapping beginning with VASA Provider 7.2 with continuing improvements throughout later versions. This section focuses on this new approach. Earlier releases supported a greater number of capabilities and allowed them to be mapped individually to storage policies, but this approach is no longer supported.

Storage Capability Profile capabilities by ONTAP tools release

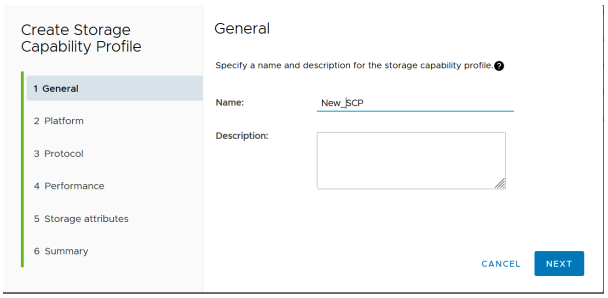
SCP Capability	Capability Values	Release Supported	Notes
Compression	Yes, No, Any	All	Mandatory for AFF in 7.2 and later.
Deduplication	Yes, No, Any	All	M andatory for AFF in 7.2 and later.

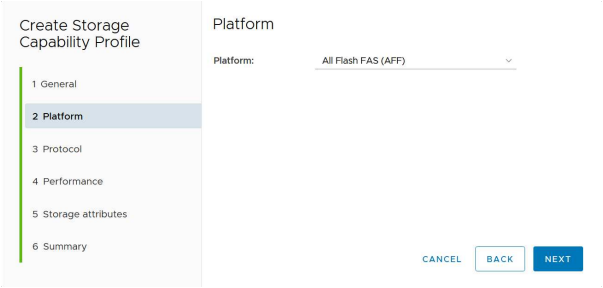
SCP Capability	Capability Values	Release Supported	Notes
Encryption	Yes, No, Any	7.2 and later	Selects/creates encrypted FlexVol volume.. ONTAP license required.
Max IOPS	<number>	7.1 and later, but differences	Listed under QoS Policy Group for 7.2 and later. See Performance management with ONTAP tools 9.10 and later for more information.
Personality	A FF, FAS	7.2 and later	FAS also includes other non-AFF systems, such as ONTAP Select. AFF includes ASA.
Protocol	NFS, NFS 4.1, iSCSI, FCP, NVMe/FC, Any	7.1 and earlier, 9.10 and later	7.2-9.8 is effectively “Any”. Beginning again in 9.10 where NFS 4.1 and NVMe/FC were added to the original list.
Space Reserve (Thin Provisioning)	Thin, Thick, (Any)	All, but differences	Called Thin Provisioning in 7.1 and earlier, which also allowed value of Any. Called Space Reserve in 7.2. All releases default to Thin.
Tiering Policy	Any, None, Snapshot, Auto	7.2 and later	Used for FabricPool - requires AFF or ASA with ONTAP 9.4 or later. Only Snapshot is recommended unless using an on-premise S3 solution like NetApp StorageGRID.

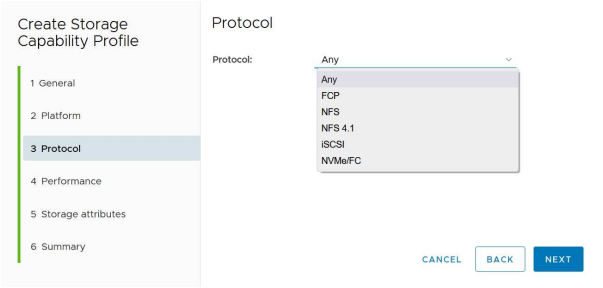
Creating Storage Capability Profiles

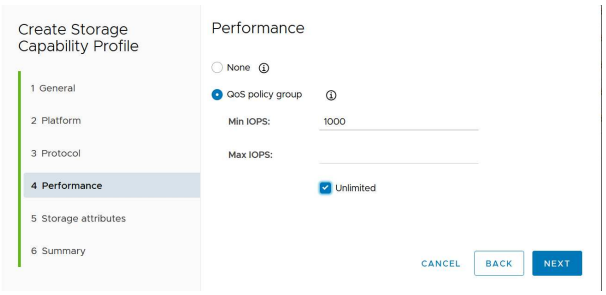
The NetApp VASA Provider comes with several pre-defined SCPs. New SCPs may be created manually, using the vCenter UI, or via automation using REST APIs. By specifying capabilities in a new profile, cloning an existing profile, or by auto-generating profile(s) from existing traditional datastores. This is done using the menus under ONTAP tools. Use *Storage Capability Profiles* to create or clone a profile, and *Storage Mapping* to auto-generate a profile.

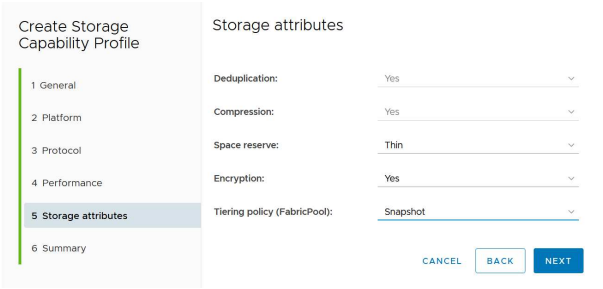
Storage Capabilities for ONTAP tools 9.10 and later

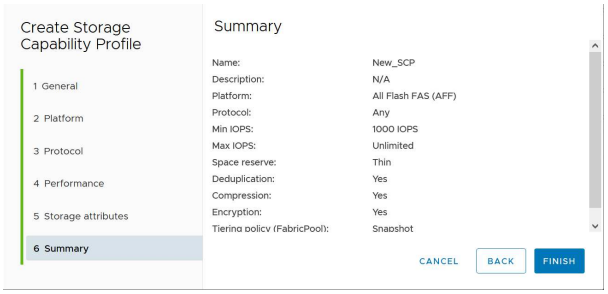

Create Storage Capability Profile
1 General
2 Platform
3 Protocol
4 Performance
5 Storage attributes
6 Summary
General
Specify a name and description for the storage capability profile.
Name:
New_SCP
Description:
CANCEL
NEXT


Create Storage Capability Profile
1 General
2 Platform
3 Protocol
4 Performance
5 Storage attributes
6 Summary
Platform
Platform:
All Flash FAS (AFF)
CANCEL
BACK
NEXT


Create Storage Capability Profile
1 General
2 Platform
3 Protocol
4 Performance
5 Storage attributes
6 Summary
Protocol
Protocol:
Any
Any
FCP
NFS
NFS 4.1
iSCSI
NVMe/FC
CANCEL
BACK
NEXT


Create Storage Capability Profile
1 General
2 Platform
3 Protocol
4 Performance
5 Storage attributes
6 Summary
Performance
None
QoS policy group
Min IOPS:
1000
Max IOPS:
Unlimited
CANCEL
BACK
NEXT


Create Storage Capability Profile
1 General
2 Platform
3 Protocol
4 Performance
5 Storage attributes
6 Summary
Storage attributes
Deduplication:
Yes
Compression:
Yes
Space reserve:
Thin
Encryption:
Yes
Tiering policy (FabricPool):
Snapshot
CANCEL
BACK
NEXT



Creating vVols Datastores

Once the necessary SCPs have been created, they may be used to create the vVols datastore (and optionally, FlexVol volumes for the datastore). Right-click on the host, cluster, or datacenter on which you want to create the vVols datastore, then select *ONTAP tools* > *Provision Datastore*. Select one or more SCPs to be supported by the datastore, then select from existing FlexVol volumes and/or provision new FlexVol volumes for the datastore. Finally, specify the default SCP for the datastore, which will be used for VMs that do not have an SCP specified by policy, as well as for swap vVols (these do not require high performance storage).

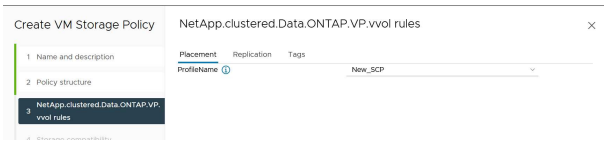
Creating VM Storage Policies

VM Storage Policies are used in vSphere to manage optional features such as Storage I/O Control or vSphere Encryption. They are also used with vVols to apply specific storage capabilities to the VM. Use the “NetApp.clustered.Data.ONTAP.VP.vvol” storage type and “ProfileName” rule to apply a specific SCP to VMs through use of the Policy. See example network configuration using vVols over NFS v3 for an example of this with the ONTAP tools VASA Provider. Rules for “NetApp.clustered.Data.ONTAP.VP.VASA10” storage are to be used with non-vVols based datastores.

Earlier releases are similar, but as mentioned in Storage Capability Profile capabilities by ONTAP tools release, your options will vary.

Once the storage policy has been created, it can be used when provisioning new VMs as shown in Deploy VM using Storage Policy. Guidelines for using performance management capabilities with VASA Provider 7.2 are covered in Performance management with ONTAP tools 9.10 and later.

VM storage policy creation with ONTAP tools VASA Provider 9.10



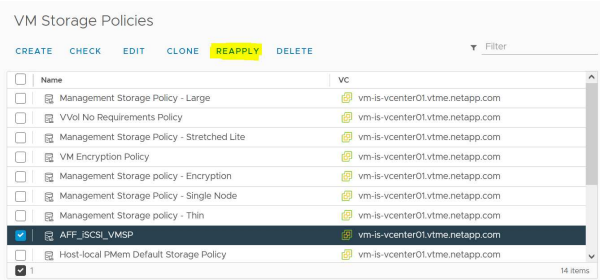
Performance management with ONTAP tools 9.10 and later

- ONTAP tools 9.10 uses its own balanced placement algorithm to place a new vVol in the best FlexVol volume within a vVols datastore. Placement is based on the specified SCP and matching FlexVol volumes. This makes sure that the datastore and backing storage can meet the specified performance requirements.
- Changing Performance capabilities such as Min and Max IOPS requires some attention to the specific configuration.
  - **Min and Max IOPS** may be specified in an SCP and used in a VM Policy.
    - Changing the IOPS in the SCP will not change QoS on the vVols until the VM Policy is edited, and then reapplied to the VMs that use it (see Storage Capabilities for ONTAP tools 9.10 and later). Or create a new SCP with the desired IOPS and change the policy to use it (and reapply to VMs). Generally it is recommended to simply define separate SCPs and VM storage policies for different tiers of service and simply change the VM storage policy on the VM.

    - AFF and FAS personalities have different IOPs settings. Both Min and Max are available on AFF. However non-AFF systems can only use Max IOPs settings.

- In some cases, a vVol may need to be migrated after a policy change (either manually, or automatically by VASA Provider and ONTAP):
  - Some changes require no migration (such as changing Max IOPS, which can be applied immediately to the VM as outlined above).
  - If the policy change cannot be supported by the current FlexVol volume that stores the vVol (for example, the platform does not support the encryption or tiering policy requested), you will need to manually migrate the VM in vCenter.
- ONTAP tools creates individual non-shared QoS policies with currently supported versions of ONTAP. Therefore, each individual VMDK will receive its own allocation of IOPs.

Reapplying VM Storage Policy



Protecting vVols

The following sections outline the procedures and best practices for using VMware vVols with ONTAP storage.

VASA Provider High Availability

The NetApp VASA Provider runs as part of the virtual appliance together with the vCenter plugin and REST API server (formerly known as the Virtual Storage Console [VSC]) and Storage Replication Adapter. If the VASA Provider is not available, VMs using vVols will continue to run. However, new vVols datastores cannot be created, and vVols cannot be created or bound by vSphere. This means that VMs using vVols cannot be powered on as vCenter will not be able to request creation of the swap vVol. And running VMs cannot use vMotion for migration to another host because the vVols cannot be bound to the new host.

VASA Provider 7.1 and later support new capabilities to make sure the services are available when needed. It includes new watchdog processes that monitor VASA Provider and integrated database services. If it detects a failure, it updates the log files and then restarts the services automatically.

Further protection must be configured by the vSphere administrator using the same availability features used to protect other mission critical VMs from faults in software, host hardware and network. No additional configuration is required on the virtual appliance to use these features; simply configure them using standard vSphere approaches. They have been tested and are supported by NetApp.

vSphere High Availability is easily configured to restart a VM on another host in the host cluster in the event of failure. vSphere Fault Tolerance provides higher availability by creating a secondary VM that is continuously replicated and can take over at any point. Additional information on these features is available in the ONTAP tools for VMware vSphere documentation (Configure high availability for ONTAP tools), as well as VMware vSphere documentation (look for vSphere Availability under ESXi and vCenter Server).

The ONTAP tools VASA Provider automatically backs up the vVols configuration in real time to managed ONTAP systems where the vVols information is stored within FlexVol volume metadata. In the event that the ONTAP tools appliance becomes unavailable for any reason, you can easily and quickly deploy a new one and import the configuration. Refer to this KB article for more information on VASA Provider recovery steps:

How to perform a VASA Provider Disaster Recovery - Resolution Guide

vVols Replication

Many ONTAP customers replicate their traditional datastores to secondary storage systems using NetApp SnapMirror, and then use the secondary system to recover individual VMs or an entire site in the event of a disaster. In most cases, customers use a software tool to manage this, such as a backup software product like the NetApp SnapCenter plugin for VMware vSphere or a disaster recovery solution such as VMware's Site Recovery Manager (together with the Storage Replication Adapter in ONTAP tools).

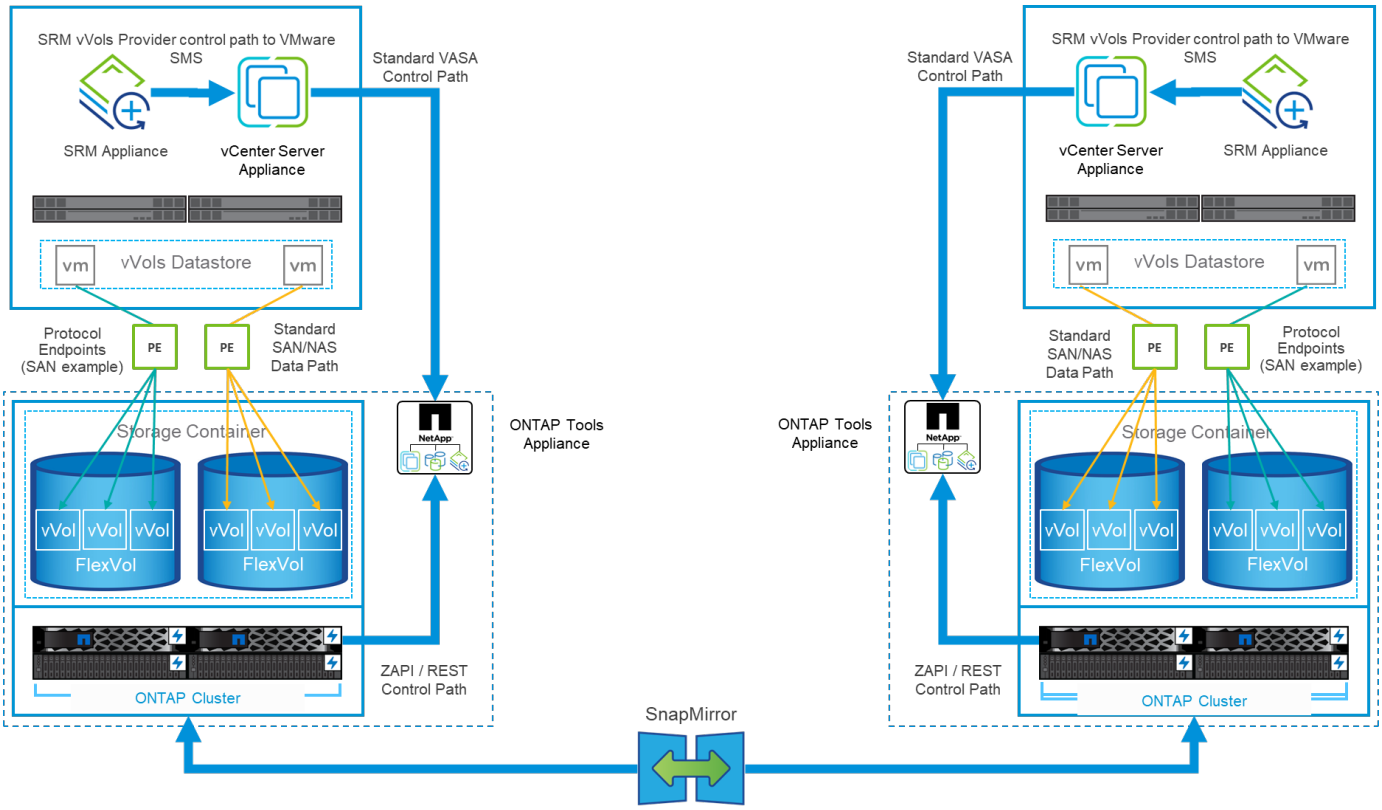
This requirement for a software tool is even more important to manage vVols replication. While some aspects can be managed by native capabilities (for example, VMware managed snapshots of vVols are offloaded to ONTAP which uses quick, efficient file or LUN clones), in general orchestration is needed to manage replication and recovery. Metadata about vVols is protected by ONTAP as well as the VASA Provider, but additional processing is needed to use them at a secondary site.

ONTAP tools 9.7.1 in conjunction with the VMware Site Recovery Manager (SRM) 8.3 release added support for disaster recovery and migration workflow orchestration taking advantage of NetApp SnapMirror technology.

In the initial release of SRM support with ONTAP tools 9.7.1 it was a requirement to pre-create FlexVols and enable SnapMirror protection before using them as backing volumes for a vVols datastore. Beginning in ONTAP tools 9.10 that process is no longer required. You can now add SnapMirror protection to existing backing volumes and update your VM storage policies to take advantage of policy based management with disaster recovery and migration orchestration and automation integrated with SRM.

Currently, VMware SRM is the only disaster recovery and migration automation solution for vVols supported by NetApp, and ONTAP tools will check for the existence of an SRM 8.3 or later server registered with your vCenter before allowing you to enable vVols replication, although it is possible to leverage the ONTAP tools REST APIs to create your own services.

vVols replication with SRM



MetroCluster Support

Although ONTAP tools is not capable of triggering a MetroCluster switchover, it does support NetApp MetroCluster systems for vVols backing volumes in a uniform vSphere Metro Storage Cluster (vMSC) configuration. Switchover of a MetroCluster system is handled in the normal manner.

While NetApp SnapMirror Business Continuity (SM-BC) can also be used as the basis for a vMSC configuration, it is not currently supported with vVols.

Refer to these guides for more information on NetApp MetroCluster:

TR-4689 MetroCluster IP Solution architecture and design

TR-4705 NetApp MetroCluster Solution architecture and design

VMware KB 2031038 VMware vSphere Support with NetApp MetroCluster

vVols Backup Overview

There are several approaches to protecting VMs such as using in-guest backup agents, attaching VM data files to a backup proxy, or using defined APIs such as VMware VADP. vVols may be protected using the same mechanisms and many NetApp partners support VM backups, including vVols.

As mentioned earlier, VMware vCenter managed snapshots are offloaded to space efficient and fast ONTAP file/LUN clones. These may be used for quick, manual backups, but are limited by vCenter to a maximum of 32 snapshots. You may use vCenter to take snapshots and revert as needed.

Beginning with SnapCenter Plugin for VMware vSphere (SCV) 4.6 when used in conjunction with ONTAP tools 9.10 and later adds support for crash consistent backup and recovery of vVols based VMs leveraging ONTAP FlexVol volume snapshots with support for SnapMirror and SnapVault replication. Up to 1023 snapshots are supported per volume. SCV can also store more snapshots with longer retention on secondary volumes using

SnapMirror with a mirror-vault policy.

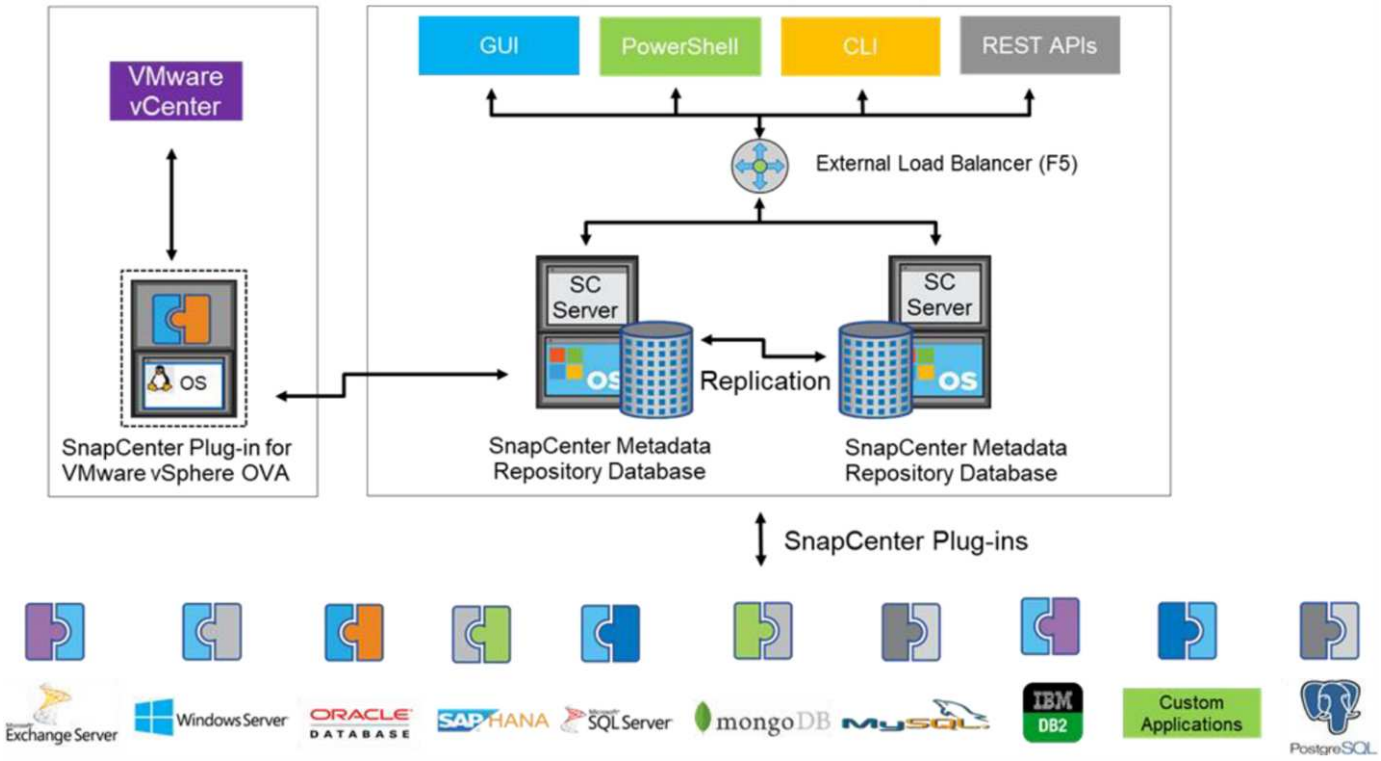
vSphere 8.0 support was introduced with SCV 4.7, which used an isolated local plugin architecture. vSphere 8.0U1 support was added to SCV 4.8 which fully transitioned to the new remote plugin architecture.

vVols Backup with SnapCenter plugin for VMware vSphere

With NetApp SnapCenter you can now create resource groups for vVols based on tags and/or folders to automatically take advantage of ONTAP's FlexVol based snapshots for vVols based VMs. This allows you to define backup and recovery services which will protect VMs automatically as they get dynamically provisioned within your environment.

SnapCenter plugin for VMware vSphere is deployed as a standalone appliance registered as a vCenter extension, managed through the vCenter UI or via REST APIs for backup and recovery service automation.

SnapCenter architecture



Since the other SnapCenter plugins don't yet support vVols at the time of this writing, we will focus on the standalone deployment model in this document.

Because SnapCenter uses ONTAP FlexVol snapshots there is no overhead placed on vSphere, nor is there any performance penalty as one might see with traditional VMs using vCenter managed snapshots. Furthermore, because SCV's functionality is exposed via REST APIs, it makes it easy to create automated workflows using tools like VMware Aria Automation, Ansible, Terraform, and virtually any other automation tool that is capable of using standard REST APIs.

For information on SnapCenter REST APIs, see Overview of REST APIs

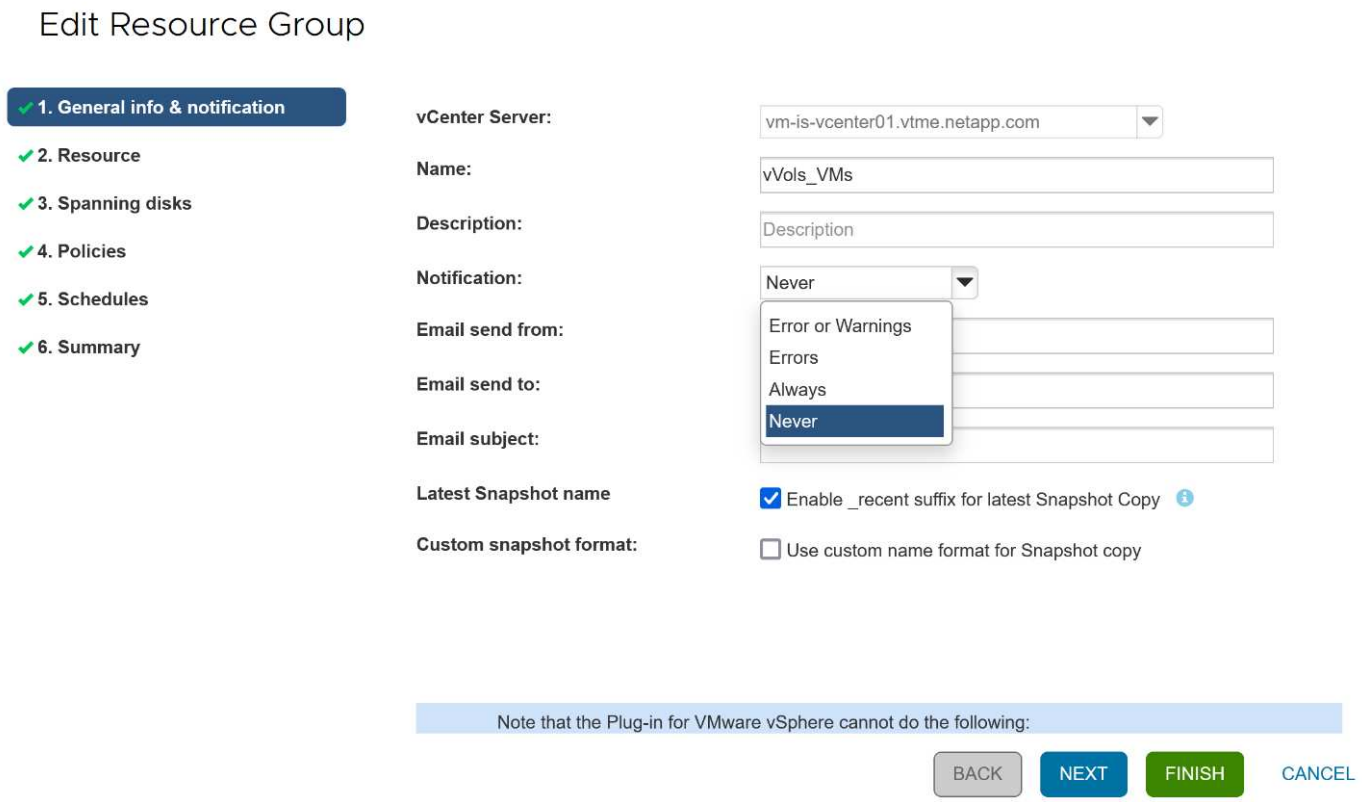
For information on SnapCenter Plug-in for VMware vSphere REST APIs, see SnapCenter Plug-in for VMware vSphere REST APIs

Best Practices

The following best practices can help you get the most out of your SnapCenter deployment.

• SCV supports both vCenter Server RBAC and ONTAP RBAC and includes predefined vCenter roles which are automatically created for you when the plugin is registered. You can read more about the supported types of RBAC here. ◦ Use the vCenter UI to assign least privileged account access using the predefined roles described here. ◦ If you use SCV with SnapCenter Server, you must assign the *SnapCenterAdmin* role. ◦ ONTAP RBAC refers to the user account used to add and manage the storage systems used by SCV. ONTAP RBAC doesn't apply to vVols based backups. Read more about ONTAP RBAC and SCV here.
• Replicate your backup datasets to a second system using SnapMirror for complete replicas of source volumes. As previously mentioned, you may also use mirror-vault policies for longer term retention of backup data independent of source volume snapshot retention settings. Both mechanisms are supported with vVols.
• Because SCV also requires ONTAP tools for VMware vSphere for vVols functionality, always check the NetApp Interoperability Matrix Tool (IMT) for specific version compatibility
• If you are using vVols replication with VMware SRM, be mindful of your policy RPO and backup schedule
• Design your backup policies with retention settings that meet your organizations defined recovery point objectives (RPOs)
• Configure notification settings on your resource groups to be notified of the status when backups run (see figure 10 below)

Resource group notification options



Get started with SCV using these documents

Learn about SnapCenter Plug-in for VMware vSphere

Deploy SnapCenter Plug-in for VMware vSphere

Troubleshooting

There are several troubleshooting resources available with additional information.

NetApp Support Site

In addition to a variety of Knowledgebase articles for NetApp virtualization products, the NetApp Support site also offers a convenient landing page for the ONTAP tools for VMware vSphere product. This portal provides links to articles, downloads, technical reports, and VMware Solutions Discussions on NetApp Community. It is available at:

NetApp Support Site

Additional solution documentation is available here:

NetApp Solutions for Virtualization

Product Troubleshooting

The various components of ONTAP tools, such as the vCenter plugin, VASA Provider, and Storage Replication Adapter are all documented together in the NetApp documents repository. However, each has a separate subsection of the Knowledge Base and may have specific troubleshooting procedures. These address the most common issues that may be encountered with the VASA Provider.

VASA Provider UI Problems

Occasionally the vCenter vSphere Web Client encounters problems with the Serenity components, causing the VASA Provider for ONTAP menu items not to display. See Resolving VASA Provider registration issues in the Deployment Guide, or this Knowledgebase article.

vVols Datastore Provisioning Fails

Occasionally vCenter services may time out when creating the vVols datastore. To correct it, restart the vmware-sps service, and re-mount the vVols datastore using the vCenter menus (Storage > New Datastore). This is covered under vVols datastore provisioning fails with vCenter Server 6.5 in the Administration Guide.

Upgrading Unified Appliance Fails to Mount ISO

Due to a bug in vCenter, the ISO used to upgrade the Unified Appliance from one release to the next may fail to mount. If the ISO is able to be attached to the appliance in vCenter, follow the process in this Knowledgebase article to resolve.

VMware Site Recovery Manager with ONTAP

VMware Site Recovery Manager with ONTAP

ONTAP has been a leading storage solution for VMware vSphere environments since its introduction into the modern datacenter in 2002, and it continues to add innovative capabilities to simplify management while reducing costs.

This document introduces the ONTAP solution for VMware Site Recovery Manager (SRM), VMware's industry leading disaster recovery (DR) software, including the latest product information and best practices to streamline deployment, reduce risk, and simplify ongoing management.

> This documentation replaces previously published technical report *TR-4900: VMware Site Recovery Manager with ONTAP*

Best practices supplement other documents such as guides and compatibility tools. They are developed based on lab testing and extensive field experience by NetApp engineers and customers. In some cases, recommended best practices might not be the right fit for your environment; however, they are generally the simplest solutions that meet the needs of the most customers.

This document is focused on capabilities in recent releases of ONTAP 9 when used in conjunction with ONTAP tools for VMware vSphere 9.12 (which includes the NetApp Storage Replication Adapter [SRA] and VASA Provider [VP]), as well as VMware Site Recovery Manager 8.7.

Why use ONTAP with SRM?

NetApp data management platforms powered by ONTAP software are some of the most widely adopted storage solutions for SRM. The reasons are plentiful: A secure, high performance, unified protocol (NAS and SAN together) data management platform that provides industry defining storage efficiency, multitenancy, quality of service controls, data protection with space-efficient snapshots and replication with SnapMirror. All leveraging native hybrid multi-cloud integration for the protection of VMware workloads and a plethora of automation and orchestration tools at your fingertips.

When you use SnapMirror for array-based replication, you take advantage of one of ONTAP's most proven and mature technologies. SnapMirror gives you the advantage of secure and highly efficient data transfers, copying

only changed file system blocks, not entire VMs or datastores. Even those blocks take advantage of space savings, such as deduplication, compression, and compaction. Modern ONTAP systems now use version-independent SnapMirror, allowing you flexibility in selecting your source and destination clusters. SnapMirror has truly become one of the most powerful tools available for disaster recovery.

Whether you are using traditional NFS, iSCSI, or Fibre Channel- attached datastores (now with support for vVols datastores), SRM provides a robust first party offering that leverages the best of ONTAP capabilities for disaster recovery or datacenter migration planning and orchestration.

How SRM leverages ONTAP 9

SRM leverages the advanced data management technologies of ONTAP systems by integrating with ONTAP tools for VMware vSphere, a virtual appliance that includes three primary components:

- The vCenter plug-in, formerly known as Virtual Storage Console (VSC), simplifies storage management and efficiency features, enhances availability, and reduces storage costs and operational overhead, whether you are using SAN or NAS. It uses best practices for provisioning datastores and optimizes ESXi host settings for NFS and block storage environments. For all these benefits, NetApp recommends this plug-in when using vSphere with systems running ONTAP software.
- The VASA Provider for ONTAP supports the VMware vStorage APIs for Storage Awareness (VASA) framework. VASA Provider connects vCenter Server with ONTAP to aid in provisioning and monitoring VM storage. It enables VMware Virtual Volumes (vVols) support and the management of storage capability profiles (including vVols replication capabilities) and individual VM vVols performance. It also provides alarms for monitoring capacity and compliance with the profiles. When used in conjunction with SRM, the VASA Provider for ONTAP enables support for vVols- based virtual machines without requiring the installation of an SRA adapter on the SRM server.
- The SRA is used together with SRM to manage the replication of VM data between production and disaster recovery sites for traditional VMFS and NFS datastores and also for the nondisruptive testing of DR replicas. It helps automate the tasks of discovery, recovery, and reprotection. It includes both an SRA server appliance and SRA adapters for the Windows SRM server and the SRM appliance.

After you have installed and configured the SRA adapters on the SRM server for protecting non-vVols datastores and/or enabled vVols replication in the VASA Provider settings, you can begin the task of configuring your vSphere environment for disaster recovery.

The SRA and VASA Provider deliver a command-and-control interface for the SRM server to manage the ONTAP FlexVols that contain your VMware Virtual Machines (VMs), as well as the SnapMirror replication protecting them.

Starting with SRM 8.3, a new SRM vVols Provider control path was introduced into the SRM server, allowing it to communicate with the vCenter server and, through it, to the VASA Provider without needing an SRA. This enabled the SRM server to leverage much deeper control over the ONTAP cluster than was possible before, because VASA provides a complete API for closely coupled integration.

SRM can test your DR plan nondisruptively using NetApp's proprietary FlexClone technology to make nearly instantaneous clones of your protected datastores at your DR site. SRM creates a sandbox to safely test so that your organization, and your customers, are protected in the event of a true disaster, giving you confidence in your organizations ability to execute a failover during a disaster.

In the event of a true disaster or even a planned migration, SRM allows you to send any last-minute changes to the dataset via a final SnapMirror update (if you choose to do so). It then breaks the mirror and mounts the datastore to your DR hosts. At that point, your VMs can be automatically powered up in any order according to your pre-planned strategy.

SRM with ONTAP and other use cases: hybrid cloud and migration

Integrating your SRM deployment with ONTAP advanced data management capabilities allows for vastly improved scale and performance when compared with local storage options. But more than that, it brings the flexibility of the hybrid cloud. The hybrid cloud enables you to save money by tiering unused data blocks from your high-performance array to your preferred hyperscaler using FabricPool, which could be an on-premises S3 store such as NetApp StorageGRID. You can also use SnapMirror for edge-based systems with software-defined ONTAP Select or cloud-based DR using Cloud Volumes ONTAP (CVO) or NetApp Private Storage in Equinix for Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) to create a fully integrated storage, networking, and compute- services stack in the cloud.

You could then perform test failover inside a cloud service provider's datacenter with near-zero storage footprint thanks to FlexClone. Protecting your organization can now cost less than ever before.

SRM can also be used to execute planned migrations by leveraging SnapMirror to efficiently transfer your VMs from one datacenter to another or even within the same datacenter, whether your own, or via any number of NetApp partner service providers.

Deployment best practices

The following sections outline the deployment best practices with ONTAP and VMware SRM.

SVM layout and segmentation for SMT

With ONTAP, the concept of the storage virtual machine (SVM) provides strict segmentation in secure multitenant environments. SVM users on one SVM cannot access or manage resources from another. In this way, you can leverage ONTAP technology by creating separate SVMs for different business units who manage their own SRM workflows on the same cluster for greater overall storage efficiency.

Consider managing ONTAP using SVM-scoped accounts and SVM management LIFs to not only improve security controls, but also improve performance. Performance is inherently greater when using SVM-scoped connections because the SRA is not required to process all the resources in an entire cluster, including physical resources. Instead, it only needs to understand the logical assets that are abstracted to the particular SVM.

When using NAS protocols only (no SAN access), you can even leverage the new NAS optimized mode by setting the following parameter (note that the name is such because SRA and VASA use the same backend services in the appliance):

1. Log into the control panel at `https://<IP address>:9083` and click Web based CLI interface.
2. Run the command `vp updateconfig -key=enable.qtree.discovery -value=true`.
3. Run the command `vp updateconfig -key=enable.optimised.sra -value=true`.
4. Run the command `vp reloadconfig`.

Deploy ONTAP tools and considerations for vVols

If you intend to use SRM with vVols, you must manage the storage using cluster- scoped credentials and a cluster management LIF. This is because the VASA Provider must understand the underlying physical architecture to satisfy the policy requires for VM storage policies. For example, if you have a policy that requires all- flash storage, the VASA Provider must be able to see which systems are all flash.

Another deployment best practice is to never store your ONTAP tools appliance on a vVols datastore that it is managing. This could lead to a situation whereby you cannot power on the VASA Provider because you cannot create the swap vVol for the appliance because the appliance is offline.

Best practices for managing ONTAP 9 systems

As previously mentioned, you can manage ONTAP clusters using either cluster or SVM scoped credentials and management LIFs. For optimum performance, you may want to consider using SVM- scoped credentials whenever you aren't using vVols. However, in doing so, you should be aware of some requirements, and that you do lose some functionality.

- The default vsadmin SVM account does not have the required access level to perform ONTAP tools tasks. Therefore, you need to create a new SVM account.
- If you are using ONTAP 9.8 or later, NetApp recommends creating an RBAC least privileged user account using ONTAP System Manager's users menu together with the JSON file available on your ONTAP tools appliance at `https://<IP address>:9083/vsc/config/`. Use your administrator password to download the JSON file. This can be used for SVM or cluster scoped accounts.

  If you are using ONTAP 9.6 or earlier, you should use the RBAC User Creator (RUC) tool available in the NetApp Support Site Toolchest.

- Because the vCenter UI plugin, VASA Provider, and SRA server are all fully integrated services, you must add storage to the SRA adapter in SRM the same way you add storage in the vCenter UI for ONTAP tools. Otherwise, the SRA server might not recognize the requests being sent from SRM via the SRA adapter.
- NFS path checking is not performed when using SVM-scoped credentials. This is because the physical location is logically abstracted from the SVM. This is not a cause for concern though, as modern ONTAP systems no longer suffer any noticeable performance decline when using indirect paths.
- Aggregate space savings due to storage efficiency might not be reported.
- Where supported, load-sharing mirrors cannot be updated.
- EMS logging might not be performed on ONTAP systems managed with SVM scoped credentials.

Operational best practices

The following sections outline the operational best practices for VMware SRM and ONTAP storage.

Datastores and protocols

- If possible, always use ONTAP tools to provision datastores and volumes. This makes sure that volumes, junction paths, LUNs, igroups, export policies, and other settings are configured in a compatible manner.
- SRM supports iSCSI, Fibre Channel, and NFS version 3 with ONTAP 9 when using array-based replication through SRA. SRM does not support array-based replication for NFS version 4.1 with either traditional or vVols datastores.
- To confirm connectivity, always verify that you can mount and unmount a new test datastore at the DR site from the destination ONTAP cluster. Test each protocol you intend to use for datastore connectivity. A best practice is to use ONTAP tools to create your test datastore, since it is doing all the datastore automation as directed by SRM.
- SAN protocols should be homogeneous for each site. You can mix NFS and SAN, but the SAN protocols should not be mixed within a site. For example, you can use FCP in site A, and iSCSI in site B. You should not use both FCP and iSCSI at site A. The reason for this is that the SRA does not create mixed igroups at

the recovery site and SRM does not filter the initiator list given to the SRA.

- Previous guides advised creating LIF to data locality. That is to say, always mount a datastore using a LIF located on the node that physically owns the volume. That is no longer a requirement in modern versions of ONTAP 9. Whenever possible, and if given cluster-scoped credentials, ONTAP tools will still choose to load balance across LIFs local to the data, but it is not a requirement for high availability or performance.
- ONTAP 9 can be configured to automatically remove snapshots to preserve uptime in the event of an out-of-space condition when autosize is not able to supply sufficient emergency capacity. The default setting for this capability does not automatically delete the snapshots that are created by SnapMirror. If SnapMirror snapshots are deleted, then the NetApp SRA cannot reverse and resynchronize replication for the affected volume. To prevent ONTAP from deleting SnapMirror snapshots, configure the Snapshot autodelete capability to try.

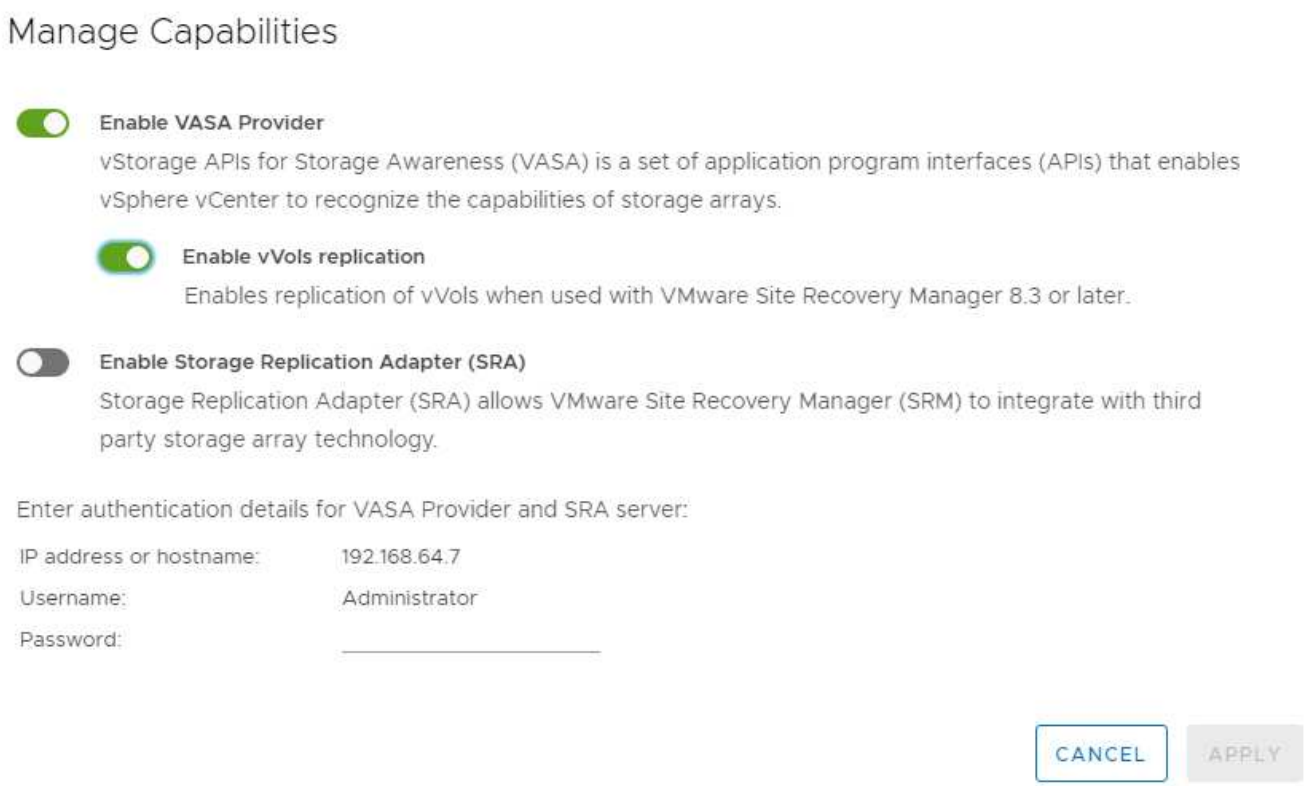
```
snap autodelete modify -volume -commitment try
```

- Volume autosize should be set to `grow` for volumes containing SAN datastores and `grow_shrink` for NFS datastores. Learn more about configuring volumes to automatically grow or shrink.
- SRM performs best when the number of datastores and thus protection groups is minimized in your recovery plans. Therefore you should consider optimizing for VM density in SRM-protected environments where RTO is of key importance.
- Use Distributed Resource Scheduler (DRS) to help balance the load on your protected and recovery ESXi clusters. Remember that if you plan to failback, when you run a reprotect the previously protected clusters will become the new recovery clusters. DRS will help balance placement going in both directions.
- Where possible, avoid using IP customization with SRM as this can increase your RTO.

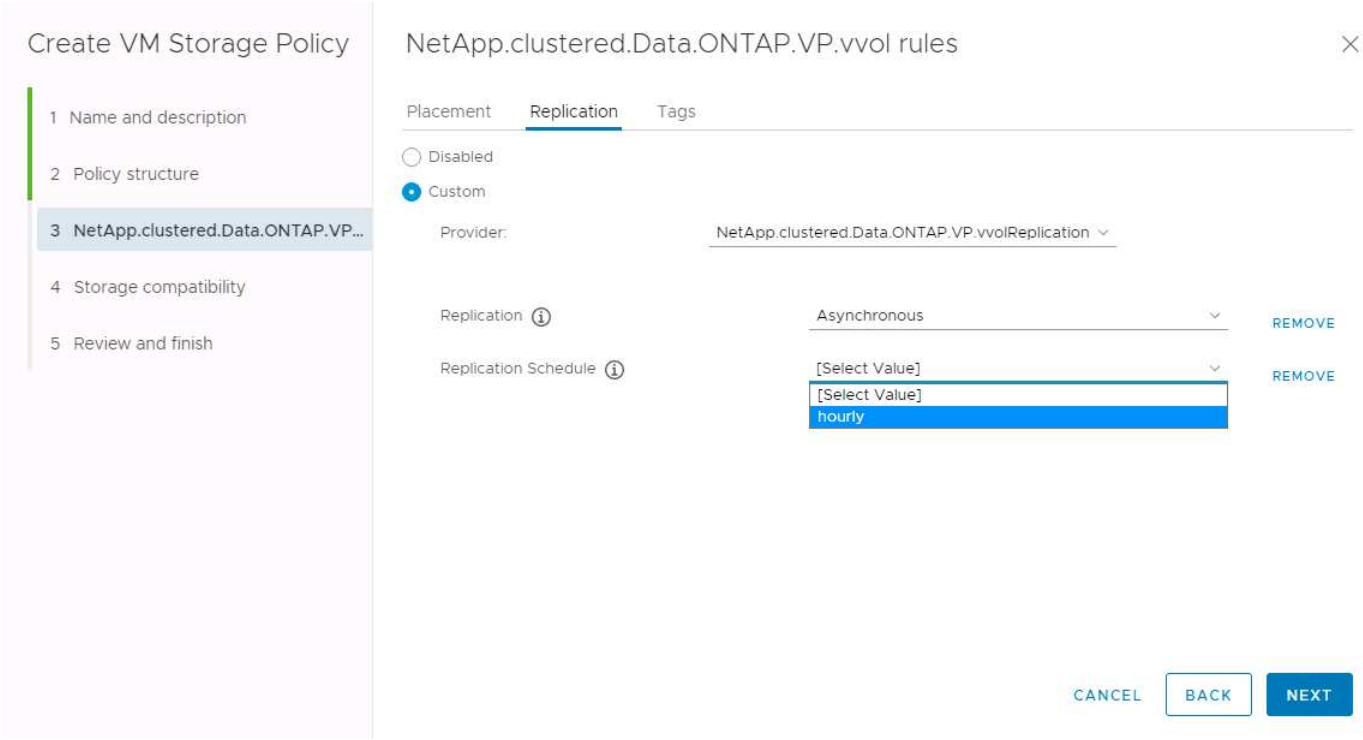
Storage Policy Based Management (SPBM) and vVols

Starting with SRM 8.3, the protection of VMs using vVols datastores is supported. SnapMirror schedules are exposed to VM storage policies by the VASA Provider when vVols replication is enabled in the ONTAP tools settings menu, as shown in the following screenshots.

The following example shows the enablement of vVols replication.

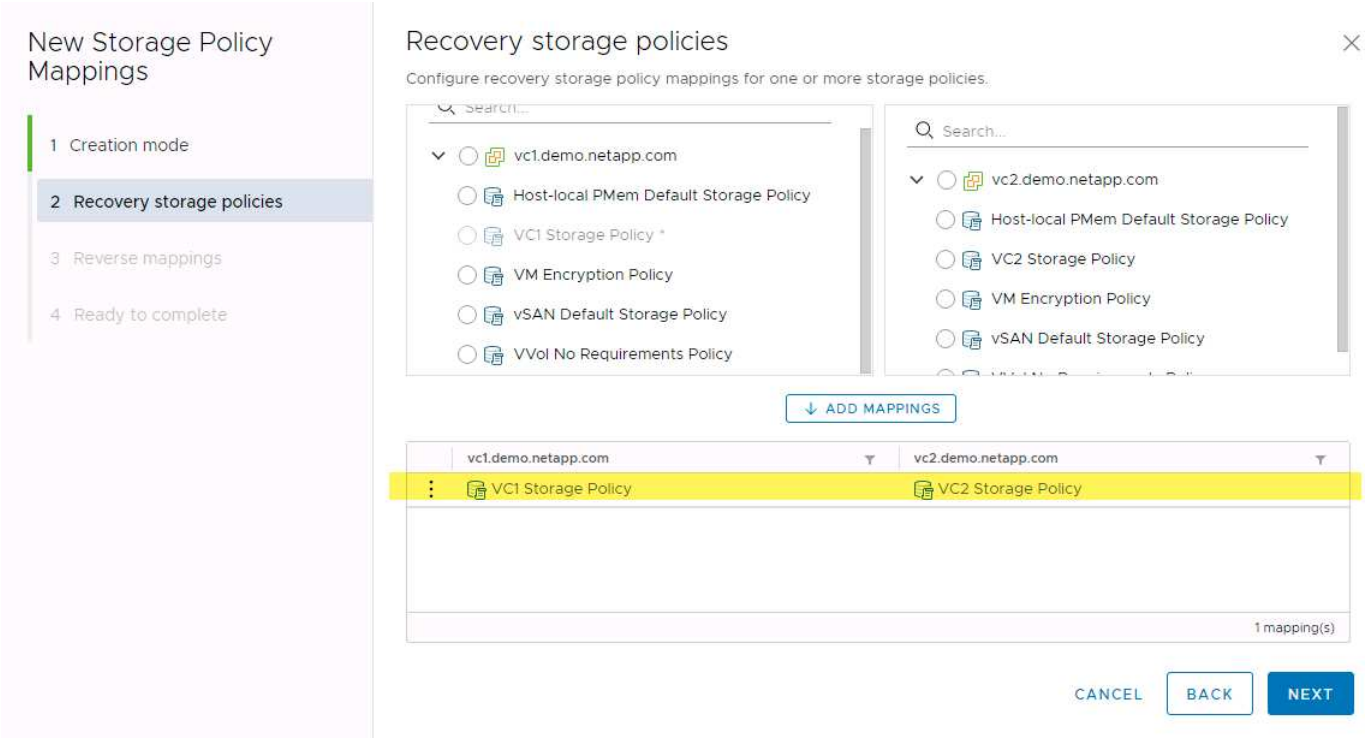


The following screenshot provides an example of SnapMirror schedules displayed in the Create VM Storage Policy wizard.



The ONTAP VASA Provider supports failover to dissimilar storage. For example, the system can fail over from ONTAP Select at an edge location to an AFF system in the core datacenter. Regardless of storage similarity, you must always configure storage policy mappings and reverse mappings for replication-enabled VM storage policies to make sure that services provided at the recovery site meet expectations and requirements. The

following screenshot highlights a sample policy mapping.



Create replicated volumes for vVols datastores

Unlike previous vVols datastores, replicated vVols datastores must be created from the start with replication enabled, and they must use volumes that were pre-created on the ONTAP systems with SnapMirror relationships. This requires pre-configuring things like cluster peering and SVM peering. These activities should be performed by your ONTAP administrator because this facilitates a strict separation of responsibilities between those who manage the ONTAP systems across multiple sites and those who are primarily responsible for vSphere operations.

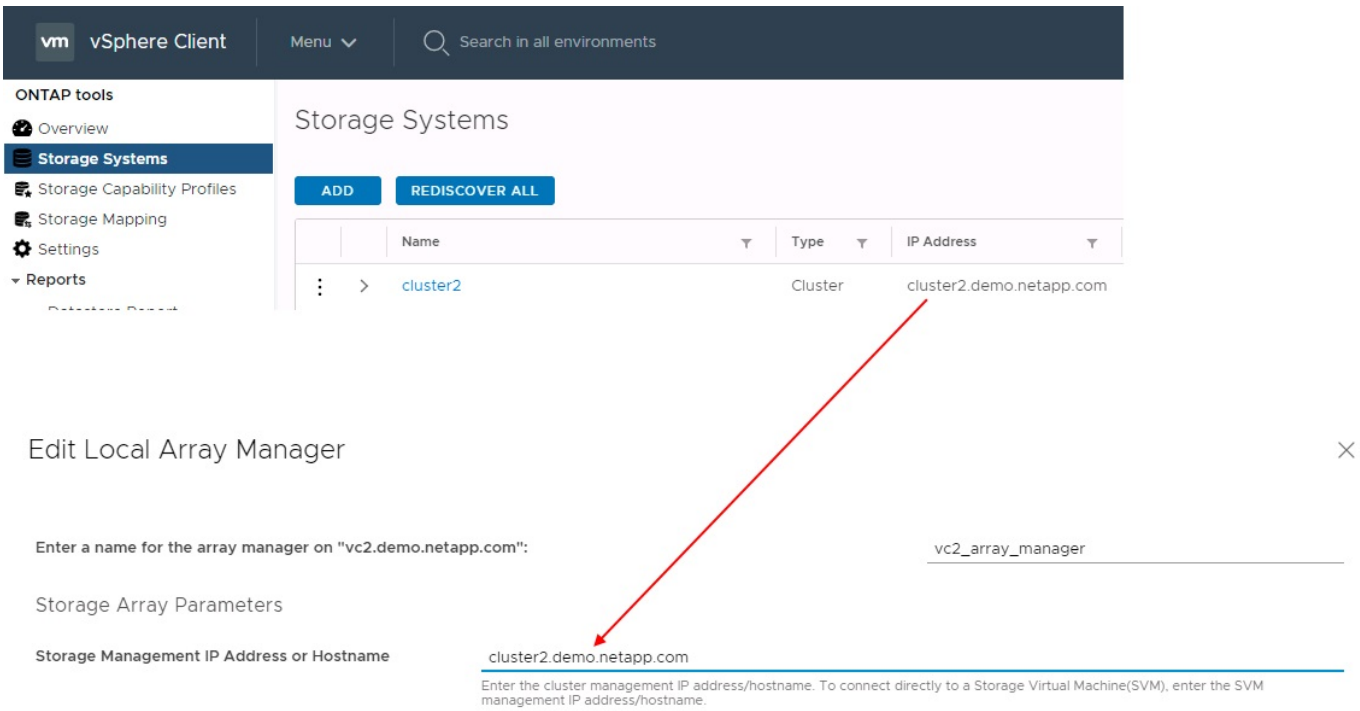
This does come with a new requirement on behalf of the vSphere administrator. Because volumes are being created outside the scope of ONTAP tools, it is unaware of the changes your ONTAP administrator has made until the regularly scheduled rediscovery period. For that reason, it is a best practice to always run rediscovery whenever you create a volume or SnapMirror relationship to be used with vVols. Simply right-click on the host or cluster and select ONTAP tools > Update Host and Storage Data, as shown in the following screenshot.

One caution should be taken when it comes to vVols and SRM. Never mix protected and unprotected VMs in the same vVols datastore. The reason for this is that when you use SRM to failover to your DR site, only those VMs that are part of the protection group are brought online in DR. Therefore, when you reprotect (reverse the SnapMirror from DR back to production again), you may overwrite the VMs that were not failed over and could contain valuable data.

About array pairs

An array manager is created for each array pair. With SRM and ONTAP tools, each array pairing is done with the scope of an SVM, even if you are using cluster credentials. This allows you to segment DR workflows between tenants based on which SVMs they have been assigned to manage. You can create multiple array managers for a given cluster, and they can be asymmetric. You can fan out or fan in between different ONTAP 9 clusters. For example, you can have SVM-A and SVM-B on Cluster-1 replicating to SVM-C on Cluster-2, SVM-D on Cluster-3, or vice-versa.

When configuring array pairs in SRM, you should always add them in SRM the same way as you added them to ONTAP Tools, meaning, they must use the same username, password, and management LIF. This requirement ensures that SRA communicates properly with the array. The following screenshot illustrates how a cluster might appear in ONTAP Tools and how it might be added to an array manager.



About replication groups

Replication groups contain logical collections of virtual machines that are recovered together. The ONTAP tools VASA Provider automatically creates replication groups for you. Because ONTAP SnapMirror replication occurs at the volume level, all VMs in a volume are in the same replication group.

There are several factors to consider with replication groups and how you distribute VMs across FlexVol volumes. Grouping similar VMs in the same volume can increase storage efficiency with older ONTAP systems that lack aggregate-level deduplication, but grouping increases the size of the volume and reduces volume I/O concurrency. The best balance of performance and storage efficiency can be achieved in modern ONTAP systems by distributing VMs across FlexVol volumes in the same aggregate, thereby leveraging aggregate-level deduplication and gaining greater I/O parallelization across multiple volumes. You can recover VMs in the volumes together because a protection group (discussed below) can contain multiple replication groups. The downside to this layout is that blocks might be transmitted over the wire multiple times because volume SnapMirror doesn't take aggregate deduplication into account.

One final consideration for replication groups is that each one is by its nature a logical consistency group (not to be confused with SRM consistency groups). This is because all VMs in the volume are transferred together using the same snapshot. So if you have VMs that must be consistent with each other, consider storing them in

the same FlexVol.

About protection groups

Protection groups define VMs and datastores in groups that are recovered together from the protected site. The protected site is where the VMs that are configured in a protection group exist during normal steady-state operations. It is important to note that even though SRM might display multiple array managers for a protection group, a protection group cannot span multiple array managers. For this reason, you should not span VM files across datastores on different SVMs.

About recovery plans

Recovery plans define which protection groups are recovered in the same process. Multiple protection groups can be configured in the same recovery plan. Also, to enable more options for the execution of recovery plans, a single protection group can be included in multiple recovery plans.

Recovery plans allow SRM administrators to define recovery workflows by assigning VMs to a priority group from 1 (highest) to 5 (lowest), with 3 (medium) being the default. Within a priority group, VMs can be configured for dependencies.

For example, your company could have a tier-1 business-critical application that relies on a Microsoft SQL server for its database. So, you decide to place your VMs in priority group 1. Within priority group 1, you begin planning the order to bring up services. You probably want your Microsoft Windows domain controller to boot up before your Microsoft SQL server, which would need to be online before your application server, and so on. You would add all these VMs to the priority group and then set the dependencies because dependencies only apply within a given priority group.

NetApp strongly recommends working with your application teams to understand the order of operations required in a failover scenario and to construct your recovery plans accordingly.

Test failover

As a best practice, always perform a test failover whenever a change is made to the configuration of a protected VM storage. This ensures that, in the event of a disaster, you can trust that Site Recovery Manager can restore services within the expected RTO target.

NetApp also recommends confirming in-guest application functionality occasionally, especially after reconfiguring VM storage.

When a test recovery operation is performed, a private test bubble network is created on the ESXi host for the VMs. However, this network is not automatically connected to any physical network adapters and therefore does not provide connectivity between the ESXi hosts. To allow communication among VMs that are running on different ESXi hosts during DR testing, a physical private network is created between the ESXi hosts at the DR site. To verify that the test network is private, the test bubble network can be separated physically or by using VLANs or VLAN tagging. This network must be segregated from the production network because as the VMs are recovered, they cannot be placed on the production network with IP addresses that could conflict with actual production systems. When a recovery plan is created in SRM, the test network that was created can be selected as the private network to connect the VMs to during the test.

After the test has been validated and is no longer required, perform a cleanup operation. Running cleanup returns the protected VMs to their initial state and resets the recovery plan to the Ready state.

Failover considerations

There are several other considerations when it comes to failing over a site in addition to the order of operations mentioned in this guide.

One issue you might have to contend with is networking differences between sites. Some environments might be able to use the same network IP addresses at both the primary site and the DR site. This ability is referred to as a stretched virtual LAN (VLAN) or stretched network setup. Other environments might have a requirement to use different network IP addresses (for example, in different VLANs) at the primary site relative to the DR site.

VMware offers several ways to solve this problem. For one, network virtualization technologies like VMware NSX-T Data Center abstract the entire networking stack from layers 2 through 7 from the operating environment, allowing for more portable solutions. Learn more about NSX-T options with SRM.

SRM also gives you the ability to change the network configuration of a VM as it is recovered. This reconfiguration includes settings such as IP addresses, gateway addresses, and DNS server settings. Different network settings, which are applied to individual VMs as they are recovered, can be specified in the property's settings of a VM in the recovery plan.

To configure SRM to apply different network settings to multiple VMs without having to edit the properties of each one in the recovery plan, VMware provides a tool called the dr-ip-customizer. Learn how to use this utility, refer to VMware's documentation.

Reprotect

After a recovery, the recovery site becomes the new production site. Because the recovery operation broke the SnapMirror replication, the new production site is not protected from any future disaster. A best practice is to protect the new production site to another site immediately after a recovery. If the original production site is operational, the VMware administrator can use the original production site as a new recovery site to protect the new production site, effectively reversing the direction of protection. Reprotection is available only in non-catastrophic failures. Therefore, the original vCenter Servers, ESXi servers, SRM servers, and corresponding databases must be eventually recoverable. If they are not available, a new protection group and a new recovery plan must be created.

Failback

A failback operation is fundamentally a failover in a different direction than before. As a best practice, you verify that the original site is back to acceptable levels of functionality before attempting to failback, or, in other words, failover to the original site. If the original site is still compromised, you should delay failback until the failure is sufficiently remediated.

Another failback best practice is to always perform a test failover after completing reprotect and before doing your final failback. This verifies that the systems in place at the original site can complete the operation.

Reprotecting the original site

After failback, you should confirm with all stakeholders that their services have been returned to normal before running reprotect again,

Running reprotect after failback essentially puts the environment back in the state it was in at the beginning, with SnapMirror replication again running from the production site to the recovery site.

Replication topologies

In ONTAP 9, the physical components of a cluster are visible to cluster administrators, but they are not directly visible to the applications and hosts that use the cluster. The physical components provide a pool of shared resources from which the logical cluster resources are constructed. Applications and hosts access data only through SVMs that contain volumes and LIFs.

Each NetApp SVM is treated as an array in VMware vCenter Site Recovery Manager. SRM supports certain array-to-array (or SVM-to-SVM) replication layouts.

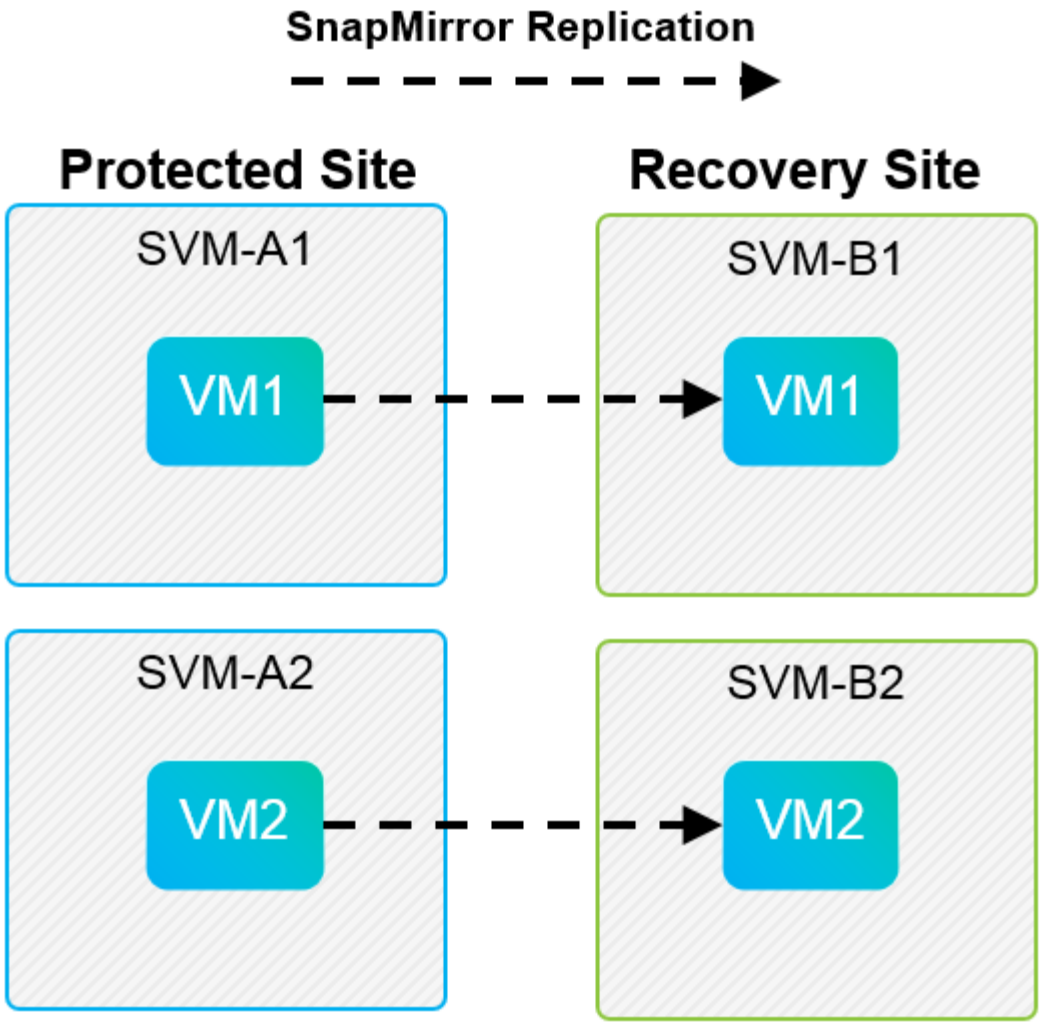
A single VM cannot own data—Virtual Machine Disk (VMDK) or RDM—on more than one SRM array for the following reasons:

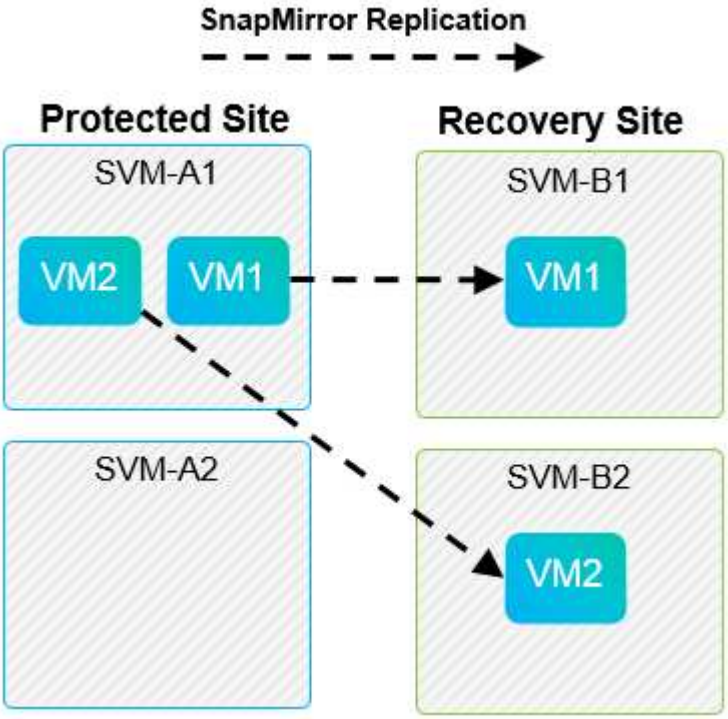
- SRM sees only the SVM, not an individual physical controller.
- An SVM can control LUNs and volumes that span multiple nodes in a cluster.

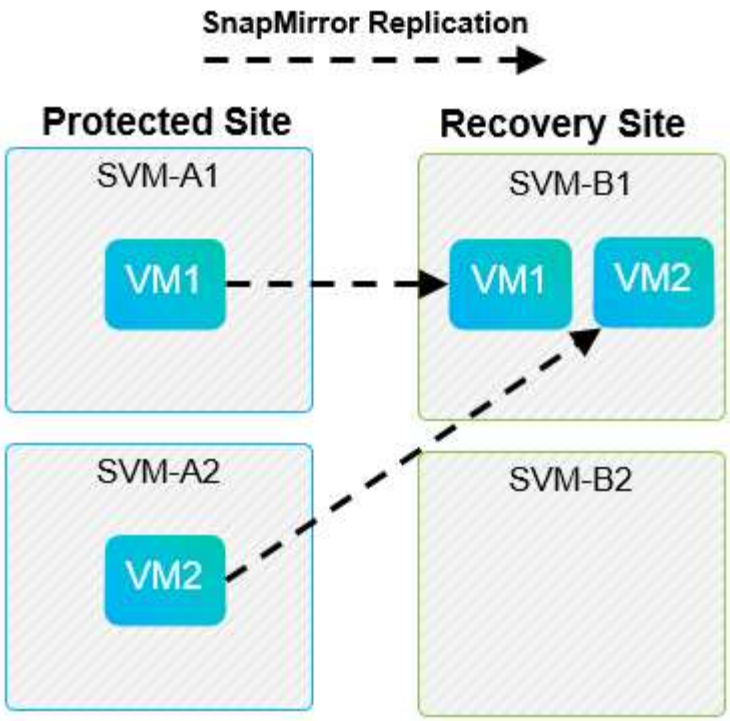
Best Practice
To determine supportability, keep this rule in mind: to protect a VM by using SRM and the NetApp SRA, all parts of the VM must exist on only one SVM. This rule applies at both the protected site and the recovery site.

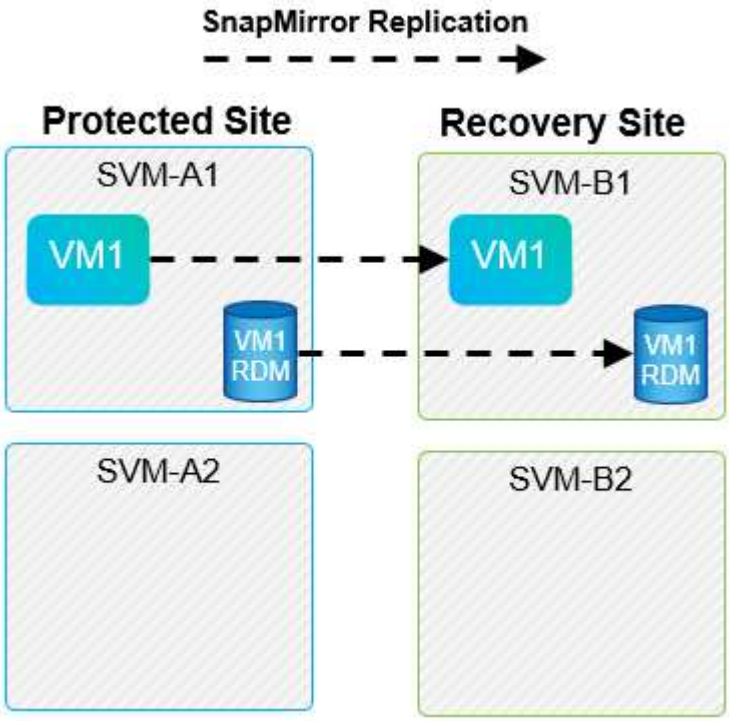
Supported SnapMirror layouts

The following figures show the SnapMirror relationship layout scenarios that SRM and SRA support. Each VM in the replicated volumes owns data on only one SRM array (SVM) at each site.


SnapMirror Replication
Protected Site
Recovery Site
SVM-A1
VM1
SVM-B1
VM1
SVM-A2
VM2
SVM-B2
VM2

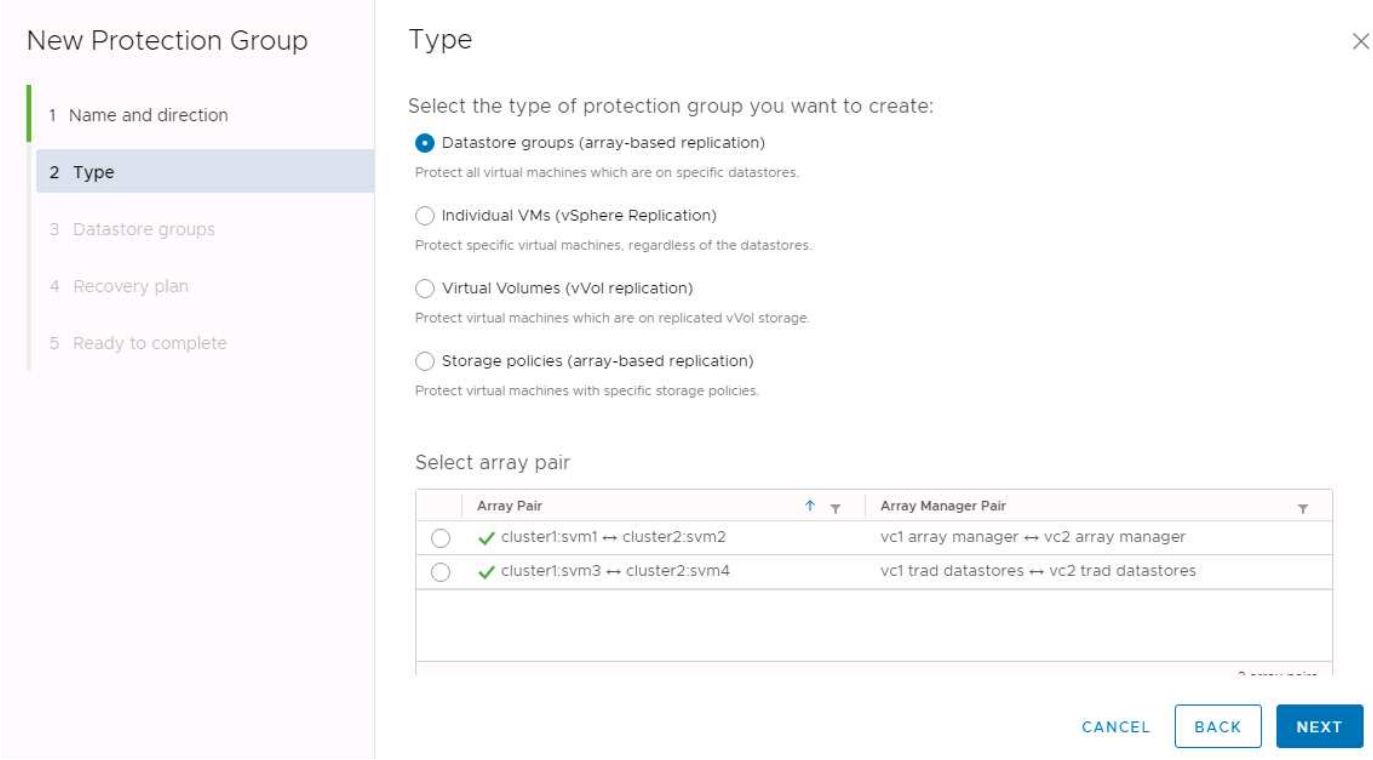

SnapMirror Replication
Protected Site
Recovery Site
SVM-A1
VM2
VM1
SVM-B1
VM1
SVM-A2
SVM-B2
VM2





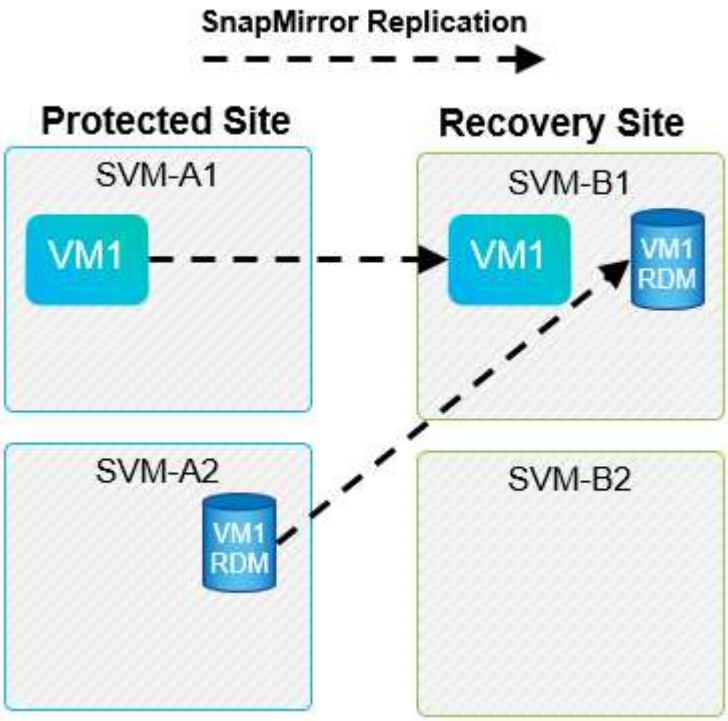
Supported Array Manager layouts

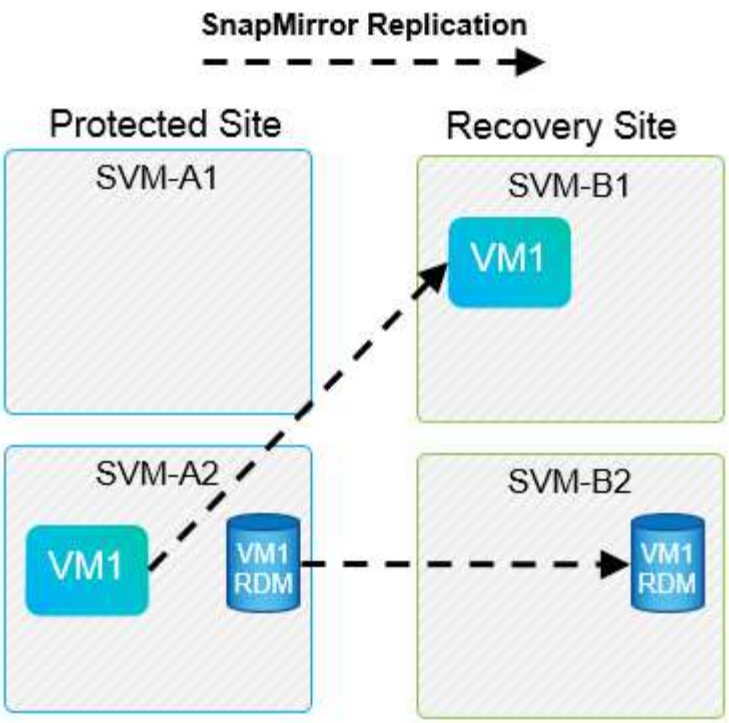
When you use array-based replication (ABR) in SRM, protection groups are isolated to a single array pair, as shown in the following screenshot. In this scenario, `SVM1` and `SVM2` are peered with `SVM3` and `SVM4` at the recovery site. However, you can select only one of the two array pairs when you create a protection group.



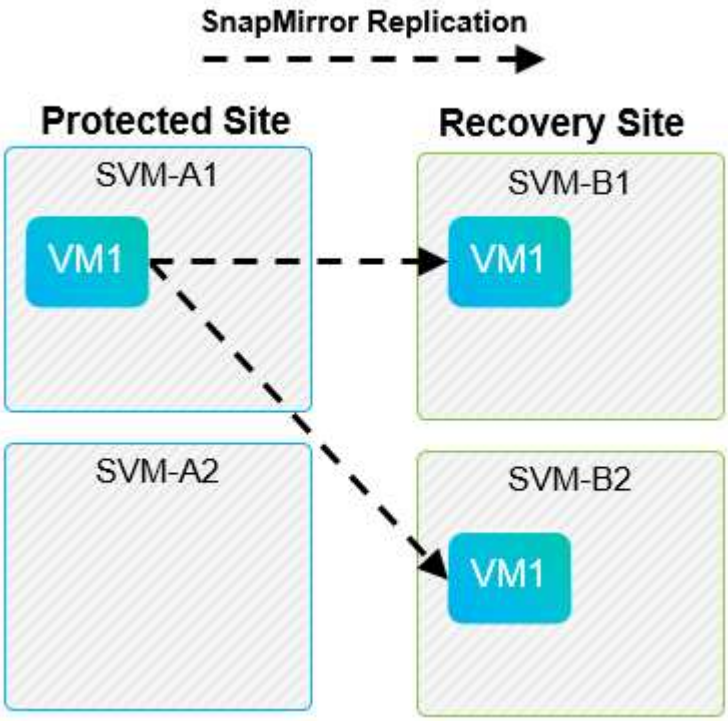
Unsupported layouts

Unsupported configurations have data (VMDK or RDM) on multiple SVMs that is owned by an individual VM. In the examples shown in the following figures, `VM1` cannot be configured for protection with SRM because `VM1` has data on two SVMs.



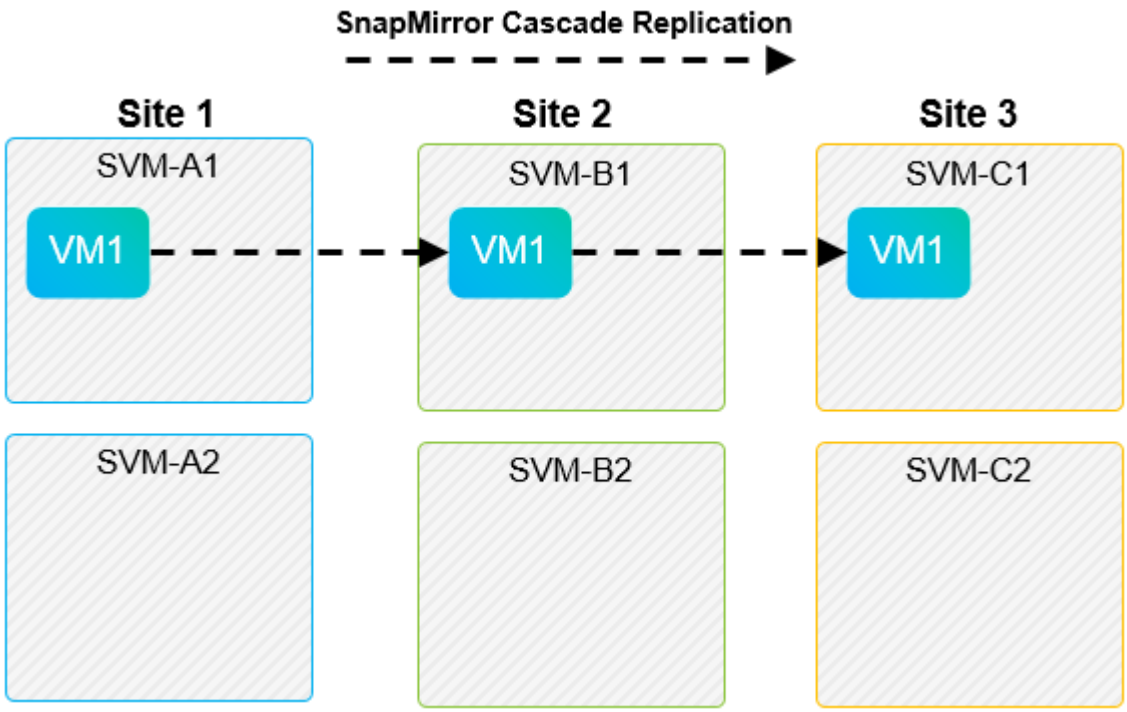


Any replication relationship in which an individual NetApp volume is replicated from one source SVM to multiple destinations in the same SVM or in different SVMs is referred to as SnapMirror fan-out. Fan-out is not supported with SRM. In the example shown in the following figure, `VM1` cannot be configured for protection in SRM because it is replicated with SnapMirror to two different locations.



SnapMirror cascade

SRM does not support cascading of SnapMirror relationships, in which a source volume is replicated to a destination volume and that destination volume is also replicated with SnapMirror to another destination volume. In the scenario shown in the following figure, SRM cannot be used for failover between any sites.



SnapMirror and SnapVault

NetApp SnapVault software enables disk-based backup of enterprise data between NetApp storage systems. SnapVault and SnapMirror can coexist in the same environment; however, SRM supports the failover of only the SnapMirror relationships.

> The NetApp SRA supports the `mirror-vault` policy type.

SnapVault was rebuilt from the ground up for ONTAP 8.2. Although former Data ONTAP 7-Mode users should find similarities, major enhancements have been made in this version of SnapVault. One major advance is the ability to preserve storage efficiencies on primary data during SnapVault transfers.

An important architectural change is that SnapVault in ONTAP 9 replicates at the volume level as opposed to at the qtree level, as is the case in 7-Mode SnapVault. This setup means that the source of a SnapVault relationship must be a volume, and that volume must replicate to its own volume on the SnapVault secondary system.

In an environment in which SnapVault is used, specifically named snapshots are created on the primary storage system. Depending on the configuration implemented, the named snapshots can be created on the primary system by a SnapVault schedule or by an application such as NetApp Active IQ Unified Manager. The named snapshots that are created on the primary system are then replicated to the SnapMirror destination, and from there they are vaulted to the SnapVault destination.

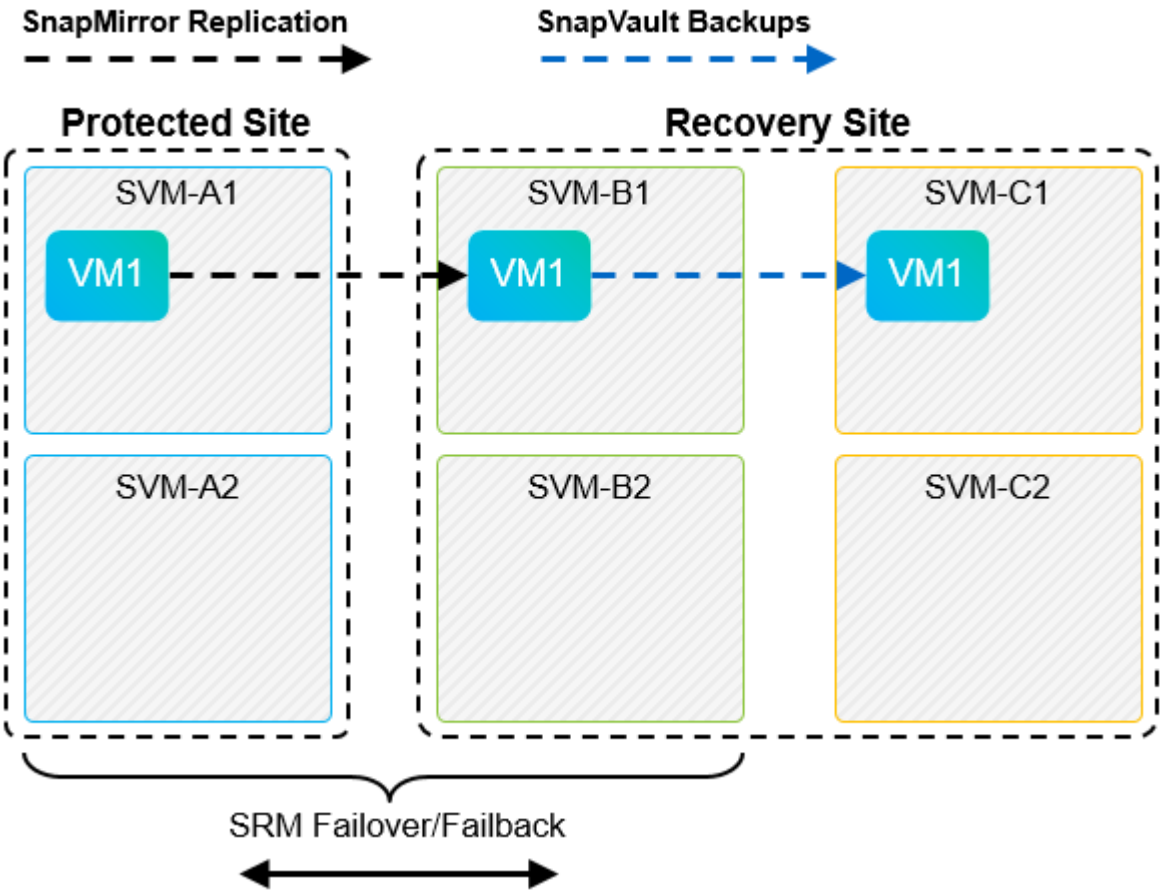
A source volume can be created in a cascade configuration in which a volume is replicated to a SnapMirror destination in the DR site, and from there it is vaulted to a SnapVault destination. A source volume can also be created in a fan-out relationship in which one destination is a SnapMirror destination and the other destination is a SnapVault destination. However, SRA does not automatically reconfigure the SnapVault relationship to use the SnapMirror destination volume as the source for the vault when SRM failover or replication reversal occurs.

For the latest information about SnapMirror and SnapVault for ONTAP 9, see TR-4015 SnapMirror Configuration Best Practice Guide for ONTAP 9.

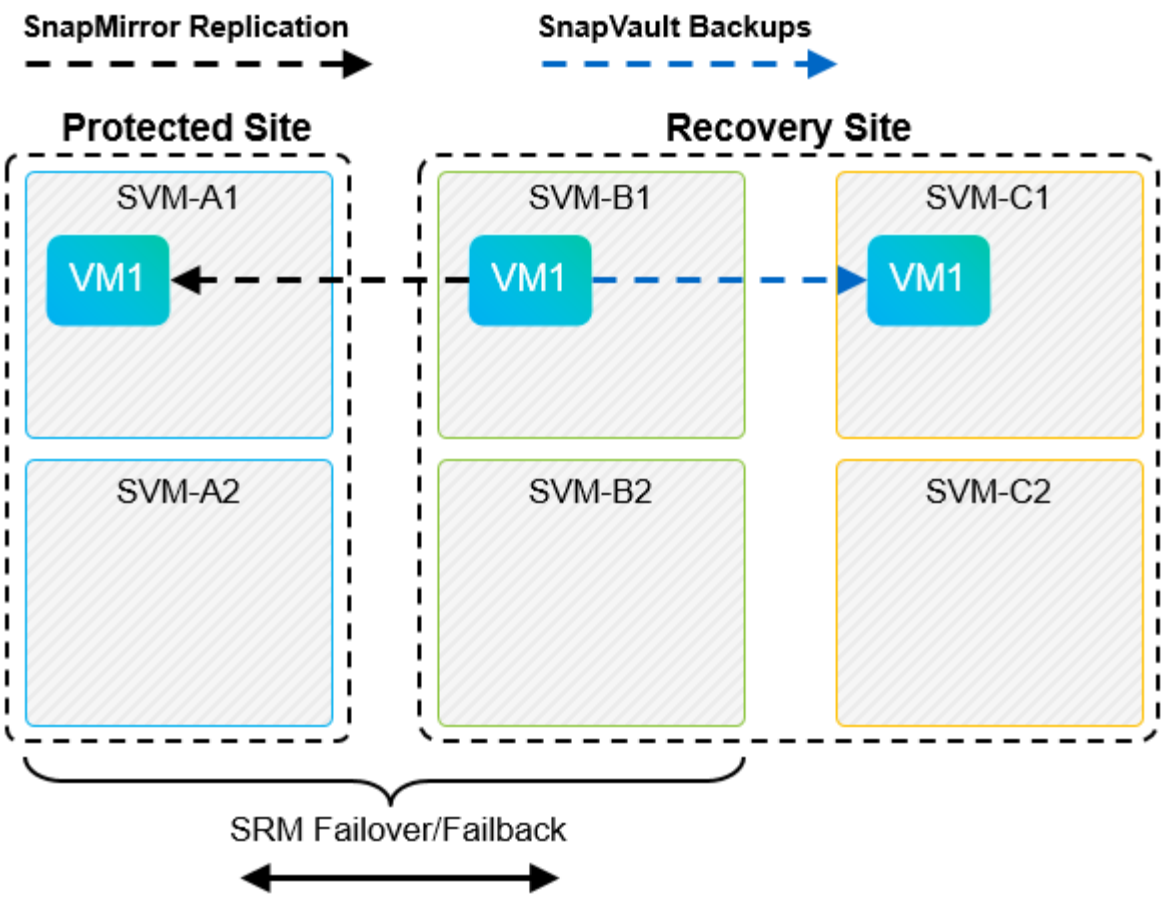
Best Practice
If SnapVault and SRM are used in the same environment, NetApp recommends using a SnapMirror to SnapVault cascade configuration in which SnapVault backups are normally performed from the SnapMirror destination at the DR site. In the event of a disaster, this configuration makes the primary site inaccessible. Keeping the SnapVault destination at the recovery site allows SnapVault backups to be reconfigured after failover so that SnapVault backups can continue while operating at the recovery site.

In a VMware environment, each datastore has a universal unique identifier (UUID), and each VM has a unique managed object ID (MOID). These IDs are not maintained by SRM during failover or failback. Because datastore UUIDs and VM MOIDs are not maintained during failover by SRM, any applications that depend on these IDs must be reconfigured after SRM failover. An example application is NetApp Active IQ Unified Manager, which coordinates SnapVault replication with the vSphere environment.

The following figure depicts a SnapMirror to SnapVault cascade configuration. If the SnapVault destination is at the DR site or at a tertiary site that is not affected by an outage at the primary site, the environment can be reconfigured to allow backups to continue after failover.



The following figure depicts the configuration after SRM has been used to reverse SnapMirror replication back to the primary site. The environment has also been reconfigured such that SnapVault backups are occurring from what is now the SnapMirror source. This setup is a SnapMirror SnapVault fan-out configuration.



After SRM performs failback and a second reversal of the SnapMirror relationships, the production data is back at the primary site. This data is now protected in the same way that it was before the failover to the DR site—through SnapMirror and SnapVault backups.

Use of Qtrees in Site Recovery Manager environments

Qtrees are special directories that allow the application of file system quotas for NAS. ONTAP 9 allows the creation of qtrees, and qtrees can exist in volumes that are replicated with SnapMirror. However, SnapMirror does not allow replication of individual qtrees or qtree-level replication. All SnapMirror replication is at the volume level only. For this reason, NetApp does not recommend the use of qtrees with SRM.

Mixed FC and iSCSI environments

With the supported SAN protocols (FC, FCoE, and iSCSI), ONTAP 9 provides LUN services—that is, the ability to create and map LUNs to attached hosts. Because the cluster consists of multiple controllers, there are multiple logical paths that are managed by multipath I/O to any individual LUN. Asymmetric logical unit access (ALUA) is used on the hosts so that the optimized path to a LUN is selected and is made active for data transfer. If the optimized path to any LUN changes (for example, because the containing volume is moved), ONTAP 9 automatically recognizes and nondisruptively adjusts for this change. If the optimized path becomes unavailable, ONTAP can nondisruptively switch to any other available path.

VMware SRM and NetApp SRA support the use of the FC protocol at one site and the iSCSI protocol at the other site. It does not support having a mix of FC-attached datastores and iSCSI-attached datastores in the same ESXi host or in different hosts in the same cluster, however. This configuration is not supported with SRM because, during the SRM failover or test failover, SRM includes all FC and iSCSI initiators in the ESXi hosts in the request.

Best Practice
SRM and SRA support mixed FC and iSCSI protocols between the protected and recovery sites. However, each site should be configured with only one protocol, either FC or iSCSI, not both protocols at the same site. If a requirement exists to have both FC and iSCSI protocols configured at the same site, NetApp recommends that some hosts use iSCSI and other hosts use FC. NetApp also recommends in this case that SRM resource mappings be set up so that the VMs are configured to fail over into one group of hosts or the other.

Troubleshooting SRM when using vVols replication

The workflow within SRM is significantly different when using vVols replication from what is used with SRA and traditional datastores. For example, there is no array manager concept. As such, `discoverarrays` and `discoverdevices` commands are never seen.

When troubleshooting, it is beneficial to understand the new workflows, which are listed below:

1. queryReplicationPeer: Discovers the replication agreements between two fault domains.
2. queryFaultDomain: Discovers fault domain hierarchy.
3. queryReplicationGroup: Discovers the replication groups present in the source or target domains.
4. syncReplicationGroup: Synchronizes the data between source and target.
5. queryPointInTimeReplica: Discovers the point in time replicas on a target.
6. testFailoverReplicationGroupStart: Begins test failover.
7. testFailoverReplicationGroupStop: Ends test failover.
8. promoteReplicationGroup: Promotes a group currently in test to production.
9. prepareFailoverReplicationGroup: Prepares for a disaster recovery.
10. failoverReplicationGroup: Executes disaster recovery.
11. reverseReplicateGroup: Initiates reverse replication.
12. queryMatchingContainer: Finds containers (along with Hosts or Replication Groups) that might satisfy a provisioning request with a given policy.
13. queryResourceMetadata: Discovers the metadata of all resources from the VASA provider, the resource utilization can be returned as an answer to the queryMatchingContainer function.

The most common error seen when configuring vVols replication is a failure to discover the SnapMirror relationships. This occurs because the volumes and SnapMirror relationships are created outside of the purview of ONTAP Tools. Therefore, it is a best practice to always make sure your SnapMirror relationship is fully initialized and that you have run a rediscovery in ONTAP Tools at both sites before attempting to create a replicated vVols datastore.

Additional Information

To learn more about the information that is described in this document, review the following documents and/or websites:

- TR-4597: VMware vSphere for ONTAP
  https://docs.netapp.com/us-en/ontap-apps-dbs/vmware/vmware-vsphere-overview.html

- TR-4400: VMware vSphere Virtual Volumes with ONTAP
  https://docs.netapp.com/us-en/ontap-apps-dbs/vmware/vmware-vvols-overview.html
- TR-4015 SnapMirror Configuration Best Practice Guide for ONTAP 9
  https://www.netapp.com/media/17229-tr4015.pdf?v=127202175503P
- RBAC User Creator for ONTAP
  https://mysupport.netapp.com/site/tools/tool-eula/rbac
- ONTAP tools for VMware vSphere Resources
  https://mysupport.netapp.com/site/products/all/details/otv/docsandkb-tab
- VMware Site Recovery Manager Documentation
  https://docs.vmware.com/en/Site-Recovery-Manager/index.html

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

vSphere Metro Storage Cluster with ONTAP

vSphere Metro Storage Cluster with ONTAP

VMware's industry-leading vSphere hypervisor can be deployed as a stretched cluster referred to as a vSphere Metro Storage Cluster (vMSC).

vMSC solutions are supported with both NetApp® MetroCluster™ and SnapMirror active sync (formerly known as SnapMirror Business Continuity, or SMBC) and provide advanced business continuity if one or more failure domains suffer a total outage. The resilience to different modes of failure depends on which configuration options you choose.

Continuous Availability Solutions for vSphere Environments

ONTAP architecture is a flexible and scalable storage platform that provides SAN (FCP, iSCSI, and NVMe-oF) and NAS (NFS v3 and v4.1) services for datastores. The NetApp AFF, ASA, and FAS storage systems use the ONTAP operating system to offer additional protocols for guest storage access like S3 and SMB/CIFS.

NetApp MetroCluster uses NetApp's HA (controller failover or CFO) function to protect against controller failures. It also includes local SyncMirror technology, cluster failover on disaster (controller failover on demand or CFOD), hardware redundancy, and geographical separation to achieve high levels of availability. SyncMirror synchronously mirrors data across the two halves of the MetroCluster configuration by writing data to two plexes: the local plex (on the local shelf) actively serving data and the remote plex (on the remote shelf) normally not serving data. Hardware redundancy is put in place for all MetroCluster components such as controllers, storage, cables, switches (used with fabric MetroCluster), and adapters.

NetApp SnapMirror active sync provides datastore-granular protection with FCP and iSCSI SAN protocols, allowing you to selectively protect only high-priority workloads. It offers active-active access to both local and remote sites, unlike NetApp MetroCluster which is an active-standby solution. At present, active sync is an asymmetric solution where one side is preferred over the other, providing better performance. This is achieved using ALUA (Asymmetric Logical Unit Access) functionality which automatically informs the ESXi host which controllers to prefer. However, NetApp has announced that active sync will soon enable fully symmetric access.

To create a VMware HA/DRS cluster across two sites, ESXi hosts are used and managed by a vCenter Server Appliance (VCSA). The vSphere management, vMotion®, and virtual machine networks are connected through a redundant network between the two sites. The vCenter Server managing the HA/DRS cluster can connect to the ESXi hosts at both sites and should be configured using vCenter HA.

Refer to How Do You Create and Configure Clusters in the vSphere Client to configure vCenter HA.

You should also refer to VMware vSphere Metro Storage Cluster Recommended Practices.

What is vSphere Metro Storage Cluster?

vSphere Metro Storage Cluster (vMSC) is a certified configuration that protects virtual machines (VMs) and containers against failures. This is achieved by using stretched storage concepts along with clusters of ESXi hosts, which are distributed across different failure domains such as racks, buildings, campuses, or even cities. The NetApp MetroCluster and SnapMirror active sync storage technologies are used to provide RPO=0 or near RPO=0 protection respectively to the host clusters. The vMSC configuration is designed to ensure that data is always available even if a complete physical or logical “site” fails. A storage device that is part of the vMSC configuration must be certified after undergoing a successful vMSC certification process. All the supported storage devices can be found in the VMware Storage Compatibility Guide.

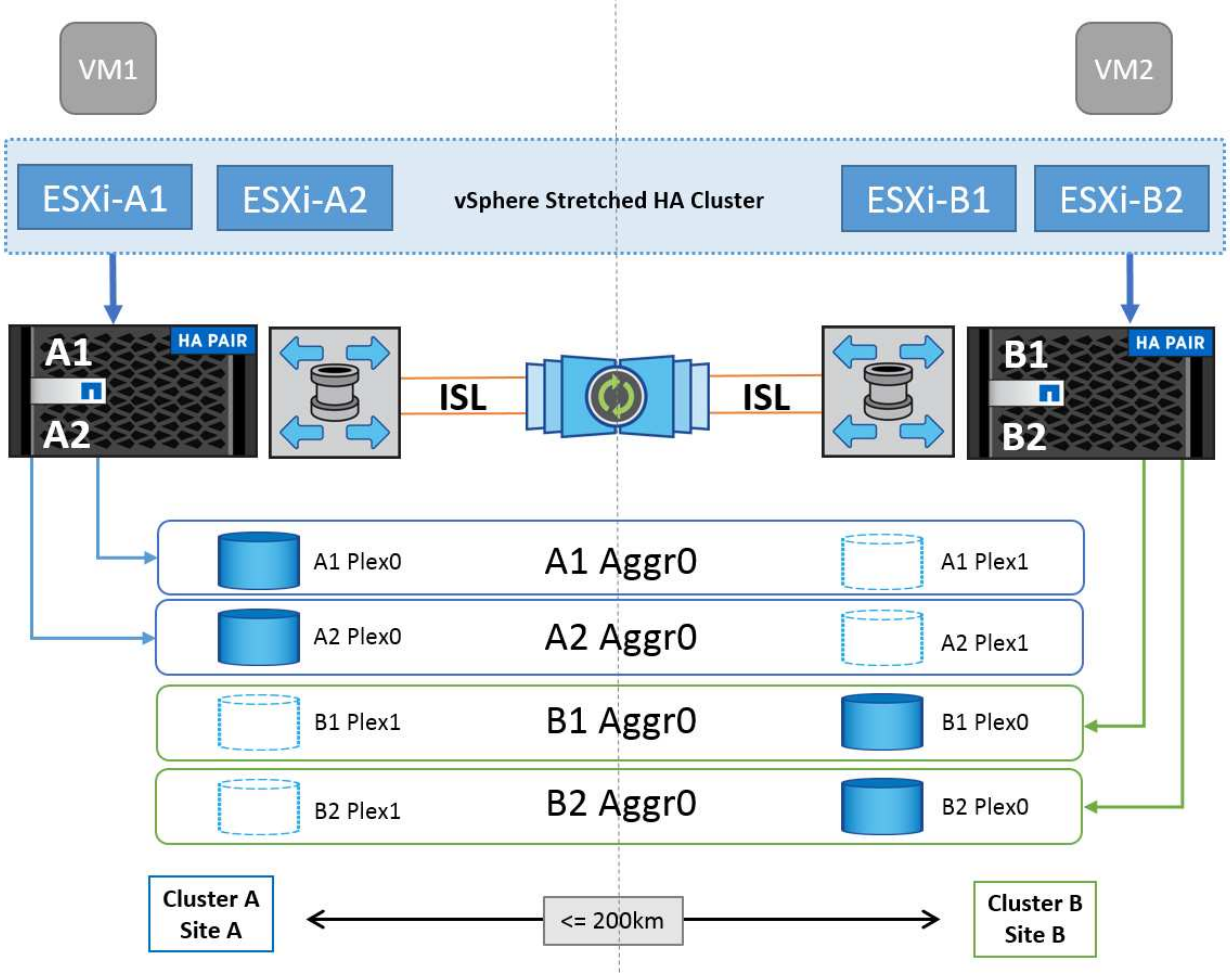
If you want more information about the design guidelines for vSphere Metro Storage Cluster, you can refer to the following documentation:

- VMware vSphere support with NetApp MetroCluster
- VMware vSphere support with NetApp SnapMirror Business Continuity (now known as SnapMirror active sync)

Depending on the latency considerations, NetApp MetroCluster can be deployed in two different configurations for use with vSphere:

- Stretch MetroCluster
- Fabric MetroCluster

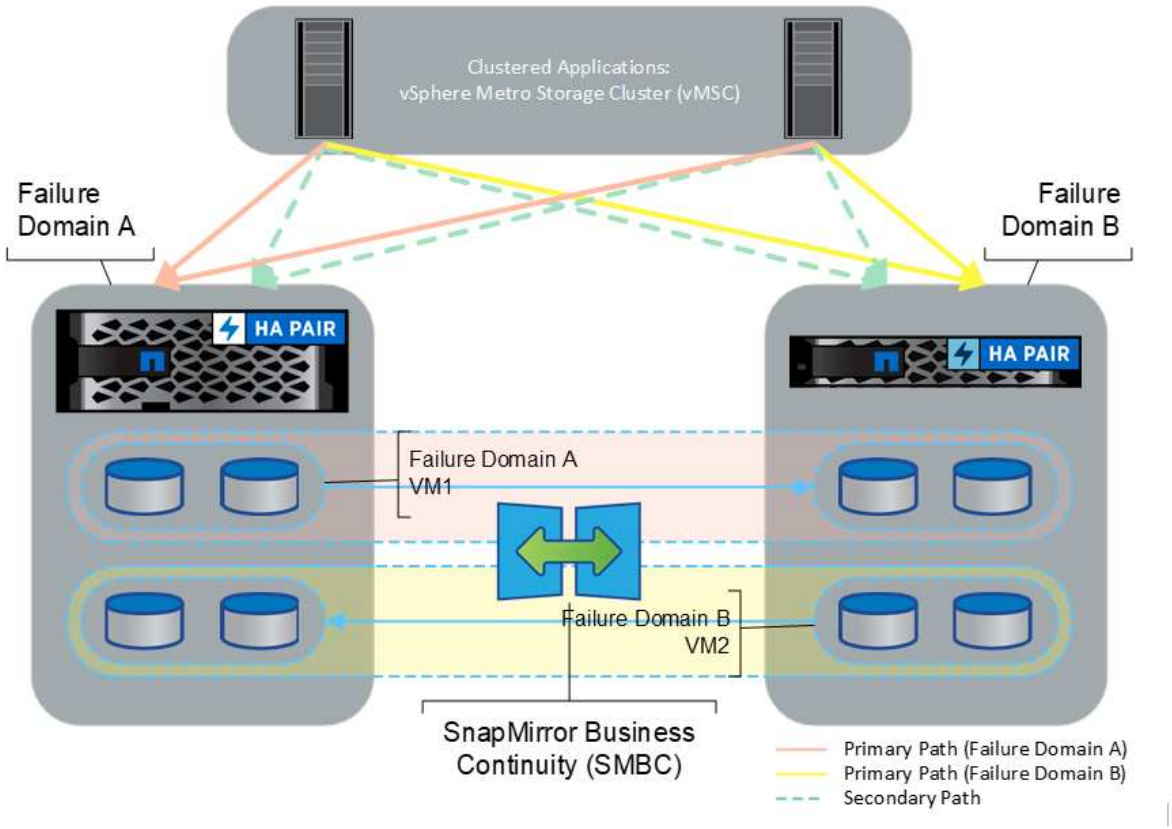
The following illustrates a high-level topology diagram of stretch MetroCluster.



Refer to MetroCluster documentation for specific design and deployment information for MetroCluster.

SnapMirror active sync can also be deployed in two different ways.

- Asymmetric
- Symmetric (private preview in ONTAP 9.14.1)



Refer to NetApp Docs for specific design and deployment information for SnapMirror active sync.

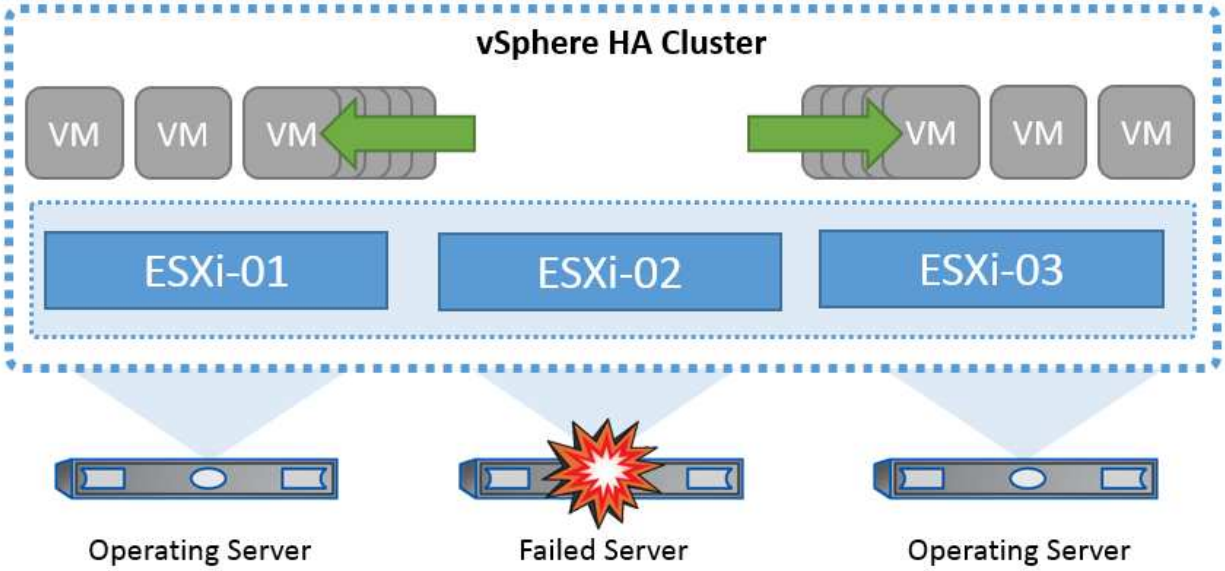
VMware vSphere Solution Overview

The vCenter Server Appliance (VCSA) is the powerful centralized management system and single pane of glass for vSphere that enables administrators to effectively operate ESXi clusters. It facilitates key functions such as VM provisioning, vMotion operation, High Availability (HA), Distributed Resource Scheduler (DRS), Tanzu Kubernetes Grid, and more. It is an essential component in VMware cloud environments and should be designed with service availability in mind.

vSphere High Availability

VMware's cluster technology groups ESXi servers into pools of shared resources for virtual machines and provides vSphere High Availability (HA). vSphere HA provides easy-to-use, high availability for applications running in virtual machines. When the HA feature is enabled on the cluster, each ESXi server maintains communication with other hosts so that if any ESXi host becomes unresponsive or isolated, the HA cluster can negotiate the recovery of the virtual machines that were running on that ESXi host among surviving hosts in the cluster. In the event of a guest operating system failure, vSphere HA restarts the affected virtual machine on the same physical server. vSphere HA makes it possible to reduce planned downtime, prevent unplanned downtime, and rapidly recover from outages.

vSphere HA cluster recovering VMs from failed server.



It's important to understand that VMware vSphere has no knowledge of NetApp MetroCluster or SnapMirror active sync and sees all ESXi hosts in the vSphere cluster as eligible hosts for HA cluster operations depending on host and VM group affinity configurations.

Host Failure Detection

As soon as the HA cluster is created, all hosts in the cluster participate in election, and one of the hosts becomes a master. Each slave performs network heartbeat to the master, and the master in turn performs network heartbeat on all slave hosts. The master host of a vSphere HA cluster is responsible for detecting the failure of slave hosts.

Depending on the type of failure detected, the virtual machines running on the hosts might need to be failed over.

In a vSphere HA cluster, three types of host failure are detected:

- Failure - A host stops functioning.
- Isolation - A host becomes network isolated.
- Partition - A host loses network connectivity with the master host.

The master host monitors the slave hosts in the cluster. This communication is done through the exchange of network heartbeats every second. When the master host stops receiving these heartbeats from a slave host, it checks for host liveness before declaring the host to have failed. The liveness check that the master host performs is to determine whether the slave host is exchanging heartbeats with one of the datastores. Also, the master host checks whether the host responds to ICMP pings sent to its management IP addresses to detect whether it is merely isolated from its master node or completely isolated from the network. It does this by pinging the default gateway. One or more isolation addresses can be specified manually to enhance the reliability of isolation validation.

Best Practice

NetApp recommends specifying a minimum of two additional isolation addresses, and that each of these addresses be site-local. This will enhance the reliability of isolation validation.

Host Isolation Response

Isolation Response is a setting in vSphere HA that determines the action triggered on Virtual Machines when a host in a vSphere HA cluster loses its management network connections but continues to run. There are three options for this setting, “Disabled”, “Shut Down and Restart VMs,” and “Power Off and Restart VMs.”

“Shut Down” is better than “Power Off”, which does not flush most recent changes to disk or commit transactions. If virtual machines have not shut down in 300 seconds they are powered off. To change the wait time, use the advanced option das.isolationshutdowntimeout.

Before HA initiates the isolation response, it first checks to see if the vSphere HA master agent owns the datastore that contains the VM config files. If not, then the host will not trigger the isolation response, because there is no master to restart the VMs. The host will periodically check the datastore state to determine if it is claimed by a vSphere HA agent that holds the master role.

Best Practice

NetApp recommends setting the “Host Isolation Response” to Disabled.

A split-brain condition can occur if a host becomes isolated or partitioned from the vSphere HA master host and the master is unable to communicate via heartbeat datastores or by ping. The master declares the isolated host dead and restarts the VMs on other hosts in the cluster. A split-brain condition now exists because there are two instances of the virtual machine running, only one of which can read or write the virtual disks. Split-brain conditions can now be avoided by configuring VM Component Protection (VMCP).

VM Component Protection (VMCP)

One of the feature enhancements in vSphere 6, relevant to HA, is VMCP. VMCP provides enhanced protection from All Paths Down (APD) and Permanent Device Loss (PDL) conditions for block (FC, iSCSI, FCoE) and file storage (NFS).

Permanent Device Loss (PDL)

PDL is a condition that occurs when a storage device permanently fails or is administratively removed and is not expected to return. The NetApp storage array issues a SCSI Sense code to ESXi declaring that the device is permanently lost. In the Failure Conditions and VM Response section of vSphere HA, you can configure what the response should be after a PDL condition is detected.

Best Practice

NetApp recommends setting the “Response for Datastore with PDL” to “**Power off and restart VMs**”. When this condition is detected a VM will be restarted instantly on a healthy host within the vSphere HA cluster.

All Paths Down (APD)

APD is a condition that occurs when a storage device becomes inaccessible to the host and no paths to the array are available. ESXi considers this a temporary problem with the device and is expecting it to become available again.

When an APD condition is detected, a timer is started. After 140 seconds, the APD condition is officially declared, and the device is marked as APD time out. When the 140 seconds have passed, HA will start counting the number of minutes specified in the Delay for VM Failover APD. When the specified time has passed, HA will restart the impacted virtual machines. You can configure VMCP to respond differently if desired (Disabled, Issue Events, or Power Off and Restart VMs).

Best Practice

NetApp recommends configuring the “Response for Datastore with APD” to “**Power off and restart VMs (conservative)**”.

Conservative refers to the likelihood of HA being able to restart VMs. When set to Conservative, HA will only restart the VM that is impacted by the APD if it knows another host can restart it. In the case of Aggressive, HA will try to restart the VM even if it doesn’t know the state of the other hosts. This can result in VMs not being restarted if there is no host with access to the datastore it is located on.

If the APD status is resolved and access to the storage is restored before the time-out has passed, HA will not unnecessarily restart the virtual machine unless you explicitly configure it to do so. If a response is desired even when the environment has recovered from the APD condition, then Response for APD Recovery After APD Timeout should be configured to Reset VMs.

Best Practice

NetApp recommends configuring Response for APD Recovery After APD Timeout to Disabled.

VMware DRS Implementation for NetApp MetroCluster

VMware DRS is a feature that aggregates the host resources in a cluster and is primarily used to load balance within a cluster in a virtual infrastructure. VMware DRS primarily calculates the CPU and memory resources to perform load balancing in a cluster. Because vSphere is unaware of stretched clustering, it considers all hosts in both sites when load balancing. To avoid cross-site traffic, NetApp recommends configuring DRS affinity rules to manage a logical separation of VMs. This will ensure that unless there is a complete site failure, HA and DRS will only use local hosts.

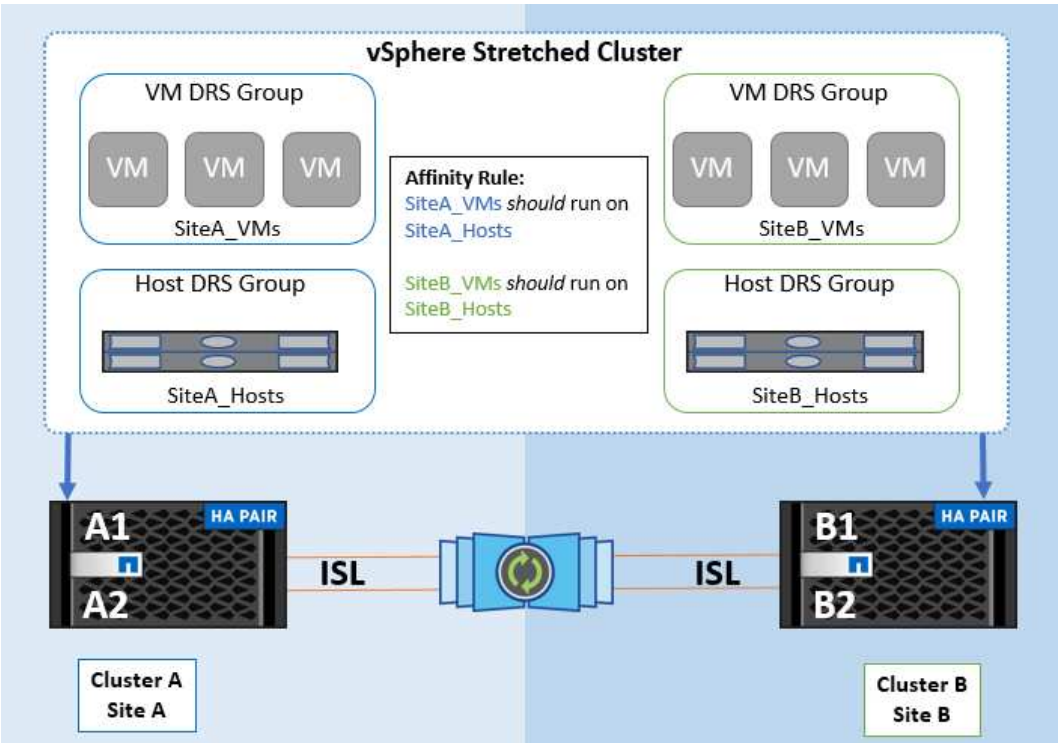
If you create a DRS affinity rule for your cluster, you can specify how vSphere applies that rule during a virtual machine failover.

There are two types of rules you can specify vSphere HA failover behavior:

- VM anti-affinity rules force specified virtual machines to remain apart during failover actions.
- VM host affinity rules place specified virtual machines on a particular host or a member of a defined group of hosts during failover actions.

Using VM host affinity rules in VMware DRS, one can have a logical separation between site A and site B so that the VM runs on the host at the same site as the array that is configured as the primary read/write controller for a given datastore. Also, VM host affinity rules enable virtual machines to stay local to the storage, which in turn ascertains the virtual machine connection in case of network failures between the sites.

The following is an example of VM host groups and affinity rules.



Best Practice

NetApp recommends implementing “should” rules instead of “must” rules because they are violated by vSphere HA in the case of a failure. Using “must” rules could potentially lead to service outages.

Availability of services should always prevail over performance. In the scenario where a full data center fails, “must” rules must choose hosts from the VM host affinity group, and when the data center is unavailable, the virtual machines will not restart.

VMware Storage DRS Implementation with NetApp MetroCluster

The VMware Storage DRS feature enables the aggregation of datastores into a single unit and balances virtual machine disks when storage I/O control thresholds are exceeded.

Storage I/O control is enabled by default on Storage DRS–enabled DRS clusters. Storage I/O control allows an administrator to control the amount of storage I/O that is allocated to virtual machines during periods of I/O congestion, which enables more important virtual machines to have preference over less important virtual machines for I/O resource allocation.

Storage DRS uses Storage vMotion to migrate the virtual machines to different datastores within a datastore cluster. In a NetApp MetroCluster environment, a virtual machine migration needs to be controlled within the datastores of that site. For example, virtual machine A, running on a host at site A, should ideally migrate within the datastores of the SVM at site A. If it fails to do so, the virtual machine will continue to operate but with degraded performance, since the virtual disk read/write will be from site B through inter-site links.

Best Practice

NetApp recommends creating datastore clusters with respect to storage site affinity; that is, datastores with site affinity for site A should not be mixed with datastore clusters with datastores with site affinity for site B.

Whenever a virtual machine is newly provisioned or migrated using Storage vMotion, NetApp recommends that all the VMware DRS rules specific to those virtual machines be manually updated, accordingly. This will

ascertain the virtual machine affinity at the site level for both host and datastore and thus reduce the network and storage overhead.

vMSC Design and Implementation Guidelines

This document outlines the design and implementation guidlines for vMSC with ONTAP storage systems.

NetApp Storage Configuration

Setup instructions for NetApp MetroCluster (referred to as an MCC configuration) are available at MetroCluster Documentation. Instructions for SnapMirror active sync are also available at SnapMirror Business Continuity overview.

Once you have configured MetroCluster, administering it is like managing a traditional ONTAP environment. You can set up Storage Virtual Machines (SVMs) using various tools like the Command Line Interface (CLI), System Manager, or Ansible. Once the SVMs are configured, create Logical Interfaces (LIFs), volumes, and Logical Unit Numbers (LUNs) on the cluster that will be used for normal operations. These objects will automatically be replicated to the other cluster using the cluster peering network.

If not using MetroCluster, you can use SnapMirror active sync which provides datastore-granular protection and active-active access across multiple ONTAP clusters in different failure domains. SnapMirror active sync uses consistency groups to ensure write-order consistency among one or more datastores and you can create multiple consistency groups depending on your application and datastore requirements. Consistency groups are especially useful for applications that require data synchronization between multiple datastores. SnapMirror active sync also supports Raw Device Mappings (RDMs) and guest-connected storage with in-guest iSCSI initiators. You can learn more about consistency groups at Consistency groups overview.

There is some difference in managing a vMSC configuration with SnapMirror active sync when compared to a MetroCluster. First, this is a SAN-only configuration, no NFS datastores can be protected with SnapMirror active sync. Second, you must map both copies of the LUNs to your ESXi hosts for them to access the replicated datastores in both failure domains.

VMware vSphere HA

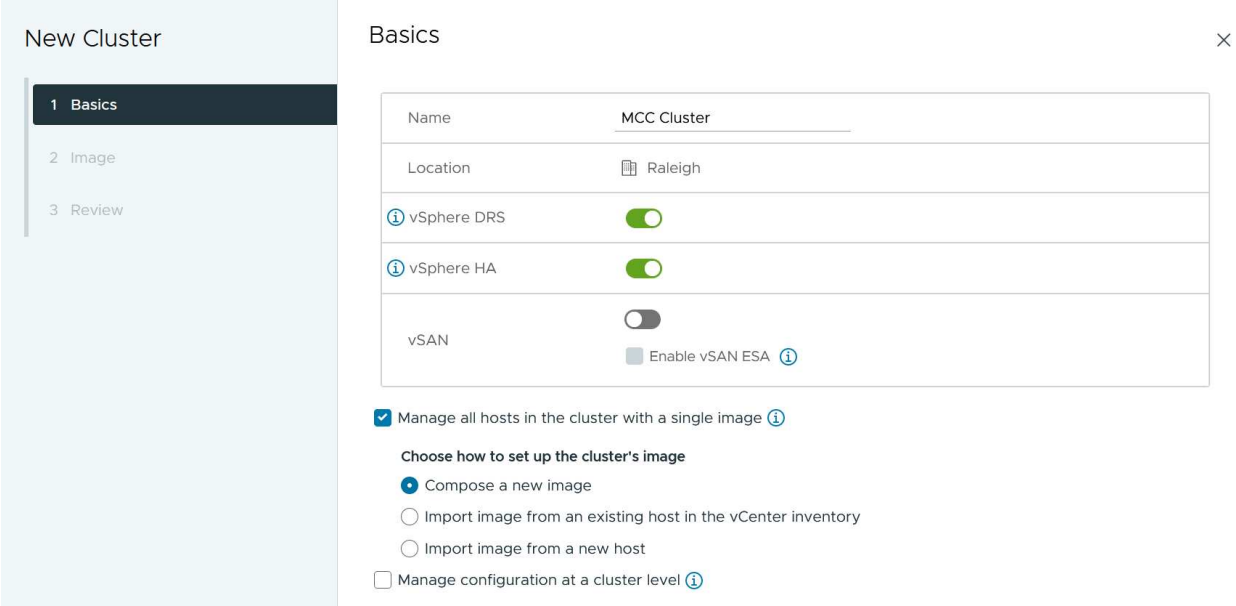
Create a vSphere HA Cluster

Creating a vSphere HA cluster is a multi-step process that is fully documented at How Do You Create and Configure Clusters in the vSphere Client on docs.vmware.com. In short, you must first create an empty cluster, and then, using vCenter, you must add hosts and specify the cluster's vSphere HA and other settings.

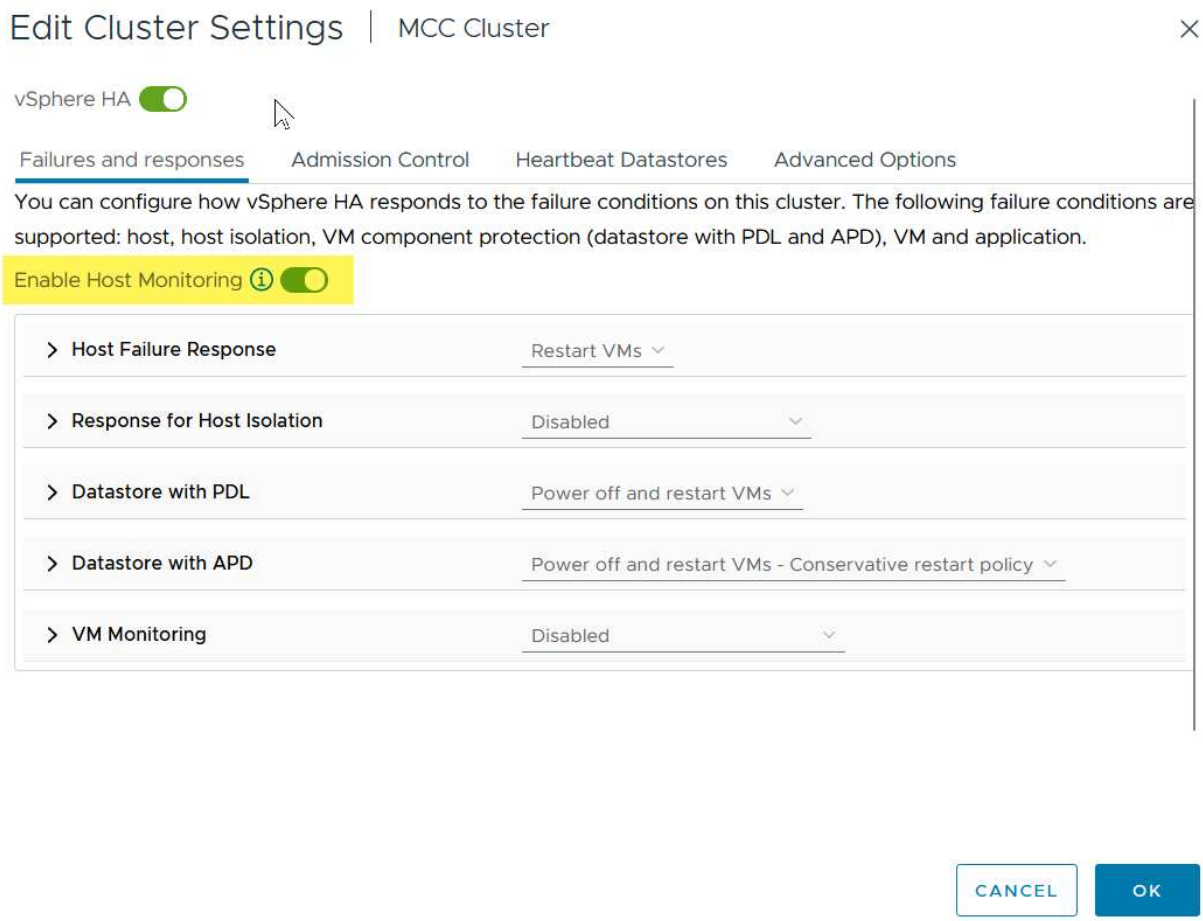
Note: Nothing in this document supersedes VMware vSphere Metro Storage Cluster Recommended Practices

To configure an HA cluster, complete the following steps:

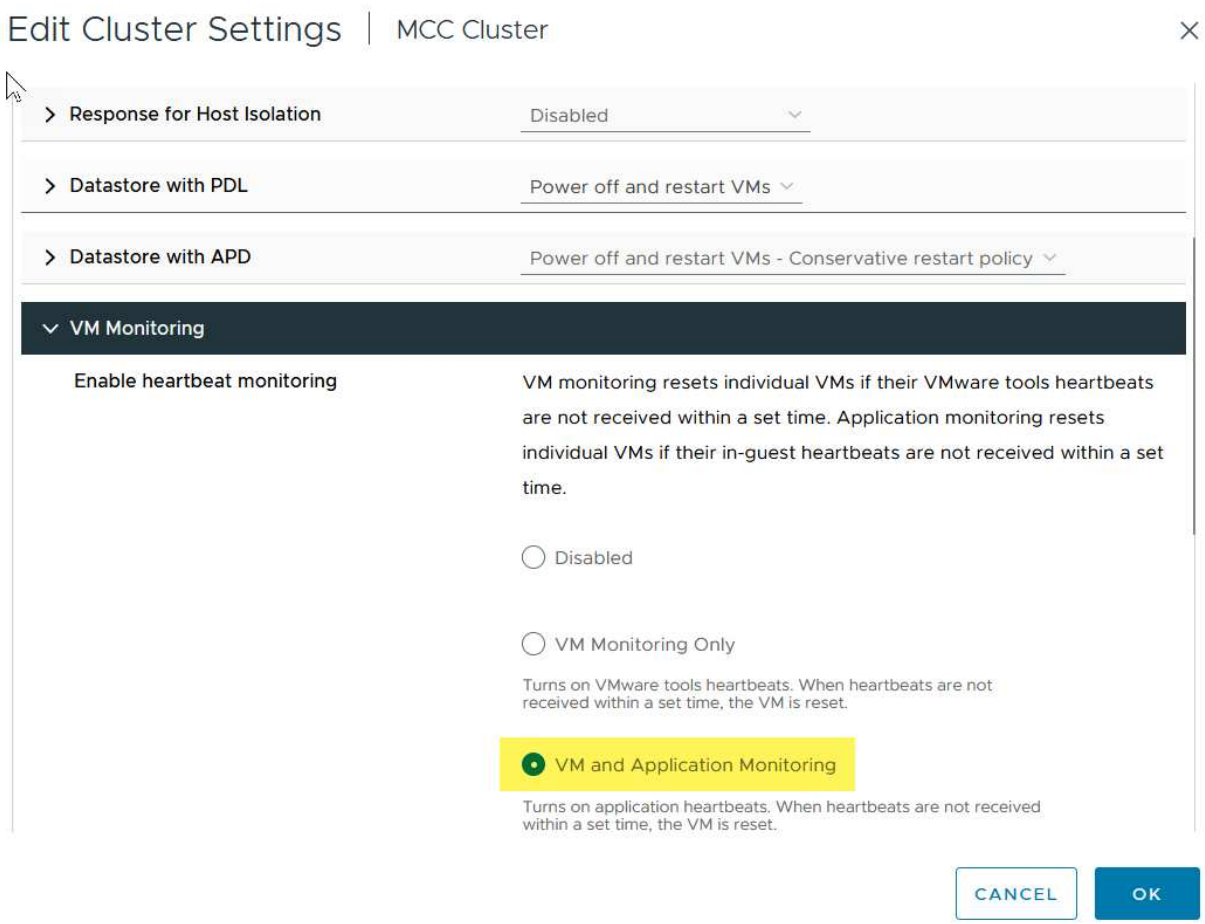
1. Connect to the vCenter UI.
2. In Hosts and Clusters, browse to the data center where you want to create your HA cluster.
3. Right-click the data center object and select New Cluster. Under basics ensure you have enabled vSphere DRS and vSphere HA. Complete the wizard.



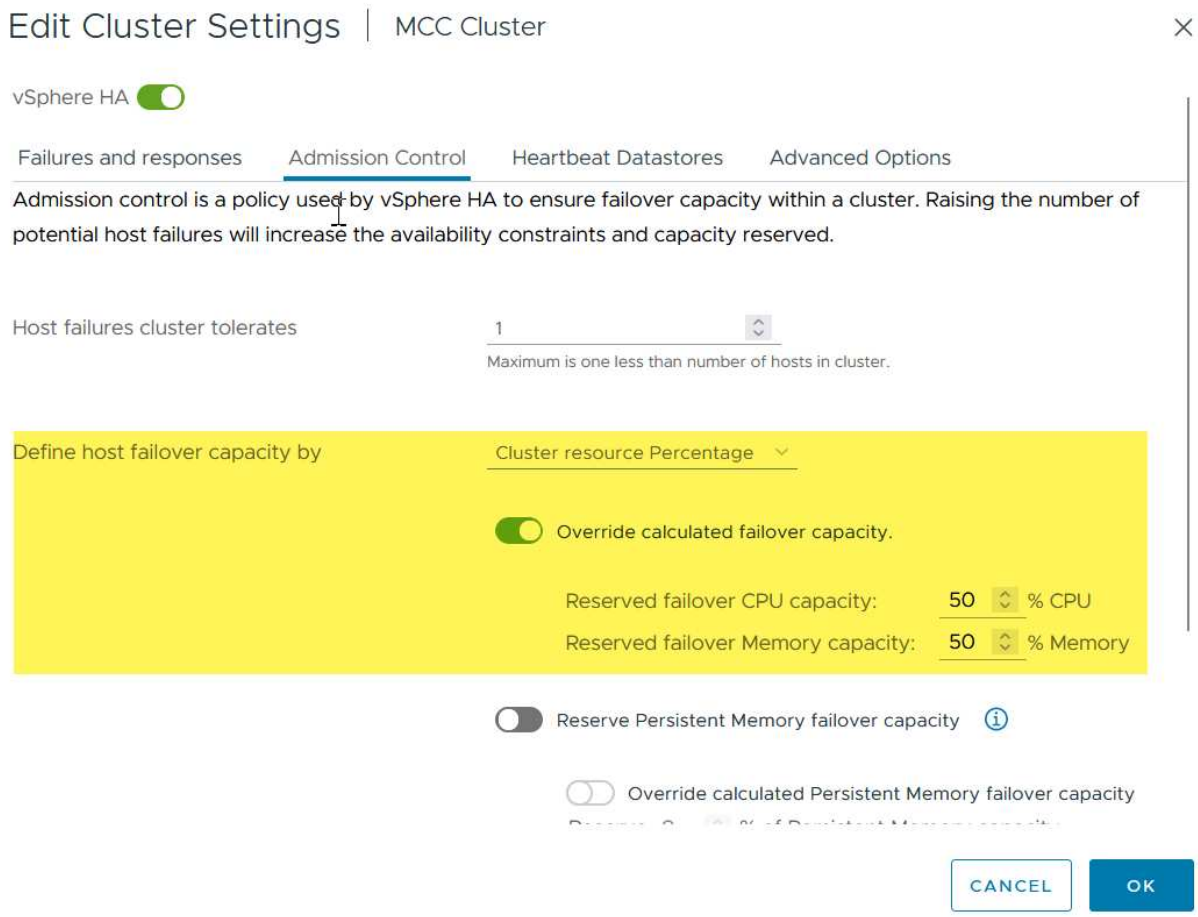
4. Select the cluster and go to the configure tab. Select vSphere HA and click edit.
5. Under Host Monitoring, select the Enable Host Monitoring option.



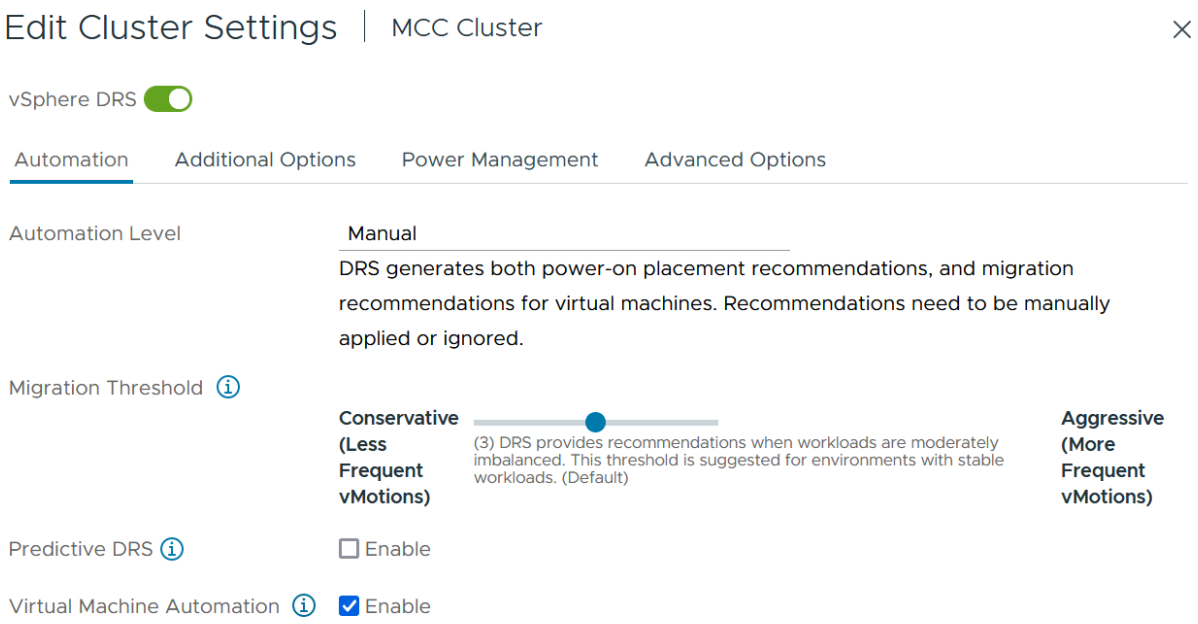
6. While still on the Failures and Responses tab, Under VM Monitoring, select the VM Monitoring Only option or VM and Application Monitoring option.



7. Under Admission Control, set the HA admission control option to cluster resource reserve; use 50% CPU/MEM.



8. Click “OK”.
9. Select DRS and click EDIT.
10. Set the automation level to manual unless required by your applications.



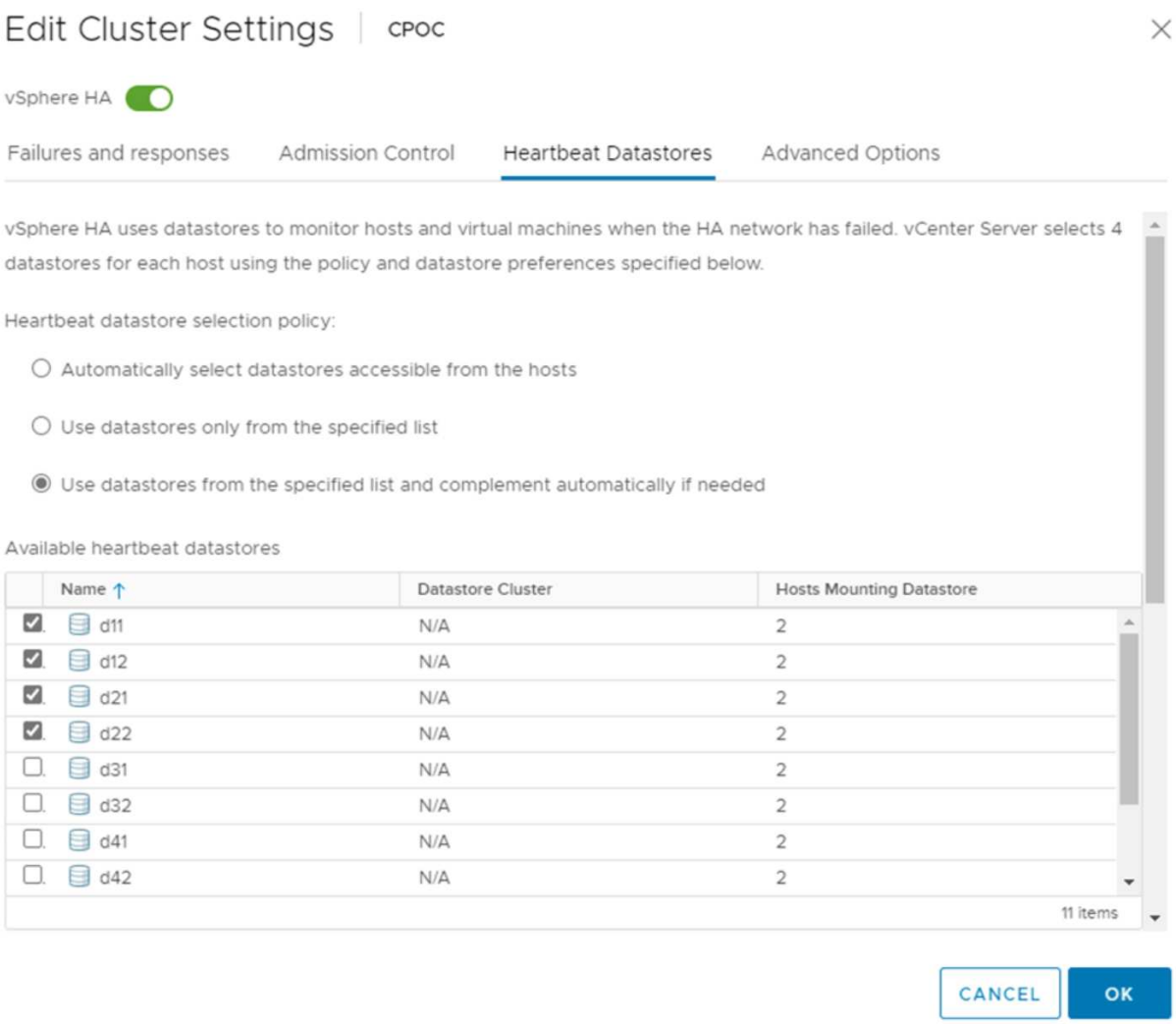
11. Enable VM Component Protection, refer to docs.vmware.com.
12. The following additional vSphere HA settings are recommended for vMSC with MCC:

Failure	Response
Host failure	Restart VMs
Host isolation	Disabled
Datastore with Permanent Device Loss (PDL)	Power off and restart VMs
Datastore with All paths Down (APD)	Power off and restart VMs
Guest not heartbeating	Reset VMs
VM restart policy	Determined by the importance of the VM
Response for host isolation	Shut down and restart VMs
Response for datastore with PDL	Power off and restart VMs
Response for datastore with APD	Power off and restart VMs (conservative)
Delay for VM failover for APD	3 minutes
Response for APD recovery with APD timeout	Disabled
VM monitoring sensitivity	Preset high

Configure Datastores for Heartbeating

vSphere HA uses datastores to monitor hosts and virtual machines when the management network has failed. You can configure how vCenter selects heartbeat datastores. To configure datastores for heartbeating, complete the following steps:

1. In the Datastore Heartbeating section, select Use Datastores from the Specified List and Compliment Automatically if Needed.
2. Select the datastores you want vCenter to use from both sites and press OK.

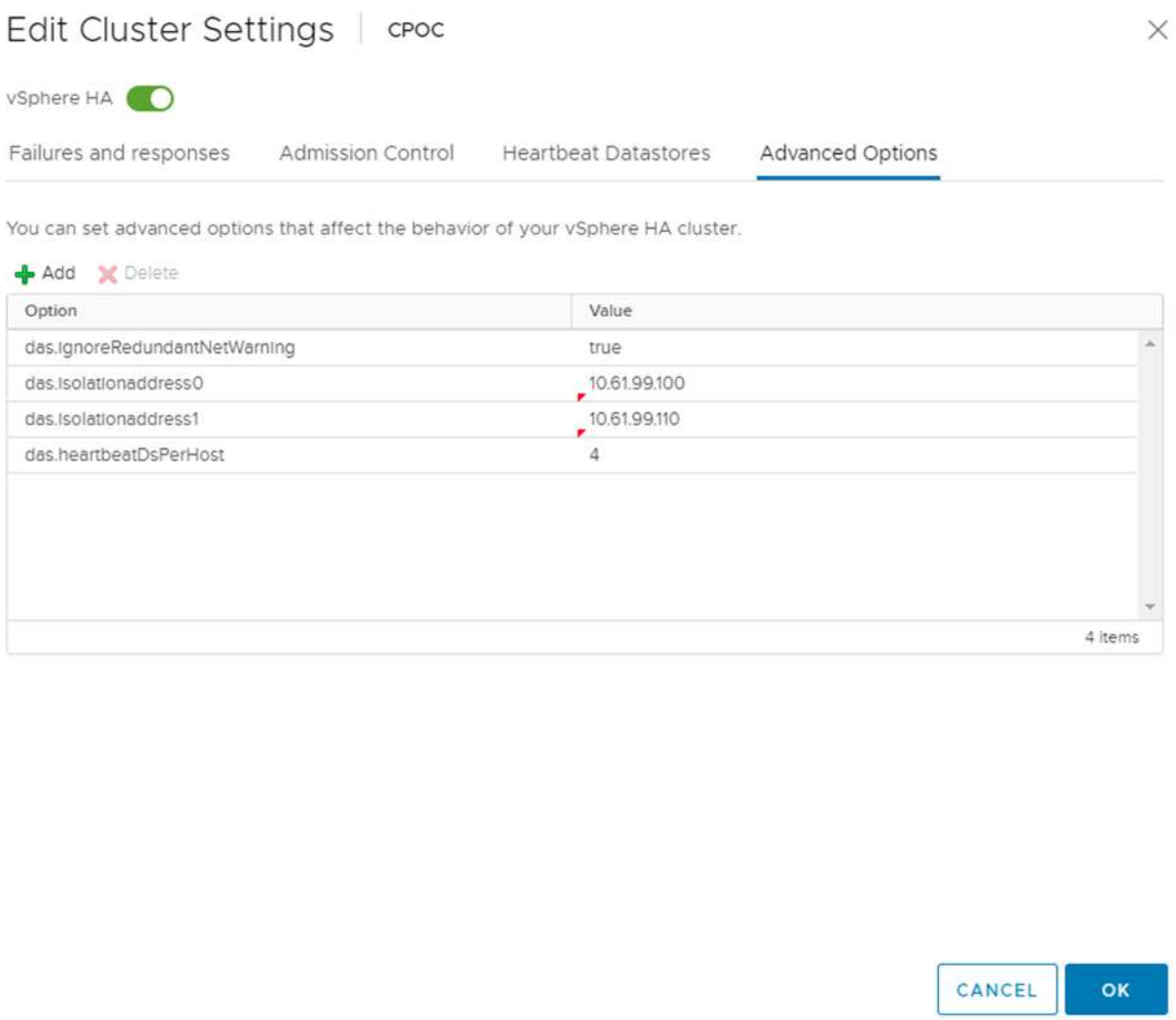


Configure Advanced Options

Host Failure Detection

Isolation events occur when hosts within an HA cluster lose connectivity to either the network or other hosts in the cluster. By default, vSphere HA will use the default gateway for its management network as the default isolation address. However, you can specify additional isolation addresses for the host to ping to determine whether an isolation response should be triggered. Add two isolation IPs that can ping, one per site. Do not use the gateway IP. The vSphere HA advanced setting used is das.isolationaddress. You can use ONTAP or Mediator IP addresses for this purpose.

Refer to core.vmware.com for more information.



Adding an advanced setting called das.heartbeatDsPerHost can increase the number of heartbeat datastores. Use four heartbeat datastores (HB DSs)—two per site. Use the “Select from List but Compliment” option. This is needed because if one site fails, you still need two HB DSs. However, those don’t have to be protected with MCC or SnapMirror active sync.

Refer to core.vmware.com for more information.

VMware DRS Affinity for NetApp MetroCluster

In this section, we create DRS groups for VMs and hosts for each site\cluster in the MetroCluster environment. Then we configure VM\Host rules to align VM host affinity with local storage resources. For example, site A VMs belong to VM group sitea_vms and site A hosts belong to host group sitea_hosts. Next, in VM\Host Rules, we state that sitea_vms should run on hosts in sitea_hosts.

Best Practice

- NetApp highly recommends the specification **Should Run on Hosts in Group** rather than the specification **Must Run on Hosts in Group**. In the event of a site A host failure, the VMs of site A need to be restarted on hosts at site B through vSphere HA, but the latter specification does not allow HA to restart VMs on site B because it’s a hard rule. The former specification is a soft rule and will be violated in the event of HA, thus enabling availability rather than performance.

Note: You can create an event-based alarm that is triggered when a virtual machine violates a VM-Host affinity rule. In the vSphere Client, add a new alarm for the virtual machine and select “VM is violating VM-Host Affinity Rule” as the event trigger. For more information about creating and editing alarms, refer to vSphere Monitoring and Performance documentation.

Create DRS Host Groups

To create DRS host groups specific to site A and site B, complete the following steps:

1. In the vSphere web client, right-click the cluster in the inventory and select Settings.
2. Click VM\Host Groups.
3. Click Add.
4. Type the name for the group (for instance, sitea_hosts).
5. From the Type menu, select Host Group.
6. Click Add and select the desired hosts from site A and click OK.
7. Repeat these steps to add another host group for site B.
8. Click OK.

Create DRS VM Groups

To create DRS VM groups specific to site A and site B, complete the following steps:

1. In the vSphere web client, right-click the cluster in the inventory and select Settings.

9. Click VM\Host Groups.
10. Click Add.
11. Type the name for the group (for instance, sitea_vms).
12. From the Type menu, select VM Group.
13. Click Add and select the desired VMs from site A and click OK.
14. Repeat these steps to add another host group for site B.
15. Click OK.

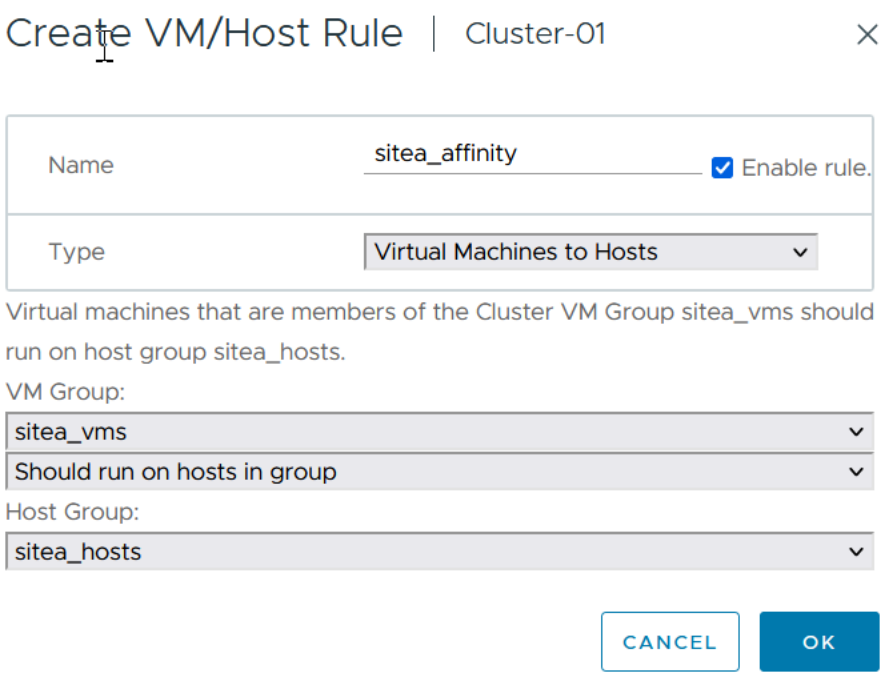
Create VM Host Rules

To create DRS affinity rules specific to site A and site B, complete the following steps:

1. In the vSphere web client, right-click the cluster in the inventory and select Settings.

1. Click VM\Host Rules.
2. Click Add.
3. Type the name for the rule (for instance, sitea_affinity).
4. Verify the Enable Rule option is checked.
5. From the Type menu, select Virtual Machines to Hosts.
6. Select the VM group (for instance, sitea_vms).
7. Select the Host group (for instance, sitea_hosts).
8. Repeat these steps to add another VM\Host Rule for site B.

9. Click OK.



VMWare vSphere Storage DRS for NetApp MetroCluster

Create Datastore Clusters

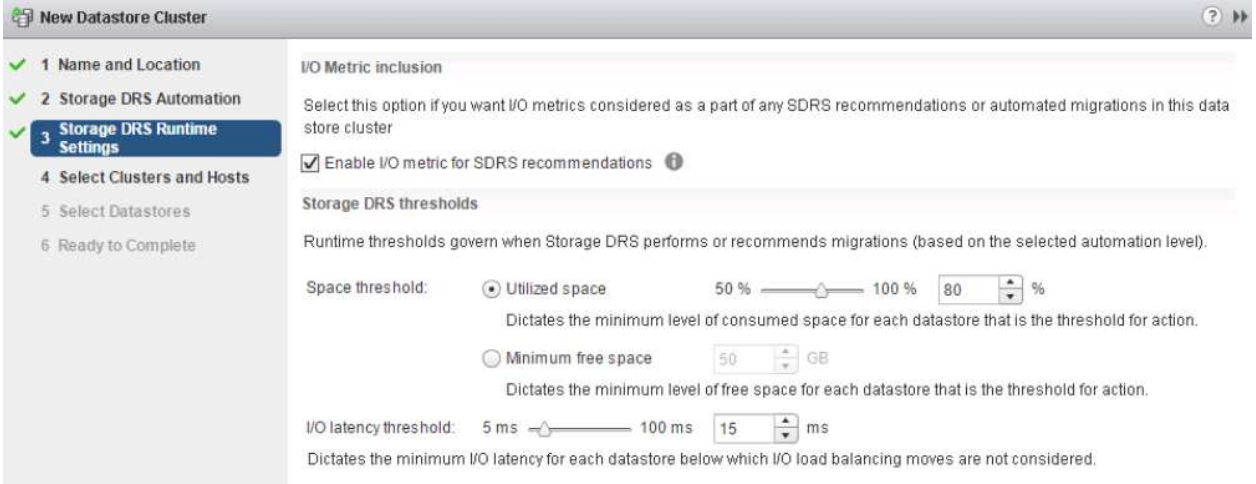
To configure a datastore cluster for each site, complete the following steps:

1. Using the vSphere web client, browse to the data center where the HA cluster resides under Storage.
2. Right-click the data center object and select Storage > New Datastore Cluster.
3. Select the Turn ON Storage DRS option and click Next.
4. Set all options to No Automation (Manual Mode) and click Next.

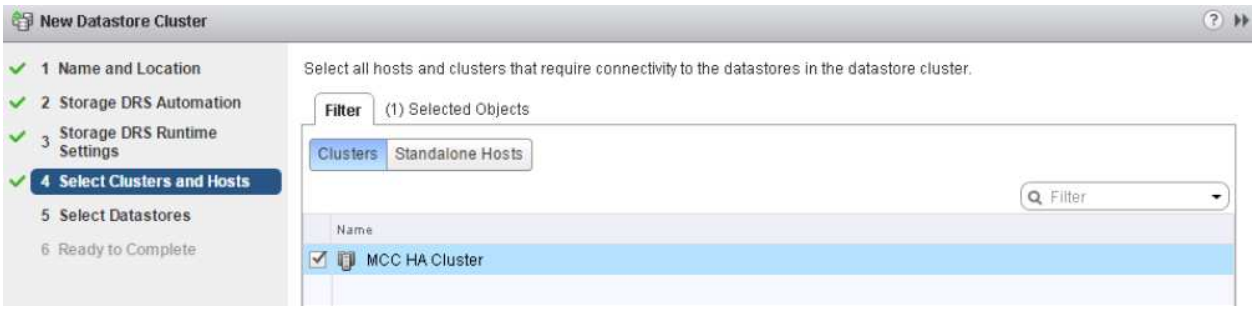
Best Practice

- NetApp recommends that Storage DRS be configured in manual mode, so that the administrator gets to decide and control when migrations need to happen.

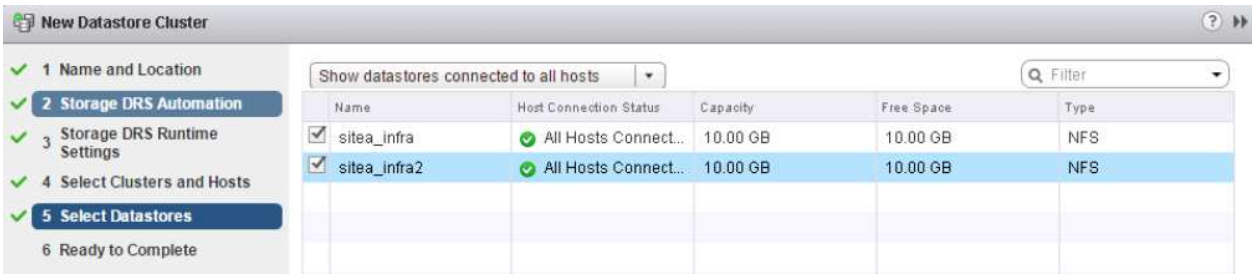
5. Verify that the Enable I/O Metric for SDRS Recommendations checkbox is checked; metric settings can be left with default values.



6. Select the HA cluster and click Next.



7. Select the datastores belonging to site A and click Next.



8. Review options and click Finish.
9. Repeat these steps to create the site B datastore cluster and verify that only datastores of site B are selected.

vCenter Server Availability

Your vCenter Server Appliances (VCSAs) should be protected with vCenter HA. vCenter HA allows you to deploy two VCSAs in an active-passive HA pair. One in each failure domain. You can read more about vCenter HA on docs.vmware.com.

Resiliency for Planned and Unplanned Events

NetApp MetroCluster and SnapMirror active sync are powerful tools that enhance the high availability and non-disruptive operations of NetApp hardware and ONTAP® software.

These tools provide site-wide protection for the entire storage environment, ensuring that your data is always available. Whether you are using standalone servers, high-availability server clusters, Docker containers, or virtualized servers, NetApp technology seamlessly maintains storage availability in the event of a total outage due to loss of power, cooling, or network connectivity, storage array shutdown, or operational error.

MetroCluster and SnapMirror active sync provide three basic methods for data continuity in the event of planned or unplanned events:

- Redundant components for protection against single-component failure
- Local HA takeover for events affecting a single controller
- Complete site protection – rapid resumption of service by moving storage and client access from the source cluster to the destination cluster

This means operations continue seamlessly in case of a single component failure and return automatically to redundant operation when the failed component is replaced.

All ONTAP clusters, except single-node clusters (typically software-defined versions, such as ONTAP Select for example), have built-in HA features called takeover and giveback. Each controller in the cluster is paired with another controller, forming an HA pair. These pairs ensure that each node is locally connected to the storage.

Takeover is an automated process where one node takes over the other's storage to maintain data services. Giveback is the reverse process that restores normal operation. Takeover can be planned, such as when performing hardware maintenance or ONTAP upgrades, or unplanned, resulting from a node panic or hardware failure.

During a takeover, Network Attached Storage Logical Interfaces (NAS LIFs) in MetroCluster configurations automatically failover. However, Storage Area Network LIFs (SAN LIFs) do not fail over; they will continue to use the direct path to the Logical Unit Numbers (LUNs).

For more information on HA takeover and giveback, please refer to the HA pair management overview. It's worth noting that this functionality is not specific to MetroCluster or SnapMirror active sync.

Site switchover with MetroCluster occurs when one site is offline or as a planned activity for site-wide maintenance. The remaining site assumes ownership of the storage resources (disks and aggregates) of the offline cluster, and the SVMs on the failed site are brought online and restarted on the disaster site, preserving their full identity for client and host access.

With SnapMirror active sync, since both copies are actively used simultaneously, your existing hosts will continue to operate. The NetApp Mediator is required to ensure site failover occurs correctly.

Failure Scenarios for vMSC with MCC

The following sections outline the expected results from various failure scenarios with vMSC and NetApp MetroCluster systems.

Single Storage Path Failure

In this scenario, if components such as the HBA port, the network port, the front-end data switch port, or an FC or Ethernet cable fails, that particular path to the storage device is marked as dead by the ESXi host. If several paths are configured for the storage device by providing resiliency at the HBA/network/switch port, ESXi ideally performs a path switchover. During this period, virtual machines remain running without getting affected, because availability to the storage is taken care of by providing multiple paths to the storage device.

Note: There is no change in MetroCluster behavior in this scenario, and all the datastores continue to be intact from their respective sites.

Best Practice

In environments in which NFS/iSCSI volumes are used, NetApp recommends having at least two network uplinks configured for the NFS vmkernel port in the standard vSwitch and the same at the port group where the NFS vmkernel interface is mapped for the distributed vSwitch. NIC teaming can be configured in either active-active or active-standby.

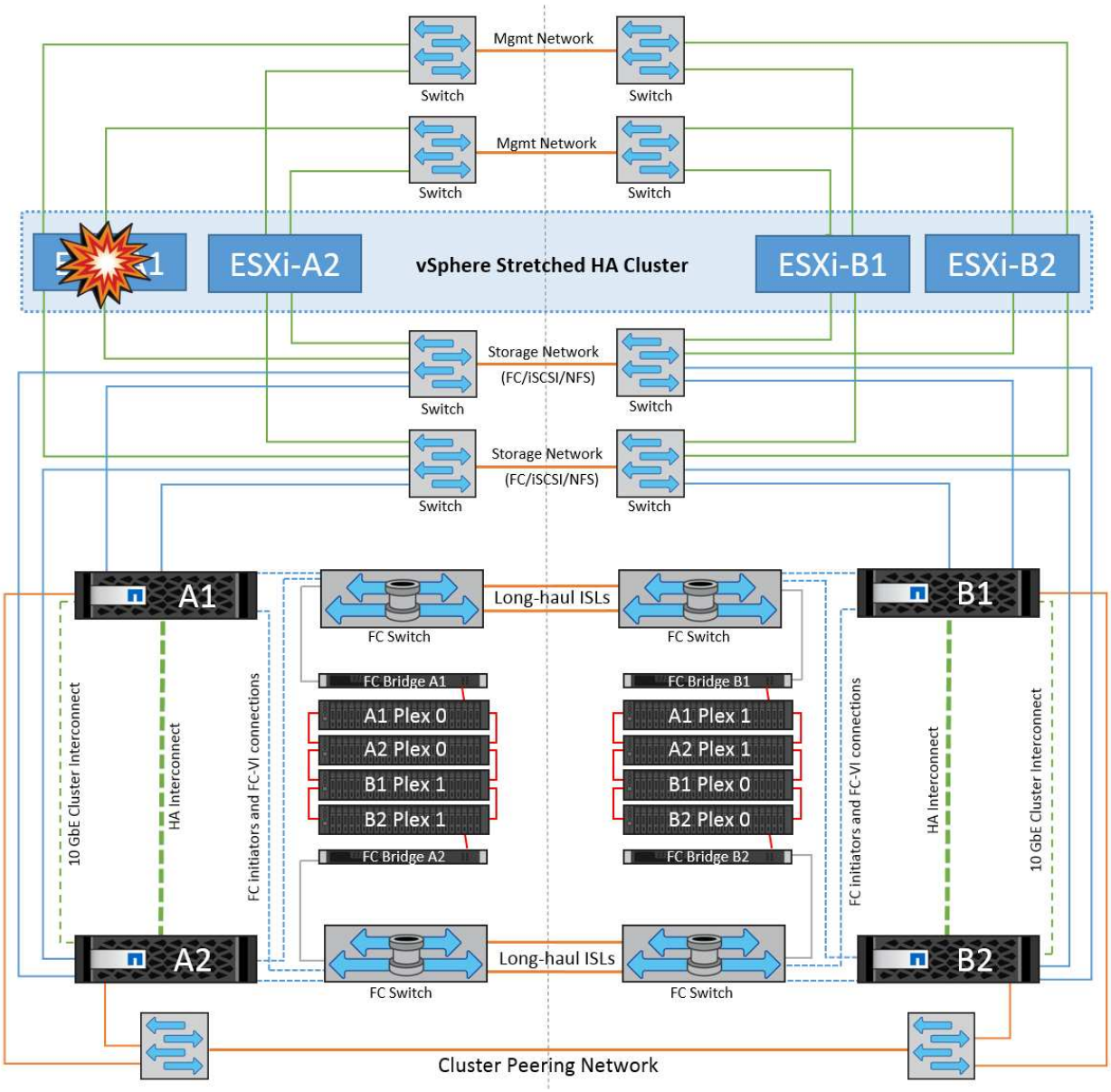
Also, for iSCSI LUNs, multipathing must be configured by binding the vmkernel interfaces to the iSCSI network adapters. For more information, refer to the vSphere storage documentation.

Best Practice

In environments in which Fibre Channel LUNs are used, NetApp recommends having at least two HBAs, which guarantees resiliency at the HBA/port level. NetApp also recommends single initiator to single target zoning as the best practice to configure zoning.

Virtual Storage Console (VSC) should be used to set multipathing policies because it sets policies for all new and existing NetApp storage devices.

Single ESXi Host Failure



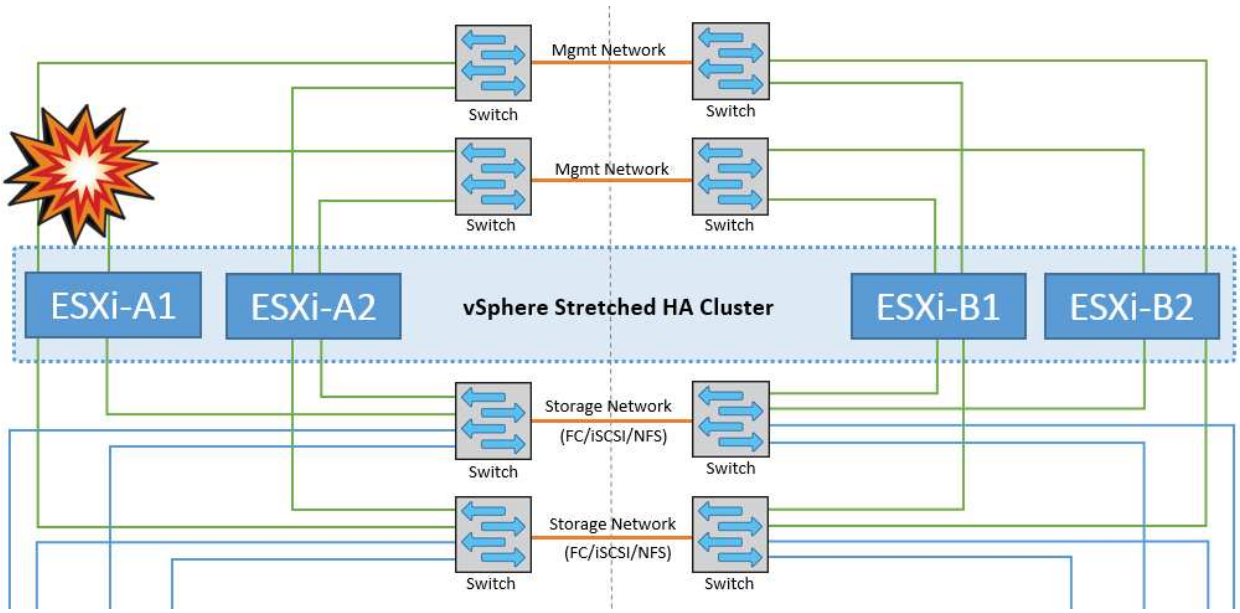
In this scenario, if there is an ESXi host failure, the master node in the VMware HA cluster detects the host failure since it no longer receives network heartbeats. To determine whether the host is really down or only a network partition, the master node monitors the datastore heartbeats and, if they are absent, it performs a final check by pinging the management IP addresses of the failed host. If all these checks are negative, then the master node declares this host a failed host and all the virtual machines that were running on this failed host are rebooted on the surviving host in the cluster.

If DRS VM and host affinity rules have been configured (VMs in VM group sitea_vms should run hosts in host group sitea_hosts), then the HA master first checks for available resources at site A. If there are no available hosts at site A, the master attempts to restart the VMs on hosts at site B.

It is possible that the virtual machines will be started on the ESXi hosts at the other site if there is a resource constraint in the local site. However, the defined DRS VM and host affinity rules will correct if any rules are violated by migrating the virtual machines back to any surviving ESXi hosts in the local site. In cases in which DRS is set to manual, NetApp recommends invoking DRS and applying the recommendations to correct the virtual machine placement.

There is no change in the MetroCluster behavior in this scenario and all the datastores continue to be intact from their respective sites.

ESXi Host Isolation

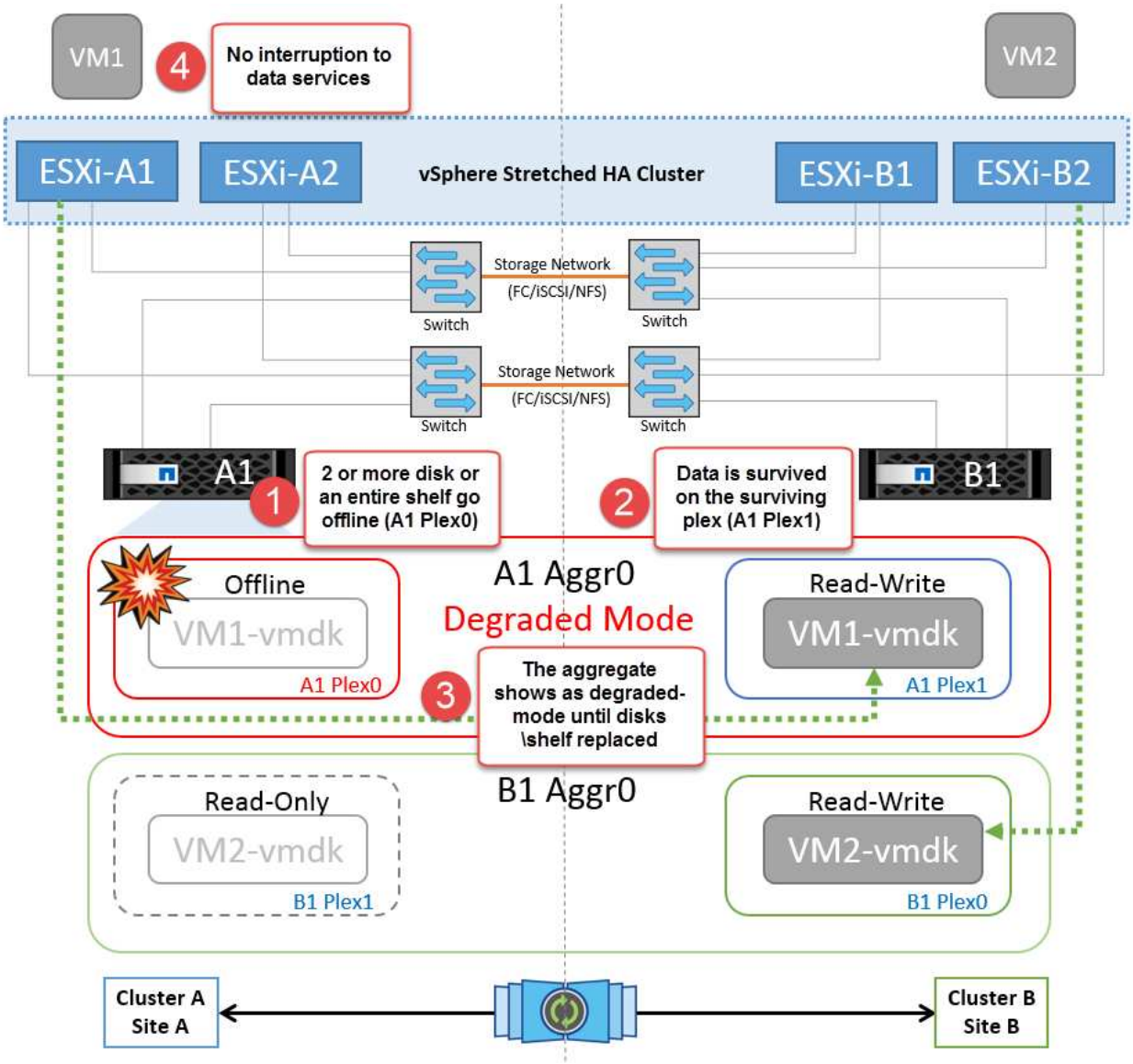


In this scenario, if the management network of the ESXi host is down, the master node in the HA cluster will not receive any heartbeats, and thus this host becomes isolated in the network. To determine whether it has failed or is only isolated, the master node starts monitoring the datastore heartbeat. If it is present then the host is declared isolated by the master node. Depending on the isolation response configured, the host may choose to power off, shut down the virtual machines, or even leave the virtual machines powered on. The default interval for the isolation response is 30 seconds.

There is no change in the MetroCluster behavior in this scenario and all the datastores continue to be intact from their respective sites.

Disk Shelf Failure

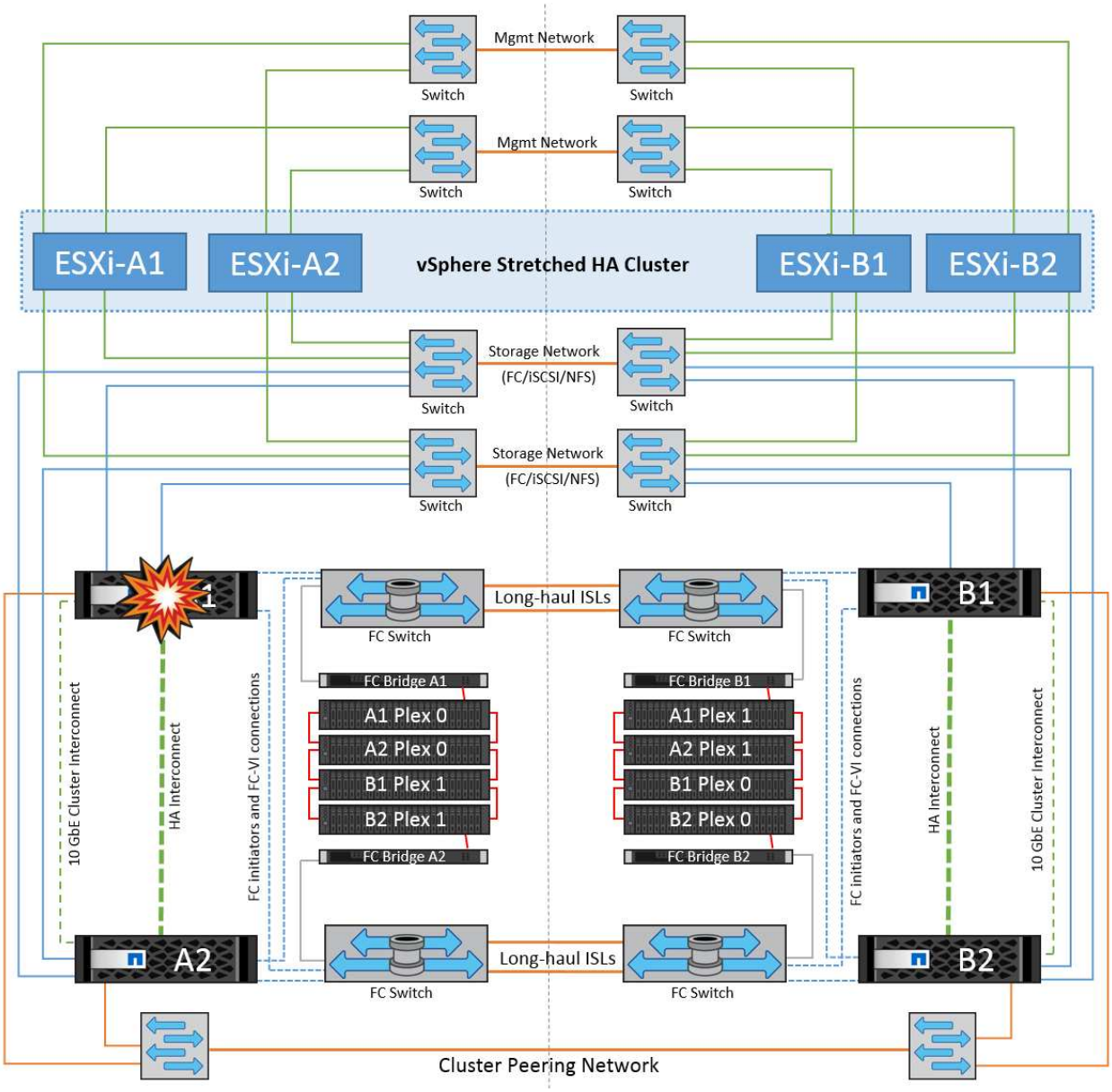
In this scenario, there is a failure of more than two disks or an entire shelf. Data is served from the surviving plex with no interruption to data services. The disk failure could affect either a local or remote plex. The aggregates will show as degraded mode because only one plex is active. Once the failed disks are replaced, the affected aggregates will automatically resync to rebuild the data. After resync, the aggregates will return automatically to normal mirrored mode. If more than two disks within a single RAID group have failed, then the plex has to be rebuilt from scratch.



Note: During this period, there is no impact on the virtual machine I/O operations, but there is degraded performance because the data is being accessed from the remote disk shelf through ISL links.

Single Storage Controller Failure

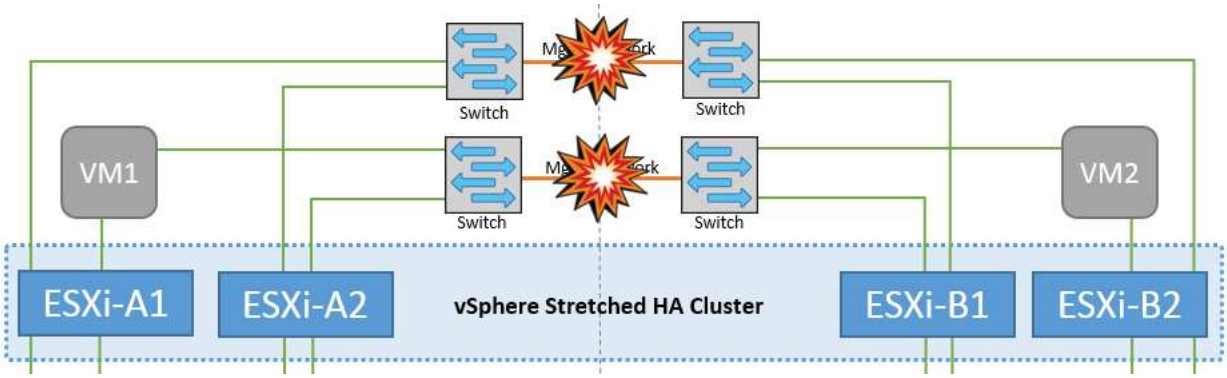
In this scenario, one of the two storage controllers fails at one site. Because there is an HA pair at each site, a failure of one node transparently and automatically triggers failover to the other node. For example, if node A1 fails, its storage and workloads are automatically transferred to node A2. Virtual machines will not be affected because all plexes remain available. The second site nodes (B1 and B2) are unaffected. In addition, vSphere HA will not take any action because the master node in the cluster will still be receiving the network heartbeats.



If the failover is part of a rolling disaster (node A1 fails over to A2), and there is a subsequent failure of A2, or the complete failure of site A, switchover following a disaster can occur at site B.

Interswitch Link Failures

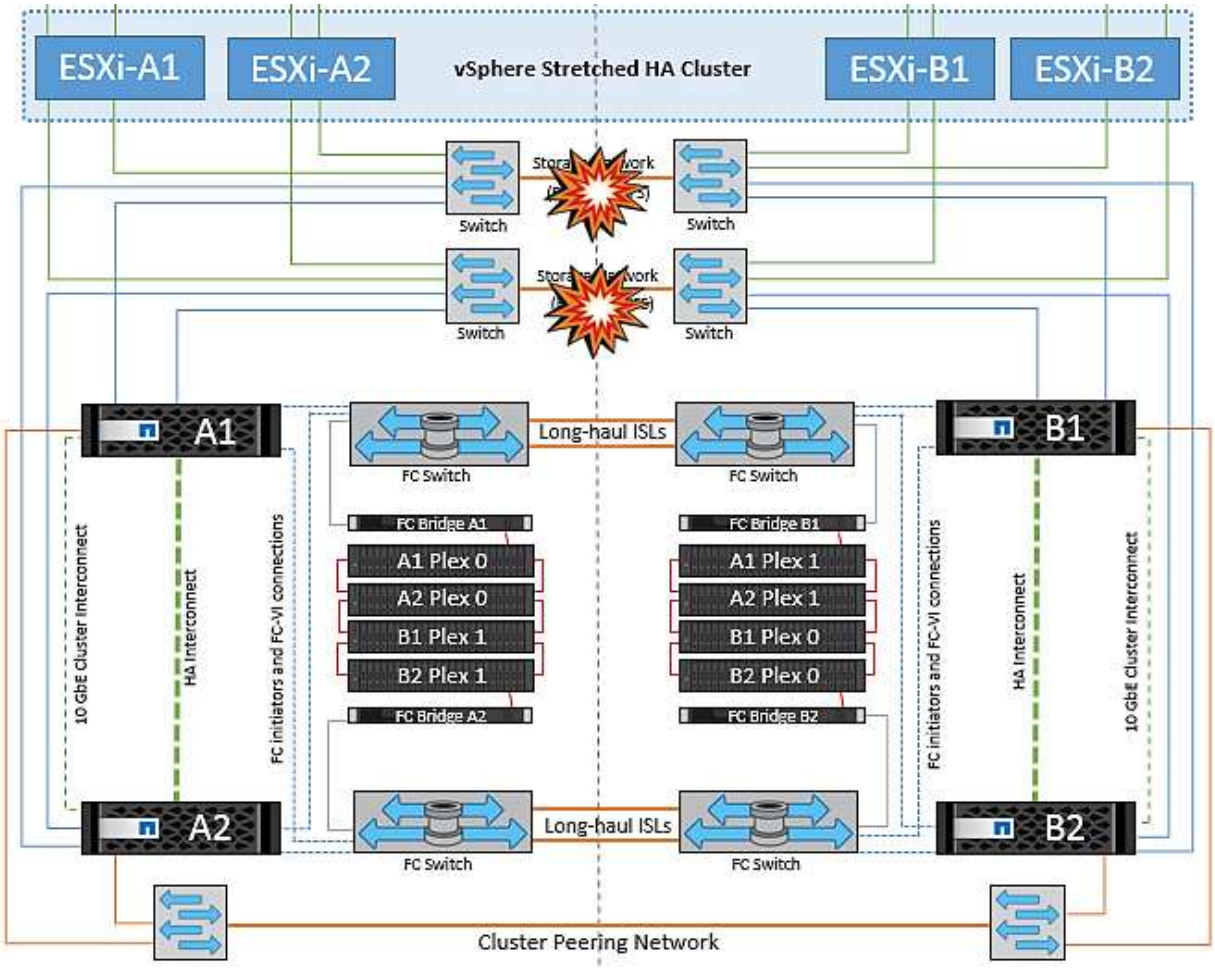
Interswitch Link Failure at Management Network



In this scenario, if the ISL links at the front-end host management network fail, the ESXi hosts at site A will not be able to communicate with ESXi hosts at site B. This will lead to a network partition because ESXi hosts at a particular site will be unable to send the network heartbeats to the master node in the HA cluster. As such, there will be two network segments because of partition and there will be a master node in each segment that will protect VMs from host failures within the particular site.

Note: During this period, the virtual machines remain running and there is no change in the MetroCluster behavior in this scenario. All the datastores continue to be intact from their respective sites.

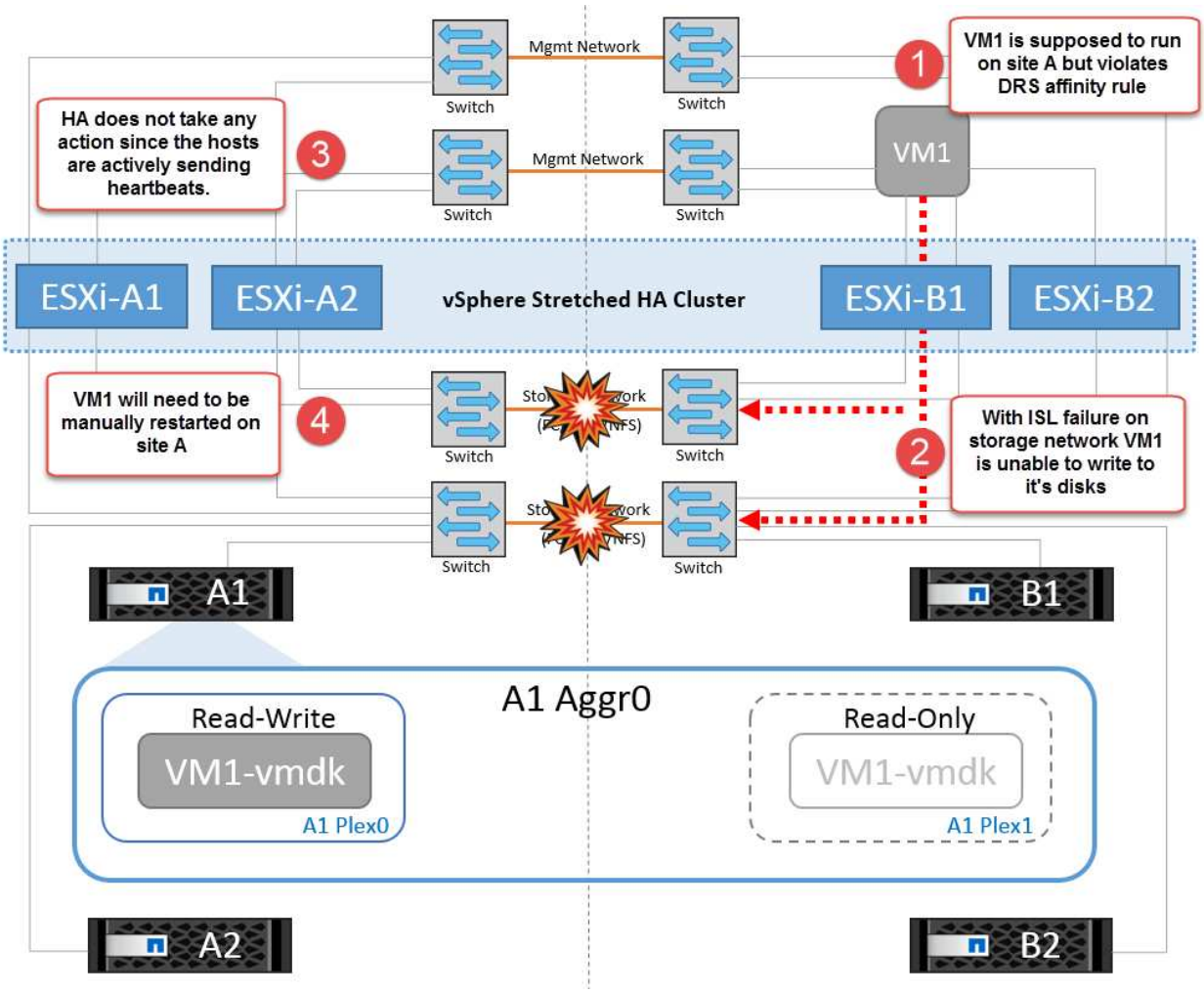
Interswitch Link Failure at Storage Network



In this scenario, if the ISL links at the backend storage network fail, the hosts at site A will lose access to the storage volumes or LUNs of cluster B at site B and vice versa. The VMware DRS rules are defined so that host-storage site affinity facilitates the virtual machines to run without impact within the site.

During this period, the virtual machines remain running in their respective sites and there is no change in the MetroCluster behavior in this scenario. All the datastores continue to be intact from their respective sites.

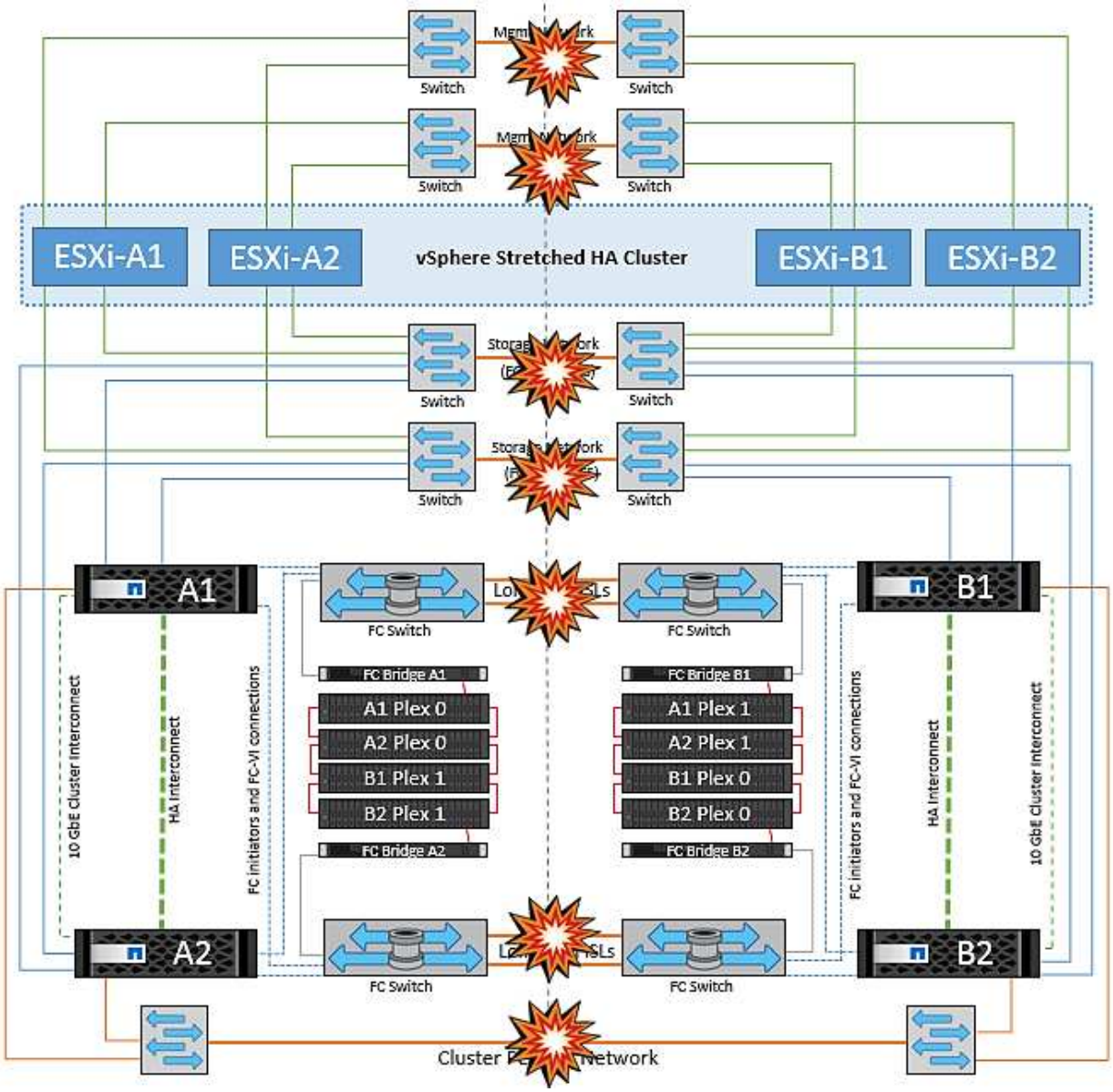
If for some reason the affinity rule was violated (for example, VM1, which was supposed to run from site A where its disks reside on local cluster A nodes, is running on a host at site B), the virtual machine's disk will be remotely accessed via ISL links. Because of ISL link failure, VM1 running at site B would not be able to write to its disks because the paths to the storage volume are down and that particular virtual machine is down. In these situations, VMware HA does not take any action since the hosts are actively sending heartbeats. Those virtual machines need to be manually powered off and powered on in their respective sites. The following figure illustrates a VM violating a DRS affinity rule.



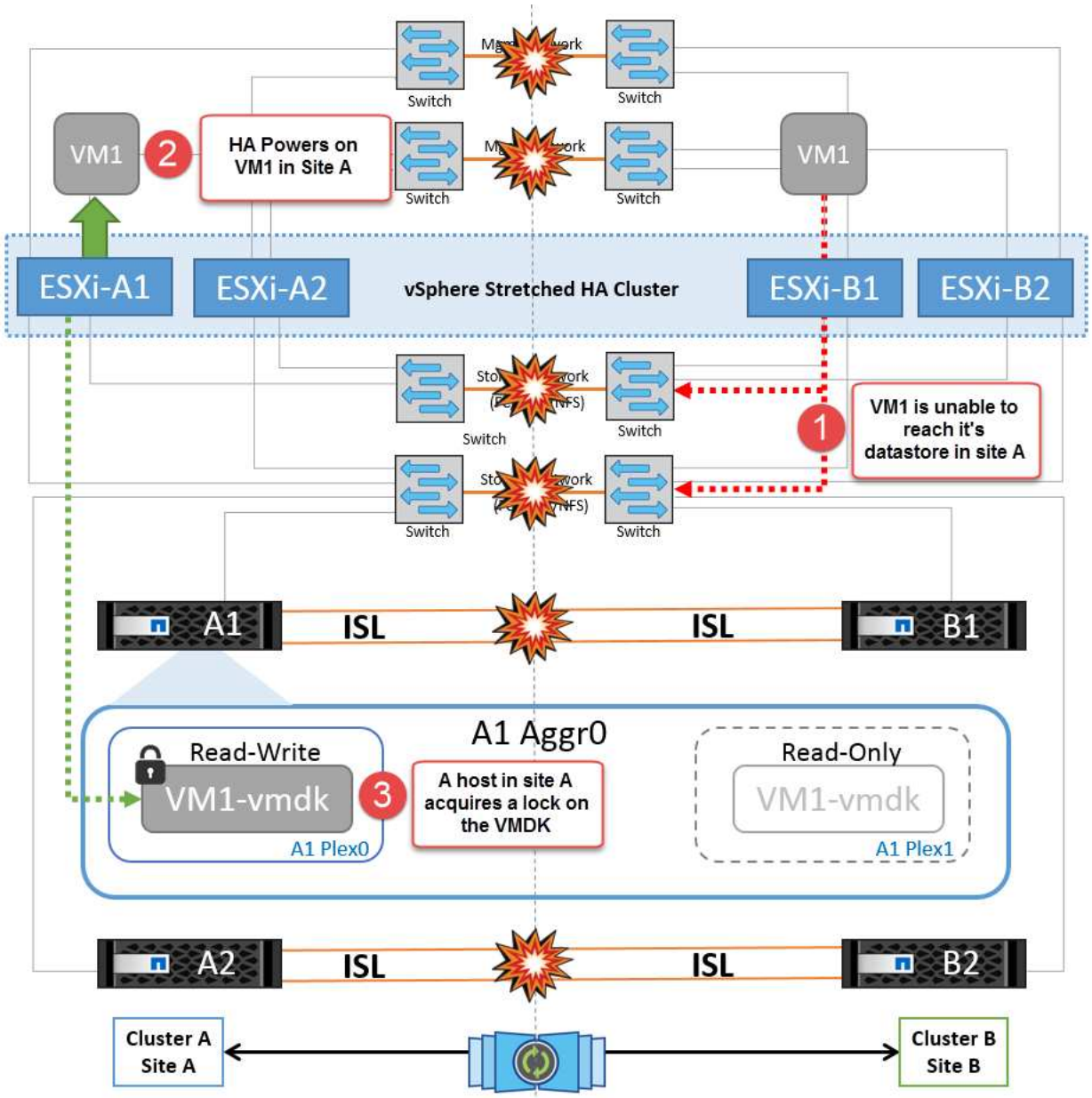
All Interswitch Failure or Complete Data Center Partition

In this scenario, all the ISL links between the sites are down and both the sites are isolated from each other. As discussed in earlier scenarios, such as ISL failure at the management network and at the storage network, the virtual machines are not affected in complete ISL failure.

After ESXi hosts are partitioned between sites, the vSphere HA agent will check for datastore heartbeats and, in each site, the local ESXi hosts will be able to update the datastore heartbeats to their respective read-write volume/LUN. Hosts in site A will assume that the other ESXi hosts at site B have failed because there are no network/datastore heartbeats. vSphere HA at site A will try to restart the virtual machines of site B, which will eventually fail because the datastores of site B will not be accessible due to storage ISL failure. A similar situation is repeated in site B.



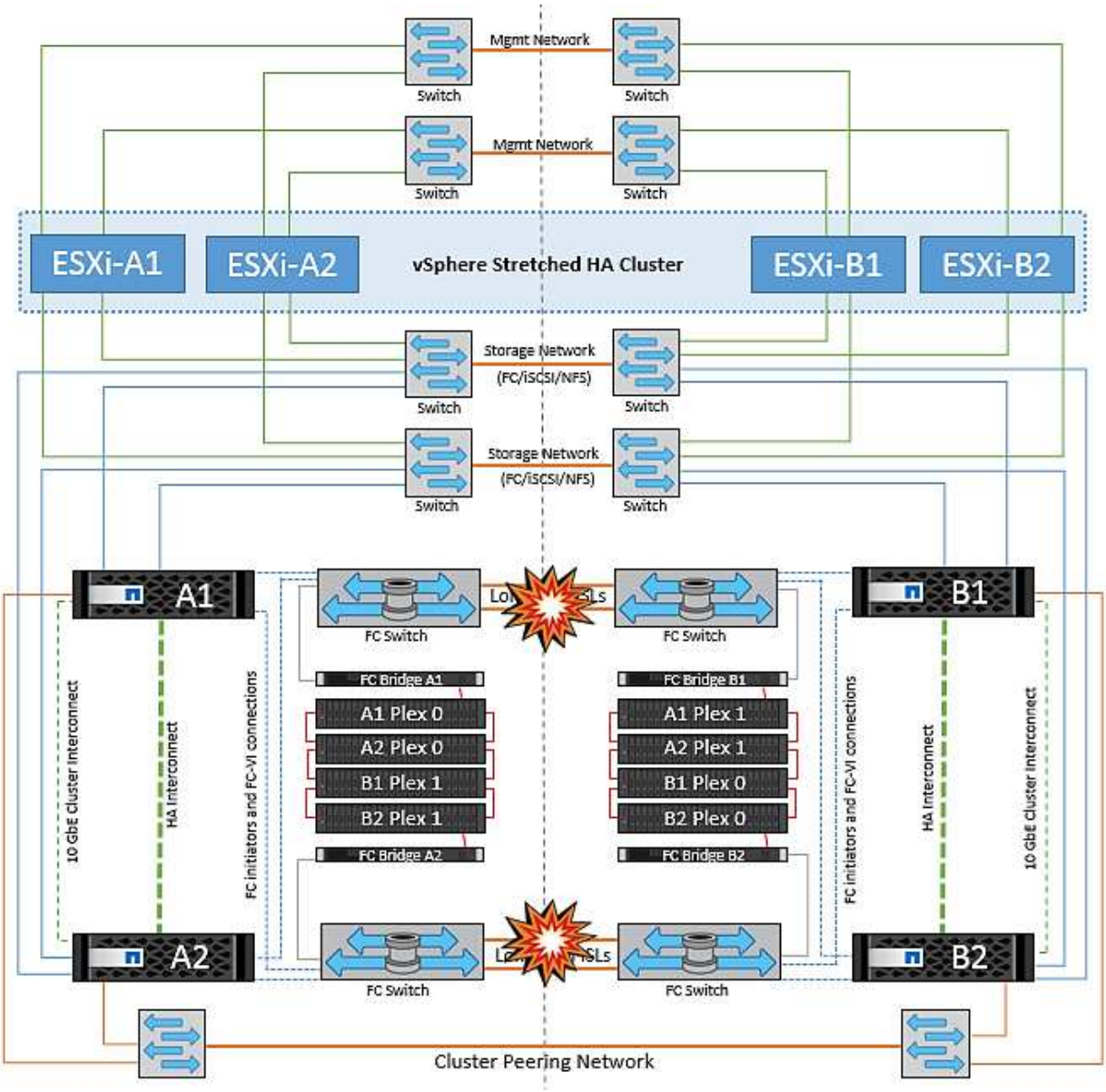
NetApp recommends determining if any virtual machine has violated the DRS rules. Any virtual machines running from a remote site will be down since they will not be able to access the datastore, and vSphere HA will restart that virtual machine on the local site. After the ISL links are back online, the virtual machine that was running in the remote site will be killed, since there cannot be two instances of virtual machines running with the same MAC addresses.



Interswitch Link Failure on Both Fabrics in NetApp MetroCluster

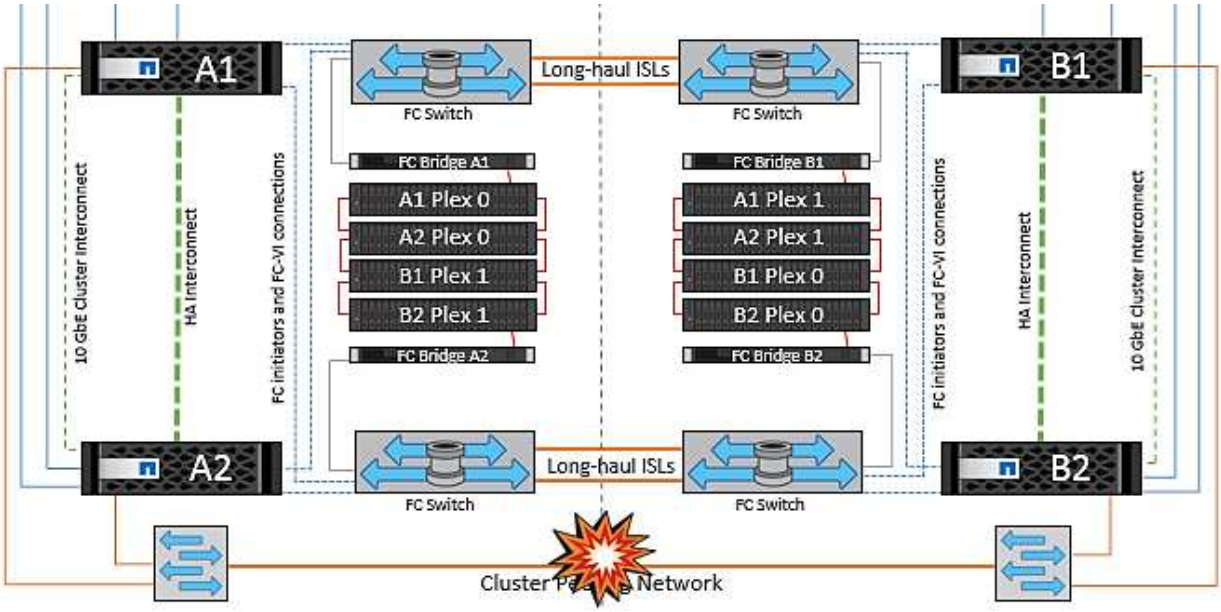
In a scenario of one or more ISLs failing, traffic continues through the remaining links. If all ISLs on both fabrics fail, such that there is no link between the sites for storage and NVRAM replication, each controller will continue to serve its local data. On restoration of a minimum of one ISL, resynchronization of all the plexes will happen automatically.

Any writes occurring after all ISLs are down will not be mirrored to the other site. A switchover on disaster, while the configuration is in this state, would therefore incur loss of the data that had not been synchronized. In this case, manual intervention is required for recovery after the switchover. If it is likely that no ISLs will be available for an extended period, an administrator can choose to shut down all data services to avoid the risk of data loss if a switchover on disaster is necessary. Performing this action should be weighed against the likelihood of a disaster requiring switchover before at least one ISL becomes available. Alternatively, if ISLs are failing in a cascading scenario, an administrator could trigger a planned switchover to one of the sites before all the links have failed.



Peered Cluster Link Failure

In a peered cluster link failure scenario, because the fabric ISLs are still active, data services (reads and writes) continue at both sites to both plexes. Any cluster configuration changes (for example, adding a new SVM, provisioning a volume or LUN in an existing SVM) cannot be propagated to the other site. These are kept in the local CRS metadata volumes and automatically propagated to the other cluster upon restoration of the peered cluster link. If a forced switchover is necessary before the peered cluster link can be restored, outstanding cluster configuration changes will be replayed automatically from the remote replicated copy of the metadata volumes at the surviving site as part of the switchover process.



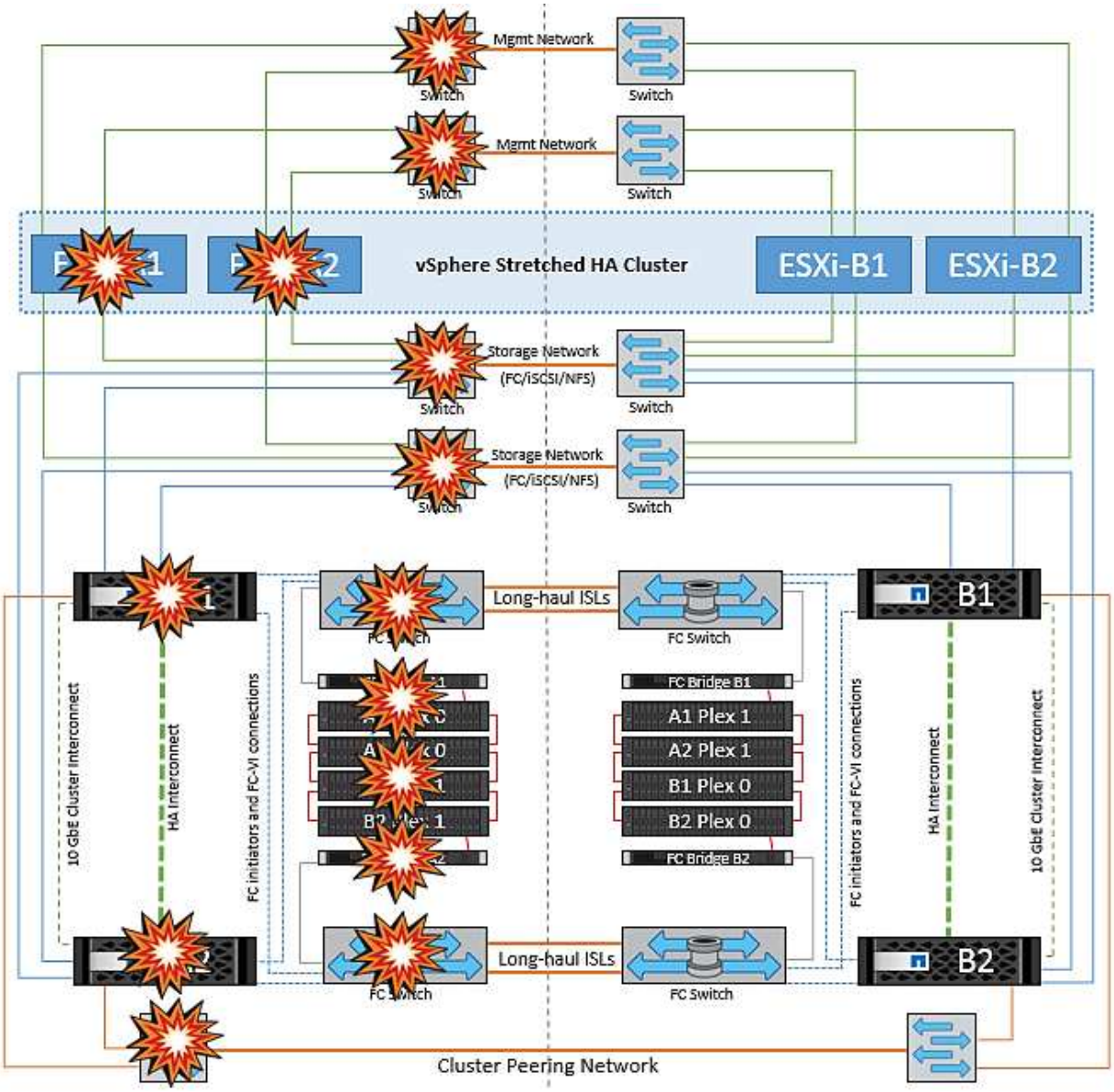
Complete Site Failure

In a complete site A failure scenario, the ESXi hosts at site B will not get the network heartbeat from the ESXi hosts at site A because they are down. The HA master at site B will verify that the datastore heartbeats are not present, declare the hosts at site A to be failed, and try to restart the site A virtual machines in site B. During this period, the storage administrator performs a switchover to resume services of the failed nodes on the surviving site which will restore all the storage services of site A at site B. After the site A volumes or LUNs are available at site B, the HA master agent will attempt to restart the site A virtual machines in site B.

If the vSphere HA master agent's attempt to restart a VM (which involves registering it and powering it on) fails, the restart is retried after a delay. The delay between restarts can be configured to up to a maximum of 30 minutes. vSphere HA attempts these restarts for a maximum number of attempts (six attempts by default).

Note: The HA master does not begin the restart attempts until the placement manager finds suitable storage, so in the case of a complete site failure, that would be after the switchover has been performed.

If site A has been switched over, a subsequent failure of one of the surviving site B nodes can be seamlessly handled by failover to the surviving node. In this case, the work of four nodes is now being performed by only one node. Recovery in this case would consist of performing a giveback to the local node. Then, when site A is restored, a switchback operation is performed to restore steady state operation of the configuration.



Product Security

ONTAP tools for VMware vSphere

Software engineering with ONTAP Tools for VMware vSphere employs the following secure development activities:

- **Threat modeling.** The purpose of threat modelling is to discover security flaws in a feature, component, or product early in the software development life cycle. A threat model is a structured representation of all the information that affects the security of an application. In essence, it is a view of the application and its environment through the lens of security.
- **Dynamic Application Security Testing (DAST).** This technology is designed to detect vulnerable conditions on applications in their running state. DAST tests the exposed HTTP and HTML interfaces of web-enable applications.
- **Third-party code currency.** As part of software development with open-source software (OSS), you must address security vulnerabilities that might be associated with any OSS incorporated into your product. This is a continuing effort because a new OSS version might have a newly discovered vulnerability reported at

any time.

- **Vulnerability scanning.** The purpose of vulnerability scanning is to detect common and known security vulnerabilities in NetApp products before they are released to customers.
- **Penetration testing.** Penetration testing is the process of evaluating a system, web application, or network to find security vulnerabilities that could be exploited by an attacker. Penetration tests (pen tests) at NetApp are conducted by a group of approved and trusted third-party companies. Their testing scope includes the launching of attacks against an application or software similar to hostile intruders or hackers using sophisticated exploitation methods or tools.

Product security features

ONTAP tools for VMware vSphere includes the following security features in each release.

- **Login banner.** SSH is disabled by default and only allows one-time logins if enabled from the VM console. The following login banner is shown after the user enters a username in the login prompt:

  WARNING: Unauthorized access to this system is forbidden and will be prosecuted by law. By accessing this system, you agree that your actions may be monitored if unauthorized usage is suspected.

  After the user completes login through the SSH channel, the following text is displayed:

```
Linux vsc1 4.19.0-12-amd64 #1 SMP Debian 4.19.152-1 (2020-10-18) x86_64
The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.
Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
```

- **Role-based access control (RBAC).** Two kinds of RBAC controls are associated with ONTAP tools:
  - Native vCenter Server privileges
  - vCenter plug-in specific privileges. For details, see this link.
- **Encrypted communications channels.** All external communication happens over HTTPS using version 1.2 of TLS.
- **Minimal port exposure.** Only the necessary ports are open on the firewall.

  The following table describes the open port details.

TCP v4/v6 port #	Direction	Function
8143	inbound	HTTPS connections for REST API
8043	inbound	HTTPS connections
9060	inbound	HTTPS connections Used for SOAP over https connections This port must be opened to allow a client to connect to the ONTAP tools API server.

TCP v4/v6 port #	Direction	Function
22	inbound	SSH (Disabled by default)
9080	inbound	HTTPS connections - VP and SRA - Internal connections from loopback only
9083	inbound	HTTPS connections - VP and SRA Used for SOAP over https connections
1162	inbound	VP SNMP trap packets
1527	internal only	Derby database port, only between this computer and itself, external connections not accepted — Internal connections only
443	bi-directional	Used for connections to ONTAP clusters

- **Support for certificate authority (CA) signed certificates.** ONTAP tools for VMware vSphere supports CA signed certificates. See this kb article for more information.
- **Audit logging.** Support bundles can be downloaded and are extremely detailed. ONTAP tools logs all user login and logout activity in a separate log file. VASA API calls are logged in a dedicated VASA audit log (local cxf.log).
- **Password policies.** The following password policies are followed:
  - Passwords are not logged in any log files.
  - Passwords are not communicated in plain text.
  - Passwords are configured during the installation process itself.
  - Password history is a configurable parameter.
  - Minimum password age is set to 24 hours.
  - Auto complete for the password fields are disabled.
  - ONTAP tools encrypts all stored credential information using SHA256 hashing.

SnapCenter Plug-in VMware vSphere

NetApp SnapCenter Plug-in for VMware vSphere software engineering uses the following secure development activities:

- **Threat modeling.** The purpose of threat modelling is to discover security flaws in a feature, component, or product early in the software development life cycle. A threat model is a structured representation of all the information that affects the security of an application. In essence, it is a view of the application and its environment through the lens of security.
- **Dynamic application security testing (DAST).** Technologies that are designed to detect vulnerable conditions on applications in their running state. DAST tests the exposed HTTP and HTML interfaces of web-enable applications.
- **Third-party code currency.** As part of developing software and using open-source software (OSS), it is

important to address security vulnerabilities that might be associated with OSS that has been incorporated into your product. This is a continuous effort as the version of the OSS component may have a newly discovered vulnerability reported at any time.

- **Vulnerability scanning.** The purpose of vulnerability scanning is to detect common and known security vulnerabilities in NetApp products before they are released to customers.
- **Penetration testing.** Penetration testing is the process of evaluating a system, web application or network to find security vulnerabilities that could be exploited by an attacker. Penetration tests (pen tests) at NetApp are conducted by a group of approved and trusted third-party companies. Their testing scope includes the launching of attacks against an application or software like hostile intruders or hackers using sophisticated exploitation methods or tools.
- **Product Security Incident Response activity.** Security vulnerabilities are discovered both internally and externally to the company and can pose a serious risk to NetApp's reputation if they are not addressed in a timely manner. To facilitate this process, a Product Security Incident Response Team (PSIRT) reports and tracks the vulnerabilities.

Product security features

NetApp SnapCenter Plug-in for VMware vSphere includes the following security features in each release:

- **Restricted shell access.** SSH is disabled by default, and one-time logins are only allowed if they are enabled from the VM console.
- **Access warning in login banner.** The following login banner is shown after the user enters a user name in the login prompt:

  WARNING: Unauthorized access to this system is forbidden and will be prosecuted by law. By accessing this system, you agree that your actions may be monitored if unauthorized usage is suspected.

  After the user completes login through the SSH channel, the following output displays:

```
Linux vsc1 4.19.0-12-amd64 #1 SMP Debian 4.19.152-1 (2020-10-18) x86_64
The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.
Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
```

- **Role-based access control (RBAC).** Two kinds of RBAC controls are associated with ONTAP tools:
  - Native vCenter Server privileges.
  - VMware vCenter plug-in specific privileges. For more information, see Role-Based Access Control (RBAC).
- **Encrypted communications channels.** All external communication happens over HTTPS by using TLS.
- **Minimal port exposure.** Only the necessary ports are open on the firewall.

The following table provides the open port details.

TCP v4/v6 port number	Function
8144	HTTPS connections for REST API

TCP v4/v6 port number	Function
8080	HTTPS connections for OVA GUI
22	SSH (disabled by default)
3306	MySQL (internal connections only; external connections disabled by default)
443	Nginx (data protection services)

- **Support for Certificate Authority (CA) signed certificates.** SnapCenter Plug-in for VMware vSphere supports the feature of CA signed certificates. See How to create and/or import an SSL certificate to SnapCenter Plug-in for VMware vSphere (SCV).
- **Password policies.** The following password policies are in effect:
  - Passwords are not logged in any log files.
  - Passwords are not communicated in plain text.
  - Passwords are configured during the installation process itself.
  - All credential information is stored using SHA256 hashing.
- **Base operating system image.** The product ships with Debian Base OS for OVA with restricted access and shell access disabled. This reduces the attack footprint. Every SnapCenter release base operating system is updated with latest security patches available for maximum security coverage.

NetApp develops software features and security patches with regards to SnapCenter Plug-in for VMware vSphere appliance and then releases them to customers as a bundled software platform. Because these appliances include specific Linux sub-operating system dependencies as well as our proprietary software, NetApp recommends that you do not make changes to the sub-operating system because this has a high potential to affect the NetApp appliance. This could affect the ability of NetApp to support the appliance. NetApp recommends testing and deploying our latest code version for appliances because they are released to patch any security-related issues.

Legal notices

Legal notices provide access to copyright statements, trademarks, patents, and more.

Copyright

https://www.netapp.com/company/legal/copyright/

Trademarks

NETAPP, the NETAPP logo, and the marks listed on the NetApp Trademarks page are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.

https://www.netapp.com/company/legal/trademarks/

Patents

A current list of NetApp owned patents can be found at:

https://www.netapp.com/pdf.html?item=/media/11887-patentspage.pdf

Privacy policy

https://www.netapp.com/company/legal/privacy-policy/

Open source

Notice files provide information about third-party copyright and licenses used in NetApp software.

ONTAP

Notice for ONTAP 9.13.1
Notice for ONTAP 9.12.1
Notice for ONTAP 9.12.0
Notice for ONTAP 9.11.1
Notice for ONTAP 9.10.1
Notice for ONTAP 9.10.0
Notice for ONTAP 9.9.1
Notice for ONTAP 9.8
Notice for ONTAP 9.7
Notice for ONTAP 9.6
Notice for ONTAP 9.5
Notice for ONTAP 9.4
Notice for ONTAP 9.3
Notice for ONTAP 9.2
Notice for ONTAP 9.1

ONTAP Mediator for MCC IP

Copyright information

Copyright © 2024 NetApp, Inc. All Rights Reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

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

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

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

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

LIMITED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (b)(3) of the Rights in Technical Data -Noncommercial Items at DFARS 252.227-7013 (FEB 2014) and FAR 52.227-19 (DEC 2007).

Data contained herein pertains to a commercial product and/or commercial service (as defined in FAR 2.101) and is proprietary to NetApp, Inc. All NetApp technical data and computer software provided under this Agreement is commercial in nature and developed solely at private expense. The U.S. Government has a non-exclusive, non-transferrable, nonsublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b) (FEB 2014).

Trademark information

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