Implementing vSphere Metro Storage Cluster using HPE 3PAR Peer Persistence
Contents

Introduction ............................................................................................................... 4
Terminology ................................................................................................................. 4
Features and benefits ................................................................................................. 4
Requirements .............................................................................................................. 5
  Host Persona 11 ........................................................................................................ 5
  Verify each VV has same WWN on both arrays ....................................................... 6
  Verify VLUN status on both arrays ........................................................................ 6
  Quorum Witness ........................................................................................................ 6
  Management port on HPE 3PAR StoreServ ............................................................. 7
  Support and coexistence of different cluster platforms with 3PAR Peer Persistence 7
VMware vSphere Metro Storage Cluster (vMSC) Configuration ..................................... 7
  vMSC uniform .......................................................................................................... 7
  vMSC non-uniform .................................................................................................. 7
Planned switchover .................................................................................................... 8
HPE 3PAR Peer Persistence failures handling ............................................................. 9
  Array to Array communication failure .................................................................. 9
  Single site to QW communication failure ............................................................... 9
  Site 1 to QW and Array to Array communication failure ........................................ 9
  Site 2 to QW and Array to Array communication failure ........................................ 9
  Site 1 and site 2 both lose communications to the QW ............................................ 9
  Site 1 and site 2 to QW and Array to Array communication failure ....................... 9
VMware vMSC and HPE 3PAR Peer Persistence best practices ..................................... 10
  VMware DRS Affinity Group Usage ...................................................................... 10
  Managing VMware Storage DRS settings .............................................................. 10
  Workload management across sites ....................................................................... 10
  Multi-VM application placement ............................................................................ 11
  Heartbeat Setting .................................................................................................... 11
  VMware HA Admission Control ........................................................................... 11
  Isolation addresses & false positive ........................................................................ 11
  Permanent Device Loss (PDL) and All Paths Down (APD) ...................................... 11
  Response for APD recovery after APD timeout ...................................................... 12
  HPE 3PAR Quorum Witness Hosting .................................................................... 12
  Remote Copy Replication links .............................................................................. 12
  HPE 3PAR Peer Persistence maximums ................................................................. 12
Introduction

HPE 3PAR StoreServ Peer Persistence software enables HPE 3PAR StoreServ systems located in different data centers at metropolitan distances to act as peers to each other, presenting continuous storage system access to hosts connected to them. This capability allows you to configure a high-availability solution between two sites or data centers where storage access failover and failback remains completely transparent to the hosts and applications running on those hosts. Compared to traditional cluster failover models where upon failover, the applications must be restarted, Peer Persistence software allows hosts to remain online serving their business applications even when they switch storage access from their original site to the storage array in the remote site, resulting in a much improved recovery time.

This paper provides information about deploying a vSphere Metro Storage Cluster (vMSC) across two data centers or sites using HPE 3PAR StoreServ Peer Persistence on StoreServ Storage systems. A vMSC configuration is designed to maintain data availability beyond a single physical or logical site.

Terminology

Throughout this paper, we identify the volume that is part of a primary remote copy group as source volume. The replicated volume in the secondary remote copy group is called the target volume. Volumes created on an HPE 3PAR StoreServ are called Virtual Volumes, or VVs. A VV exported to the host is known as a VLUN. A zoning operation creates a logical connection in the Fibre Channel (FC) SAN between an FC host bus adapter (HBA) on a server and one on a storage system.

Features and benefits

Peer Persistence is a high availability configuration between two data centers in which the ESXi hosts are setup in a metro cluster configuration with access to storage arrays in both sites. Storage volumes created on one storage array are replicated to the other array using synchronous remote copy to ensure that the volumes are in sync at all times. Peer Persistence software takes advantage of the Asymmetric Logical Unit Access (ALUA) capability that allows paths to a SCSI device to be marked as having different characteristics. With ALUA the same LUN can be exported from both arrays simultaneously, but only the paths to the side accepting write to the volume will be marked as active. The paths to the secondary side volume will be marked as standby preventing the host from performing any I/O using those paths. In the event of a non-disruptive array volume migration scenario, the standby paths are marked as active and host traffic to the primary storage array is redirected to the secondary storage array without impact to the hosts. Figure 1 shows a recommended Peer Persistence configuration.

Figure 1. Peer Persistence setup with a Uniform vMSC configuration
The Peer Persistence software supports bi-directional replication which enables each site to perform as a disaster recovery center for the other. It enables you to move your applications from one site to another based on your business and performance needs without impact on the applications running on those hosts.

An example would be the use of vMotion within a VMware vSphere environment. Peer Persistence provides the same stretched cluster configuration across data centers that are used with VMware's vSphere Metro Storage Cluster certification to enable multi-site storage high availability. In figure 1, a few virtual machines (VMs) are being serviced by a 3PAR StoreServ storage system on site 1 while other VMs are being serviced by another 3PAR StoreServ storage system at site 2 located within metropolitan distances from one another. vMotion allows customers to move VMs across sites. Peer Persistence software enables the movement of VMs across data centers to be completely transparent to the applications those VMs are running.

Peer Persistence achieves automatic transparent failover in the event of a complete array or data center failure with the help of the Quorum Witness software. The Quorum Witness software needs to be installed at a third site and will act as an independent witness should one of the arrays or data centers become unavailable.

Requirements

- Firmware version on both HPE 3PAR StoreServ arrays must be 3.12 MU2 or newer.
- The sites must be set up in a Remote Copy 1-to-1 configuration in synchronous mode.
- The round trip latency between the two sites must be 2.6 milliseconds or less. For 3PAR OS 3.2.2, the round trip latency is 5 ms or less.
- Quorum Witness software must be installed on a virtual machine at a third site.
- The HPE 3PAR StoreServ arrays communicate with the Quorum Witness via the management port over TCP/IP. The Quorum witness needs to be reachable from both sites.
- This non-disruptive failover configuration is supported with VMware ESXi 5.x or newer.
- All source and target volumes exported from the two HPE 3PAR StoreServ arrays must have the same volume WWN.
- Hosts have to be connected to the HPE StoreServ storage system using FC, iSCSI, or FCoE. Support for the iSCSI and FCoE host connectivity in HPE 3PAR Peer Persistence has been added in later version of HPE 3PAR StoreServ OS. Please refer to the Single Point of Connectivity Knowledge (SPOCK) for more details.
- All associated hosts are connected to both arrays (Uniform vMSC).
- The ESXi hosts that these volumes are exported to must be configured with host persona 11 (VMware), which supports host capability Asymmetric Logical Unit Access (ALUA) configuration.

As certain requirements change due to new software releases or supported configurations as a result of testing and qualifications, it is extremely important to check the latest supported configuration and requirement for 3PAR Peer Persistence in a vSphere environment.

Please refer to the Single Point of Connectivity Knowledge (SPOCK).

Host Persona 11

Host personas are a set of behaviors that permit hosts connected to FC or iSCSI ports on the system to deviate from the default host behavior. By assigning a persona to a host, multiple host types that require distinct customized responses can share a single system port. For example, hosts running Windows®, Linux®, and AIX operating systems can all connect to the same system port. This simplifies connecting hosts to the system and reduces management costs related to complex host connections.

Persona 11 supports ALUA path management and it must be used for host configurations that have ESXi hosts connected to both the primary and the secondary. Existing ESXi hosts defined with host persona 6 should be changed to persona 11 if these hosts are connected to both primary and secondary array.
For host configurations that have independent ESXi hosts attached to the primary or secondary arrays, host persona 6 (Generic-Legacy) should be configured. This persona does not support Asymmetric Logical Unit Access (ALUA) path management.

For more information on host persona migration, please refer to the HPE 3PAR StoreServ Storage VMware ESX Host Persona Migration Guide (HPE Part Number: QL226-97875) hp.com/go/storage/docs.

**Verify each VV has same WWN on both arrays**

It's important that the source and the target volumes on the 3PAR StoreServ arrays have the same identity when both volumes are exported to the same ESXi host. Hence both volumes must have the same WWN. User can create volumes with the same WWN automatically using the new option in admitrcopyvv command. See the 3PAR StoreServ Copy Software User's Guide for instructions on how to use the admitrcopyvv command. The 3PAR StoreServ Management Console (SSMC) also give the users option to create VVs with the same WWN when configuring RC groups.

The replicated volume WWN can also be changed later if the volumes were created using the 3PAR StoreServ management console or any method other than the admitrcopyvv. See the 3PAR StoreServ Command Line Interface Administrator’s Manual or the 3PAR StoreServ Management Console Online Help for instructions on how to edit a VV’s WWN.

**Verify VLUN status on both arrays**

Once the volumes are exported from both arrays, the showvlun command will show the state as “active” for VLUNs exported from the source side and “standby” for the VLUNs exported from the target side. ESXi hosts will send IO requests only to the LUNs reported as active.

**Quorum Witness**

The 3PAR StoreServ Quorum Witness (QW) software enables transparent automatic failover between the two sites in a vMSC cluster. The Quorum Witness software gets regularly updated with status information from the 3PAR StoreServ arrays. In the event one of the 3PAR StoreServ arrays becomes inaccessible in case of power outage or similar event in a data center, the surviving 3PAR StoreServ array detects that the Quorum Witness is not getting updated by the failed array and initiates a failover operation on any secondary groups associated with that failed array.

**Note**

QW details have to be setup at the target level, and transparent failover will occur only for Remote Copy groups that use the target, are started and have the “auto_failover” policy set. Failover is not automatic for remote copy groups between the same two HPE 3PAR StoreServ systems which do not have the “auto_failover” policy set. For HPE 3PAR OS 3.13 or later, the setting of path management is required as well. A volume group must have the path management policy set if it is to be valid for automated and transparent failover. If this policy is not set, then no automatic failover will occur if a disaster strikes.

The Quorum Witness is typically set up on a third site where events that may impact site 1 or site 2 cannot impact the Quorum Witness site at the same time. Additionally, the QW connects to arrays on both sites using non-RC links. With the above two configuration characteristics (site and link independence), QW helps determine the following nature of failures:

- **Link failure**: The QW can detect if the two arrays are alive but not communicating because of a link failure. QW would still be receiving updates from both of the arrays.

- **Array/Site failure**: The QW can detect when one of the arrays/sites fails. QW would not be receiving updates from one of the array that has failed.

The “Handling failures” section covers the various failure conditions handled by Peer Persistence.
The Quorum Witness software is packaged as an Open Virtualization Format (OVF) package for deployment in a vSphere or a Hyper-V environment. The Quorum Witness virtual machine should not be provisioned on a datastore allocated from either of the arrays that are part of the Peer Persistence configuration. If the datastore is located on one of the arrays that is part of the Peer Persistence configuration failure of the array will make the datastore unavailable which will result in the shutdown of the Quorum Witness virtual machine. The Quorum Witness software uses port 8080 to communicate with the two 3PAR StoreServ systems in the quorum. Hence firewall rules need to be changed to open port 8080 at the QW site.

The Quorum Witness is a passive component of the configuration and the software itself does not provide any high availability, however the virtual machine can be made highly available by using VMware—High Availability or Fault Tolerance, or Microsoft® Failover Clustering for the Hyper-V based QW package. For the latest supported configuration and requirement for HPE 3PAR Peer Persistence QW, refer to the Single Point of Connectivity Knowledge (SPOCK).

Management port on HPE 3PAR StoreServ
The HPE 3PAR StoreServs communicate with the Quorum Witness using the HPE 3PAR StoreServ management interface. The network should be setup such that the QW server has network connectivity to the admin interface on both arrays. It is required to connect the administration port from a minimum of two controller nodes to the network. However if the array has more than two controllers the best practice is to connect all controller nodes to the network.

Support and coexistence of different cluster platforms with 3PAR Peer Persistence
It is not unusual that customer might have a need to support different cluster platforms across data centers. As 3PAR Peer Persistence expands its support for different platforms in a metro cluster or multisite cluster deployment, customers can take advantage and deploy those clusters. With the release of the 3PAR OS 3.2.2, VMware vSphere, Microsoft Windows and Red Hat® RHEL are supported with 3PAR Peer Persistence. It is fully supported for customers to have clusters for all of these environment deployed in a stretched cluster environment with a single 3PAR Peer Persistence.

VMware vSphere Metro Storage Cluster (vMSC) Configuration
Two vMSC configurations are possible based on how the hosts are connected to the storage arrays.

vMSC uniform
In this configuration each host can access LUNs exported from both the sites. Figure 1 shows a typical Uniform configuration. This setup helps in protecting against a storage failure at a site. So, if the array on site 1 fails storage will failover to the HPE 3PAR StoreServ array on site 2 transparently and automatically allowing VMs on site 1 continued access to storage on site 2 without any interruption. Of course this implies that customers need to have good bandwidth between their sites. Outside of what happens in sync replication (ask only after write to remote array is completed), even host to storage array traffic will be routed across the FC connections to the remote site (in synchronous replication, hosts are writing only to the local array).

Note that with Uniform, the use case is typically of active load balancing being done where secondary paths/volumes are made active on site 2 array as a means to achieve load balancing. This is in addition to what Uniform provides which protection against storage failure.

vMSC non-uniform
In this configuration each host can access storage resources available only on its local site. Figure 2 shows a typical non-uniform configuration. This is geared towards protecting against a complete site failure similar to what VMware vCenter Site Recovery Manager (SRM) does.

The concept here is that if a complete site fails (storage and server cluster), Peer Persistence will be able to transparently and automatically failover the storage to site 2 while the host application fails over in parallel. Consider this really as a DR solution. Non-uniform configurations will not allow an application to transparently failover. If a switchover is executed the local ESXi servers will lose connectivity, the VMs will go offline and will be switched according to the HA setup, if they are switched to the remote site, they will restart automatically.

Note
vMSC non-uniform is not supported with HPE 3PAR Peer Persistence.
Planned switchover

A switchover operation migrates the remote copy group role from source to target without impacting host I/O. This operation requires that the associated hosts are connected to the arrays on both sites (Uniform vMSC). This is a planned operation and the remote copy groups need to already be started and synced for this command to work. This operation is useful when you want have the least possible latency for accessing storage when using VMs from a host. For example, if vMotion is used to move all VMs on a host from site 1 to site 2, one can use the switchover command to reverse the replication resulting in site 2 hosts being able to access the LUNs locally from site 2 array. The system performs the following action as part of a switchover operation:

- I/O from the Virtual Machines to the volumes in the source remote copy group is blocked and in flight I/O is allowed to drain. The remote copy group is stopped and snapshots are taken on the primary array.
- The primary array target port group is changed to transition state and it sends a remote failover request to the secondary remote copy group.
- The secondary array target port group is then changed to transition state and it takes a recovery point snapshot.
- The secondary remote copy group changes state to become primary-reversed and makes the volumes read/write.
- The pri-rev target port group is changed to active state and the array returns a failover complete message to the primary remote copy group.
- The primary array target port group is changed to standby state and any blocked I/O is returned to the host with a sense error: NOT READY, LOGICAL UNIT NOT ACCESSIBLE, TARGET PORT IN STANDBY STATE.
- The host will then perform SCSI inquiry requests to detect what target port groups have changed and which paths are now active and I/O will now be serviced on the active path to the primary-reverse remote copy group.
- All operations to this point should complete within 30 seconds to ensure that host I/O does not timeout.
- The primary remote copy group will then send a remote recover request to the primary-reverse remote copy group. The primary-reverse remote copy group will perform a recover operation and will change the primary remote copy group state to secondary-reverse state. The remote copy group will then be started from primary-reverse to secondary-reverse.
• When the primary-reverse and secondary-reverse volumes are back in sync the snapshots that were taken on both sides will be removed. The remote copy group will then undergo a reverse (- natural) operation. This operation will change the primary-reserve group to primary and secondary-reverse group to secondary. It is possible for this operation to time out or fail if the target links go down during processing. If this happens, issue the “setrcopygroup reverse - natural” command to complete the process manually. The system is now fully reversed and ready for another switchover request.

HPE 3PAR Peer Persistence failures handling

Peer Persistence with Quorum Witness can identify different types of failures and perform transparent failover automatically. Peer Persistence will perform transparent failover only for remote copy groups that have the auto failover and path management enabled. The path management policy ensure that the target port group state of volumes in the specified group will be Active on the primary and reported as Standby on the secondary. The auto_failover ensure that an automatic failover on a remote copy group when used in conjunction with the quorum witness functionality.

It is fully supported for customers to have mixed remote copy groups where certain number of them are configured for transparent failover while others are not setup for it.

Array to Array communication failure

In the case of an array to array communication failure, the HPE 3PAR StoreServ continues to send/receive communication to/from the Quorum Witness. Automatic transparent failover does not occur and host I/Os continue to go to their primary volumes; however replication of I/O across RC links will stop due to the failure.

Single site to QW communication failure

In the case where the arrays at either site 1 or site 2 lose connectivity with the Quorum Witness, an automatic failover will not occur. Host I/O will continue as before and replication of I/O across RC links will continue as normal. An automatic failover does not need to occur because the two HPE 3PAR StoreServ arrays can still communicate with one another via the replication links they share.

Site 1 to QW and Array to Array communication failure

If the array in site 1 becomes isolated due to a dual network failure (communication with the QW fails and the replication links to the array in site 2 fails), Remote copy groups that are in primary state on the array in site 1 will be transitioned to failsafe mode and will stop serving I/O to the ESXi hosts in site 1. The array in site 2 (which is still communicating with the QW) will perform an automatic failover and host I/O will be redirected to the new primary volumes in site 2.

Site 2 to QW and Array to Array communication failure

This is the same as the previous case with the difference that site 2 will become isolated due to a dual network failure.

Site 1 and site 2 both lose communications to the QW

In the case of a communication failure with the QW from both sites, but where the Remote Copy links remain operational both arrays will be aware of each other’s operational state. Automatic failover does not happen and host I/Os continue to go to the primary volumes; replication of I/O across RC links will continue as normal.

Site 1 and site 2 to QW and Array to Array communication failure.

In the case where all network communications between the arrays and the QW and the communication between both arrays also fails, both arrays will be isolated as they cannot communicate with each other over the RC links nor can they communicate with the QW. In this case remote copy groups that are primary will go into failsafe mode and stop serving I/O, and failover actions will not be performed. This will result in host I/O failure and replication of I/O across RC links will stop due to the failure.
### Table 1. HPE 3PAR Peer Persistence failures handling

<table>
<thead>
<tr>
<th>Failure scenario</th>
<th>Replication Stopped</th>
<th>Automatic Failover</th>
<th>Host I/O Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Array to Array remote copy links failure</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Single site to QW network failure</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Single Site to QW network and Array to Array remote copy link failure</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Both sites to QW network failure</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Both sites to QW network and Array to Array remote copy link failure</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### VMware vMSC and HPE 3PAR Peer Persistence best practices

The implementation of the VMware vSphere Metro Storage Cluster requires certain considerations for optimal deployment. VMware has published a technical white paper titled: “VMware vSphere Metro Storage Cluster Recommended Practices”. The paper discusses vMSC design considerations and operational procedures. It is a recommended document to review for better outcome of the deployment of 3PAR StoreServ Peer Persistence in a vMSC environment.

In the next section, we will discuss few of the considerations and recommendations that customer should consider when deploying 3PAR Peer Persistence with vMSC.

#### VMware DRS Affinity Group Usage

The VMs should be grouped together to the preferred site during normal operation. This will guarantee that they use local storage in their preferred site. However, during certain failure scenarios these VMs may be running on the storage in the other data center. The user should consider using the “should rule” rather than the “must rule” with affinity group. Making this choice will allow VMs to failover to the other site as a result of the VMware HA initiating those moves. The result of this implementation is to simply employ affinity group to create site awareness leveraging the DRS “should rule.” For more information about those setting refer to the VMware vSphere HA & DRS best practices.

#### Managing VMware Storage DRS settings

VMware vSphere 5.0 introduced a new feature called Storage DRS. The feature provides smart VM placement and load balancing methods based on space capacity utilization and I/O latency. It offers an efficient way for provisioning of VMs and monitoring of storage used by the vSphere cluster.

Since the goal during normal operation is to have VMs running on their preferred site, the setup of Storage DRS should be “Manual Mode”. In this mode, Storage DRS will make recommendations when thresholds for space utilization or I/O latency are exceeded. This will help avoid having Storage DRS moving VMs to a LUN on remote storage. One also needs to take into consideration the extra load on the links between the two sites due to storage vMotion between LUNs from the two sites. Its best to schedule these moves for non-peak hours and implement the recommendations manually if needed. Even with manual mode, the user should still take advantage of the Storage DRS on the initial placement of the VMs during the provisioning process along with the monitoring of the environment.

#### Workload management across sites

Users should consider the impact that vMotion and Storage vMotion will have on the inter-site links. Making an effort to avoid having host I/O cross sites, vMotion and Storage vMotion can be used to ensure that the inter-site bandwidth is dedicated to replicating production data.
Multi-VM application placement
User should consider grouping or placing multi-VM applications together on the same site. Such placement would ensure that VMs which have dependencies on other VMs and could have significant traffic between them are not consuming resources on replication links. There is also the added latency due to the distance between the two sites in the case where VMs sharing a datastore are split between the two sites. The added latency can impact the application user’s experience.

Heartbeat Setting
VMware vSphere 5.0 introduced Datastore Heartbeating. The datastore heartbeat mechanism is used to check if a host has failed in the VMware HA cluster or simply if it’s isolated due to management network issues. The master host in the HA cluster relies on this mechanism to assess if a host has failed and as a result if the VMs hosted on that host would need to be restarted on another host in the HA cluster. The default behavior of the VMware HA cluster is to use two datastores for the heartbeat. vCenter uses an algorithm to pick the two datastore. It is recommended that the user consider using a total of 4 with 2 at each site in case of the loss of communication between the two sites.

VMware HA Admission Control
Users should consider the resources required at each site to handle the workload in case of a failure of one site and the subsequent failover for the workload to the surviving site. VMware vCenter uses the Admission Control to guarantee that enough resources are available in the cluster to protect user workload in case of a failover. By enabling admission control and configuring its policy to be able to handle the all workload when restarted by VMware HA.

Isolation addresses & false positive
Since the network heartbeat is key in the indicating the state of a host, the user should ensure that the management network is resilient. Using the VMware “das.isolationaddress” setting the user should assign one isolation address per site as minimum to avoid having an isolation incident due to the communication loss between the two sites that might result in unnecessary failover and subsequent restart of VMs.

Permanent Device Loss (PDL) and All Paths Down (APD)
In vSphere 6.0, VMware added VM Component Protection (VMCP) to vSphere High Availability (HA). This enhancement enables HA to respond to situation where the connectivity to a VM’s datastore is affected either temporarily or permanently. The response to the datastore accessibility failures can vary between generating alarms to restarting VMs on another host.

A Permanent Device Loss (PDL) event occurs when a device presented to host becomes unavailable and the storage array issues a SCSI sense code to the host for the event. In a PDL state, the storage array can communicate with the vSphere host and will issue SCSI sense codes to indicate the status of the device. As a result, the host will stop issuing I/O requests to that device on the array when a PDL state is detected. The host considers the device permanently unavailable. An example of this scenario is a failed LUN, or an administrator inadvertently changing zone configuration.

The configuration for PDL handling in vSphere HA offers three actions that can be taken in response to a PDL event:

- Disabled—No action is taken.
- Issue events—No action is taken but notifications will be sent when a PDL event has occurred.
- Power off and restart VMs—All affected VMs will be terminated on the host and vSphere HA will attempt to restart the VMs on hosts that still have connectivity to the storage device.

Users should consider implementing VMware recommendation for vSphere HA by setting the action to “Power off and restart VMs”. When the PDL condition is detected, a VM is restarted instantly on a healthy host within the vSphere HA cluster.

An All Paths Down (APD) event occurs when the vSphere host cannot access the storage device, and there is no PDL SCSI code returned from the storage array. The device is considered to be in an APD state. The device may return, or it may not. During an APD condition, the host continues to retry I/O commands to the storage device until the APD Timeout is reached. Once the APD Timeout is expired, the host begins to fast-fail any non-virtual machine I/O to the storage device. The default APD Timeout value is 140 seconds and can be changed per host using the Misc.APDTimeout advanced setting.
There are options available for an APD response:

- **Disabled**—No action is taken.
- **Issue events**—No action is taken but notifications are sent when an APD event has occurred.
- **Power off and restart VMs (conservative)**—vSphere HA will not attempt to restart the affected VMs unless it has determined there is another host that can restart the VMs.
- **Power off and restart VMs (aggressive)**—vSphere HA will terminate the affected VMs even if it cannot determine that another host can restart the VMs.
- **Delay for VM failover for APD**—Once the APD Timeout has been reached (default: 140 seconds) VMCP will wait an additional period of time before taking action against the affected VMs. By default, the waiting period is 3 minutes. In other words, VMCP will wait 320s before taking action on VMs. The sum of the APD Timeout and the Delay for VM Failover is also known as the VMCP Timeout.

**Response for APD recovery after APD timeout**

This setting will instruct vSphere HA to take a certain action if an APD event is cleared after the APD timeout was reached but before the Delay for VM failover has been reached.

- **Disabled**—No action will be taken.
- **Reset VMs**—The VMs will be reset on the same host. This option is available because some applications or guest operating systems may be in an unstable condition after losing connection with storage services for an extended period of time. This setting will instruct vSphere HA how to handle this situation.

For more information please refer to the “VMware vSphere Metro Storage Cluster Recommended Practices” white paper. There is also an excellent VMware vSphere blog discussing VM Component Protection (VMCP) in details as well.

**HPE 3PAR Quorum Witness Hosting**

Given how critical the QW is to operation of HPE 3PAR Peer Persistence, it is highly recommended that the QW VM should be hosted outside the stretched cluster environment and should not be on the HPE 3PAR arrays used in the HPE 3PAR Peer Persistence deployment it arbitrates or the vSphere hosts that are part of the Metro Cluster.

**Remote Copy Replication links**

It is recommended that RCFC be used for the synchronous replication of data between the arrays in a Peer Persistence configuration. Use of RCFC will ensure a high bandwidth low latency connection for the replication of data between the arrays and it uses a patented protocol for synchronous replication that only requires a single round trip across the replication links to replicate an I/O smaller than 80K in size.

**HPE 3PAR Peer Persistence maximums**

It is highly recommended to consolidate VVs inside few RC Groups in a HPE 3PAR Peer Persistence configuration. Such configuration will guarantee the host I/Os will be failed over within the timing window in case of a transparent failover.

For the latest supported configuration, maximum and requirement for 3PAR Peer Persistence please refer to the Single Point of Connectivity Knowledge (SPOCK).
**HPE OneView for VMware vCenter**

HPE OneView for VMware vCenter is software for VMware's vCenter management console which enables the vSphere administrator to quickly obtain context-aware information about HPE servers and HPE storage in their VMware vSphere environment directly from within vCenter. This enables the vSphere administrator to easily manage physical servers, storage, datastores, and virtual machines. By providing the ability to clearly view and directly manage relationships between virtual machines and HPE infrastructure, the VMware administrator's productivity increases, as does the ability to ensure quality of service.

![Figure 3. HPE OneView and HPE Management—Storage](image)

**Features and benefits**

- Simplify administration with integration of the physical with the virtual infrastructure.
- Reduce downtime by automating responses to hardware events.
- Take control by launching your trusted HPE management tools.
- Proactively manage changes with detailed relationship dashboards of server, networking, and storage.
- Maintain stability and reliability with firmware inventory and deployment.
- On demand server and storage capacity provisioning.
- Visualize complex relationships:
  - How virtual machines are mapped to underlying storage.
  - How peer persistence volumes are being configured.

![Figure 4. HPE OneView and HPE Management—VMs to Storage mapping](image)
HPE OneView for VMware vCenter v 7.4 added support for new storage provisioning wizards for HPE 3PAR Peer Persistence. Using the wizard will allow for the setting of the "auto_failover" and "path_management."

![HPE OneView and HPE Management—Switch Peer Persistence](image)

**Figure 5.** HPE OneView and HPE Management—Switch Peer Persistence

### Sample 3PAR Peer Persistence environment configuration

In this example, we are deploying two HPE 3PAR StoreServ storage arrays that span two locations (Redmond and Kirkland) are presented to a VMware vSphere Metro Storage Cluster (vMSC) as a single stretched cluster using 3PAR Peer Persistence. The vSphere cluster consists of 4 hosts running ESXi 6.0 with 2 hosts at each site. The Quorum Witness is hosted at a third site.

![3PAR Peer Persistence sample cluster configuration](image)

**Figure 6.** 3PAR Peer Persistence sample cluster configuration
Configuring Remote Copy

The Remote Copy configurations and operations for 3PAR Peer Persistence deployment are based on a linked pair of storage systems, which are called a remote-copy pair. In a remote-copy pair, one system is the primary system and one is the backup system. The primary system holds the primary virtual volumes, and the backup system holds the secondary virtual volumes. Secondary virtual volumes contain copies of the primary virtual volumes. The volume that is part of a Primary Remote Copy group is the source volume; the replicated volume in the secondary Remote Copy group is called the target volume.

When you set up HPE 3PAR Remote Copy, you configure the backup system in the remote-copy pair as the target of the primary system, and the primary system as the target of the backup system.

Note

The steps below assume that hardware requirements for Remote Copy are met. Ports, cabling and zoning are correctly set up; and also, that all requirements for Peer Persistence have been met.

Figure 7. 3PAR system view for Remote Copy configuration
Creating remote copy configuration

In order to setup RC between the two 3PAR StoreServ systems in the two site, we would need to create remote copy targets on both arrays. Using the 3PAR StoreServ MC to connect to both systems as in figure 7. Then, proceed with the RC 1:1 configuration setup as in figure 8.

In this step we would select ports from both system to create an RCFC link. For this sample deployment.

![Figure 8. Pairing of the 3PAR StoreServ for RC configuration](image)

![Figure 9. 3PAR StoreServ RC configuration overview](image)
Once Remote Copy is configured between the two arrays, the next step is to deploy and configure a quorum witness in order to complete the setup the 3PAR Peer Persistence quorum.

**Deploying a Quorum Witness**

The Quorum Witness (QW) software provides the arbitrator functionality in the Peer Persistence setup. The detailed instructions for deploying Quorum Witness are outlined in the 3PAR QW Deployment chapter in “3PAR StoreServ Remote Copy Software User’s Guide”. Please refer to the User’s Guide for detailed steps and updated information to deploy it.

In this setup the QW was deployed on an ESXi 6.0 host. Once the deployment of the OVF template of the QW was completed, the password, DNS and network were configured. The IP address assigned during this QW setup will be used in the 3PAR Peer Persistence Quorum witness.

![3PAR Peer Persistence QW VM](image)

**Figure 10.** 3PAR Peer Persistence QW VM
Troubleshooting Witness Quorum Communication issues

As the communication between the 3PAR StoreServ systems and the QW is key for the 3PAR Peer Persistence quorum, it is critical that verify status of the communication once the QW is deployed. The verification is performed using 3PAR CLI, so login to one of the 3PAR StoreServ to arrays in the quorum and using the witness check command to verify the communication between Quorum Witness and management interface of the two arrays. The two commands will not report any message when successful in earlier versions of 3PAR OS 3.1.2 MUs and 3.1.3 MUs. In version 3.2.1 or newer a basic message is displayed if the verification is successful.

Figure 11. 3PAR Peer Persistence QW troubleshooting check 1

In the case the Quorum Witness server is unreachable, an error will be issued. The following is an example if the Quorum Witness server is unreachable.

Figure 12. 3PAR Peer Persistence QW troubleshooting check 2

Configuring the 3PAR Peer Persistence quorum

As the 3PAR StoreServ systems were configured for Remote Copy and the quorum witness is deployed, the next step is the 3PAR Peer Persistence quorum between these 3 components: 3PAR systems and QW.

Using the 3PAR SSMC, the QW can be configured using the Remote Copy Configuration\Action\Configure quorum witness as in figure 13.

Figure 13. Configure quorum witness
Using the IP address of the Quorum Witness that was noted down from the QW deployment, we complete the configuration of the 3PAR Peer Persistence quorum.

**Figure 14.** 3PAR Peer Persistence quorum configuration

The Quorum will be configured and started. The status can be checked by using the “showrcopy –qw targets” command from the 3PAR StoreServ CLI on one of the 3PAR StoreServ in the quorum. Or using the 3PAR SSMC.

**Figure 15.** QW Configuration–status check
Remote Copy Group Configuration

Using the Remote Copy “Create Group” wizard, a remote copy group can be created by selecting a source and target system. The user would have to specify the group name and select synchronous mode. The option to create remote VV automatically will ensure that the VVs have the same WWN. It's preferred to create new volumes on the target, this way the volumes will have the same characteristics as the source volumes. If you manually created volume beforehand, then it is important to ensure that the volume size on target matches the volume size on the source and both volumes have the same WWN.
It is also important to ensure that the 3PAR Peer Persistence policies for remote copy groups are selected for the path management and auto_failover handling. Adding the source volumes to the remote copy groups is the last option in the creation of these remote copy groups.
Once virtual volumes are added in the group and task is finished. The group state status will change to “started”, and the initial synchronization will begin. Additional groups can be created to meet business needs. A typical uniform vMSC configuration will have Remote Copy groups replicating in both directions. In this example setup, one remote copy groups primary on each array. There are two volumes in each of these remote copy groups.

**Figure 20.** Remote Copy Groups configuration

**Figure 21.** Remote Copy Groups configuration overview

**Exporting LUNs to hosts**

When creating ESXi hosts, they need to be created using Persona 11. For ease of deployment and management, user should consider adding all hosts in the cluster in a host set in both arrays. Adding all VVs in a remote copy group in a VV set. Such options will provide for an easier management of the storage provisioning for the Cluster.
After exporting the VVs from both arrays to the hosts in the cluster, a rescan might be needed for the host to display the paths and storage presented. A closer look at path details for one of the LUNs and you should see some “Active” paths and some “Standby” paths. The next step would be to create datastores on these LUNs and start using them in the vSphere cluster. After exporting the VVs from both arrays, verify in the path details of the LUNs that the expected number of Active and Standby paths are available.
Testing 3PAR Peer Persistence planned switchover

The HPE 3PAR Peer Persistence switchover operation migrates host I/O from one storage array to the other and reverses the direction of data replication. The switchover is a manual operation designed to facilitate service optimization and storage system maintenance activities within a high-availability data storage solution.

In a switchover, the operation is initiated in one of two ways:

- Using the 3PAR CLI and issuing the setrcopy switchover command from the command line interface, to a remote-copy volume group on the primary storage system.
- Using the 3PAR SSMC and issuing the switchover command from the SSMC.

The “Planned switchover” section provides more details about the workings of this operation.

In this example the remote copy group is primary on Kirkland3PAR.

**Figure 24. 3PAR Peer Persistence status before switchover**

The switchover can be executed by using the 3PAR SSMC as in figure 25

**Figure 25. 3PAR Peer Persistence switchover using the 3PAR SSMC**
After the switchover, the remote copy group is now primary on the Redmond3PAR array. This operation that is transparent to the host I/O highlight the value of 3PAR Peer Persistence as a solution for workload mobility between data centers. This scenario apply to balancing resources and maintenance use cases in a typical customer environment.

3PAR Peer Persistence Automatic Transparent Failover

Automatic transparent failover migrates the servicing of host I/O from the failed primary to the surviving secondary storage system. In an automatic transparent failover, there is no coordination of operation between the two storage systems, because the original source system failed and became inaccessible due to power outage or other failures rendering it completely inaccessible. In an automatic transparent failover, the trigger is provided to the surviving secondary storage system by the Quorum Witness infrastructure and related software, which begins state transition and activation of the secondary host paths for all applicable secondary volume groups. Only primary remote copy groups from that primary array that have “auto_failover” policy enabled will be failed over to the second array.

For an automatic failover operation, replication will have stopped, and will remain unavailable until the failed system is recovered. Recovery process is needed to restart replication and reverse replication back in the original direction.

Prior to the automatic failover event, half of the remote copy groups in this 3PAR Peer Persistence configuration are primary on Kirkland3PAR. As we are simulating the event of an array power outage impacting Kirkland3PAR, those remote copy groups will failover to the other surviving 3PAR (Redmond3PAR).

Figure 26. 3PAR Peer Persistence configuration after switchover

Figure 27. 3PAR Peer Persistence configuration prior to an automatic transparent failover
Failover simulation

In order to simulate an event causing an outage at the Kirkland site (power or network communication outage), we break the communication between the Kirkland 3PAR array and the QW as well as the Remote Copy replication links between the two arrays. Please note that the breakup of RC links and the QW is simultaneous.

As the Kirkland3PAR becomes totally inaccessible due to a simulated outage at the data center, the 3PAR Peer Persistence automatic failover will protect the applications workload running on the virtual volumes that are parts of the remote copy group RCG-KR-1. The failover is transparent to the host I/Os.

When we check the status on the surviving array Redmond3PAR, we notice that quorum is in a failover state, the remote copy links are down and that the Kirkland3PAR target is marked as failed. The remote copy group RCG-KR-1 is now “Primary-Rev” on the Redmond3PAR.

```
Redmond3PAR cli% showcopy -qw
Remote Copy System Information
Status: Started, Normal
Target Information
Name    ID Type Status Policy QW-ID Q-Status Q-Status-Blend
Kirkland3PAR 5 FC failed mirror_config 192.168.0.34 Failover

Link Information
Target Node Address Status Options
Kirkland3PAR 0:2.3 20230002A018575 Down -
Kirkland3PAR 1:2.4 21240002A018575 Down -
receive 0:2.3 20230002A018575 Up -
receive 1:2.4 21240002A018575 Up -

Group Information
Name    Target Status Role Mode Options
RCG-KR-1 89701 Kirkland3PAR Stopped Primary-Rev Sync auto_failover, path_management
LocalVV  ID RemoteVV  ID Sync Status LastSyncTime
vmsk-kir-red-vv1 1023 vmsk-kir-red-vv1 221 Stopped 2015-12-13 02:43:21 PST
vmsk-kir-red-vv2 1924 vmsk-kir-red-vv2 222 Stopped 2015-12-13 02:43:21 PST

Name    Target Status Role Mode Options
RCG-RK-2 Kirkland3PAR Stopped Primary Sync auto_failover, path_management
LocalVV  ID RemoteVV  ID Sync Status LastSyncTime
vmsk-red-kir-vv1 784 vmsk-red-kir-vv1 439 Stopped 2015-12-13 02:23:46 PST
```

Figure 28. Source System (Kirkland3PAR) for RC groups prior to an automatic transparent failover

Figure 29. 3PAR Peer Persistence configuration on surviving array after an automatic failover
Recovering from an automatic transparent failover

Once the condition that caused the array to become inaccessible are cleared and it is back online, remote copy replication links and communication with the QW are restored.

We proceed to recovering remote copy groups and restoring the environment to its state prior to the automatic transparent failure event. To recover the remote copy groups, the 3PAR SSMC or the 3PAR CLI can be used in this to perform this task. Using the 3PAR SSMC, Select the remote copy group RCG-KR-1 and choose “Recover” action. The task will reverse replication and synchronize the delta changes from the backup system (Redmond3PAR). The RCG-KR-1 will be Primary-rev on the Redmond3PAR system while is Secondary-rev on the Kirkland3PAR system that was just brought back online.

![Remote Copy recovery after an automatic failover](image)

![Remote copy group status after recovery](image)

**Figure 30.** Remote Copy recovery after an automatic failover

**Figure 31.** Remote copy group status after recovery
As we are looking to restore the remote copy group to its state prior to the ATF, it is required to reverse the volume groups by manually complete the reverse operation on the volume group by running the "setrcopygroup reverse -natural" command from the 3PAR CLI to reverse the roles of RCG-KR-1.

Once this task is completed, the RCG-KR-1 will become Primary on the Redmond3PAR and Secondary on the Kirkland3PAR. The last step in restoring the environment to its initial state prior to the ATF event, is to switchover the remote copy group RCG-KR-1 to Kirkland3PAR. Both, the reverse and switchover task are outlined in figure 32.

```
Redmond3PAR cli% setrcopygroup reverse -natural -f RCG-KR-1.r99701
reverse started with tasks: 16844
Redmond3PAR cli% setrcopygroup switchover -f RCG-KR-1.r99701
switchover started with tasks: 16845
```

Figure 32. Remote copy group reverse and switchover

Figure 33. 3PAR Peer Persistence environment restored back

**Resources**

HPE Enterprise Storage Information Library: [hp.com/Go/Storage/Docs](http://hp.com/Go/Storage/Docs)

HPE 3PAR StoreServ Storage: [hp.com/go/storeerv](http://hp.com/go/storeerv)

3PAR VMware Implementation Guide

HPE 3PAR StoreServ Storage best practice guide

HPE 3PAR Remote Copy Software User Guide OS 3.2.2

VMware vSphere Metro Storage Cluster recommended practices
Learn more at
hp.com/solutions/activeanswers
hp.com/go/3parstoreserv