Best Practices for Using Dell SRDF/Metro in a VMware vSphere Metro Storage Cluster

Abstract

This white paper discusses best practices when configuring Dell SRDF/Metro with VMware vSphere® Metro Storage Cluster (vMSC).

November 2023

Dell Engineering

Revisions

Date	Description
August 2023	New template release
November 2023	Update for PowerMaxOS 10.1.0.0, vSphere 8.0U2

Acknowledgments

Author: Drew Tonnesen, drew.tonnesen@dell.com

Table of contents

Executive Summary

The Dell PowerMax™ family provides disaster recovery and mobility solutions through its remote replication technology SRDF® (Symmetrix Remote Data Facility). SRDF has the capability to replicate between multiple sites, co-located or even thousands of miles apart depending on the type of replication desired.

As part of this replication family, Dell offers SRDF/Metro®, an active/active version of SRDF. In a traditional SRDF device pair relationship, the secondary device, or R2, is write disabled. Only the primary device, or R1, is accessible for read/write activity. It is an active/passive solution. With SRDF/Metro, the R2 is also write enabled and accessible by the host or application. The R2 takes on the personality of the R1 including the WWN. A host, therefore, would see both the R1 and R2 as the same device.

As both devices are simultaneously accessible, the hosts in a cluster, for example, can read and write to both the R1 and R2. The SRDF/Metro technology ensures that the R1 and R2 remain current and consistent, addressing any conflicts which might arise between the pairs.

When SRDF/Metro is used in conjunction with VMware vSphere across multiple hosts in a single vCenter, a VMware vSphere Metro Storage Cluster (vMSC) is formed. At its core, a VMware vMSC infrastructure is a stretched cluster. The architecture is built on the idea of extending what is defined as “local” in terms of network and storage. This enables these subsystems to span distance, presenting a single and common base infrastructure set of resources to the vSphere cluster at both sites. In essence, it stretches network and storage between sites.

With vMSC customers acquire the capability to migrate virtual machines between sites with VMware vSphere vMotion® and vSphere Storage vMotion, enabling on-demand and nonintrusive mobility of workload.

Audience

This white paper is intended for VMware administrators, server administrators, and storage administrators responsible for creating, managing, and using VMware, as well as their underlying storage devices. The paper assumes the reader is familiar with VMware infrastructure and the VMAX or PowerMax arrays and the related software.

1 Introduction

The purpose of this paper is to provide the best practices for running a VMware vSphere Metro Storage Cluster (vMSC) that utilizes Dell SRDF/Metro technology for stretched storage. It will include best practices for vMSC as well as SRDF/Metro and any special considerations when running the two technologies together.

This paper covers up to the PowerMaxOS 10.1.0.0 (6079) and vSphere 8.0U2 releases. Although some information (e.g., images) contained herein is from earlier PowerMaxOS, management, and/or vSphere versions, the best practices cover all revisions. It is always important to check product documentation for any limitations or features that are present in a particular customer's hardware/software versions.

1.1 NVMe over Fabrics (NVMeoF)

Beginning with PowerMaxOS 5978.444.444, the PowerMax array introduced NVMeoF or NVMe over Fabrics. NVMe stands for non-volatile memory. This is the media itself such as NAND-based flash or storage class memory (SCM) which compromise the PowerMax backend. NVMe or non-volatile memory express is a set of standards which define a PCI Express (PCIe) interface used to efficiently access data storage volumes on NVM. NVMe provides concurrency, parallelism, and scalability to drive performance. It replaces the SCSI interface. NVMeoF is the specification that details how to access that NVMe storage over the network between host and storage. The network transport could be Fibre Channel, TCP, RoCE, or any other number of next generation fabrics. Dell supports Fibre Channel with NVMe on the PowerMax 2000/2500/8000/8500[1], also known as FC-NVMe, and TCP with NVMe on the PowerMax 2500/8500, also known as NVMe/TCP.

To use FC-NVMe with the PowerMax requires 32 Gb/s Fibre Channel modules (SLICs). The emulation FN is assigned to ports on these modules to support FC-NVMe. In addition, in order to use FC-NVMe with a host, a supported Operating System and supported HBA card is necessary. With NVMeoF, targets are presented as namespaces (equivalent to SCSI LUNs) to a host in active/active or asymmetrical access modes (ALUA).

To use NVMe/TCP with the PowerMax requires 25 Gb/s network modules (SLICs). An OR emulation is assigned to ports on these modules to support NVMe/TCP. In addition, in order to use NVMe/TCP with a host, a supported Operating System and supported network card is necessary. With NVMeoF, targets are presented as namespaces (equivalent to SCSI LUNs) to a host in active/active or asymmetrical access modes (ALUA).

VMware vSphere 7 is the first release to support NVMeoF. It offers both NVMe over Fibre Channel (FC-NVMe) and NVMe over RDMA (RoCE v2). Dell offers FC-NVMe with vSphere 7 on the PowerMax 2000/2500/8000/8500 and NVMe/TCP with vSphere 7 U3 and higher on the PowerMax 2500/8500. ESXi hosts discover and use the presented namespaces and emulates the NVMeoF targets as SCSI targets internally and presents them.

1.1.1 Support

While SRDF supports the use of NVMeoF in active/passive configurations, the standards (specifications) for active/active solutions are not yet ported over to NVMeoF for any vendor, and as a result, SRDF/Metro is not supported.

[1] FC-NVMe requires PowerMaxOS 10.1.0.0 with PowerMax 2500/8500.

2 SRDF/Metro

SRDF/Metro is a feature available on the VMAX3 and VMAX All Flash arrays starting with HYPERMAX OS 5977.691.684 and all PowerMax arrays which provides active/active access to the R1 and R2 of an SRDF configuration. In traditional SRDF, R1 devices are Read/Write accessible while R2 devices are Read Only/Write Disabled. In SRDF/Metro configurations both the R1 and R2 are Read/Write accessible. The way this is accomplished is the R2 takes on the personality of the R1 in terms of geometry and most importantly the WWN. By sharing a WWN, the R1 and R2 appear as a shared virtual device across the two arrays for host presentation. A host or typically multiple hosts in a cluster can read and write to both the R1 and R2. SRDF/Metro ensures that each copy remains current and consistent and addresses any write conflicts which might arise. The SRDF mode is referred to as “Active”. In VMware environments, the ESXi hosts from two different data centers can be placed in the same vCenter, forming a VMware vSphere Metro Storage Cluster (vMSC). A simple diagram of the feature is shown in Figure 1.

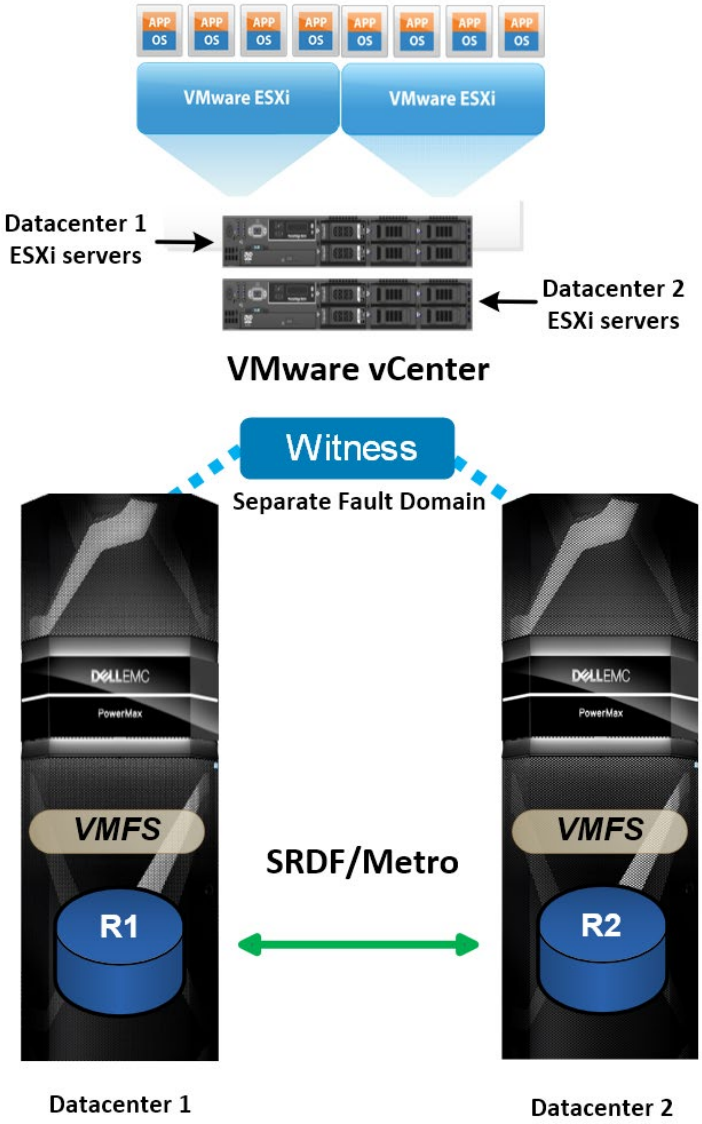

Figure 1. SRDF/Metro

This feature provides the following advantages:

- A high availability solution at Metro distances by leveraging and extending SRDF/S functionality.
- Active-Active replication capabilities on both the source and target sites.
- Witness support to enable full high availability, resiliency, and seamless failover.

2.1 Smart DR/MetroDR

SRDF/Metro supports 3-site configurations with a leg off either the R1, R2 or both members of a pair. This means that neither the R1 nor R2 is aware of the existence of an asynchronous or adaptive copy leg off the other since the pairs are independent. Up until PowerMaxOS 5978 Q3 2020, SRDF/Metro did not support a Star-like configuration where either the R1 or R2 could update a single, remote leg. Beginning with PowerMaxOS 5978 Q3 2020, Dell offers Smart DR (also known as MetroDR) which enhances SRDF/Metro and provides Star-like functionality. The Smart DR feature extends the high availability (HA) solution of SRDF/Metro to a third array and supports the ability to have Geo distance DR support for a device in a Metro configuration. Smart DR integrates SRDF/Metro and SRDF/A/ACP enabling HR DR by closely coupling the SRDF/A/ACP sessions on each side of the Metro pair to replicate to a single DR device. While only one side of the SRDF/Metro pair (R1) will update the SRDF/A/ACP leg at one time, both are capable of doing so. The asynchronous replication of data from either side of the SRDF/Metro pair to a tertiary site enables Failover/Failback to the DR site while retaining the Metro environment. Smart DR requires a minimum of PowerMaxOS 5978 Q3 2020 and Solutions Enabler 9.2 running on the three arrays with Witness (array or virtual) configured. Bias is not supported. Both asynchronous (SRDF/A) and adaptive copy (SRDF/ACP) modes are supported on the third leg.

Figure 2 is a schematic representation of the business continuity solution that integrates a VMware environment and Smart DR. The SRDF/Metro array pair hosts the VMware Metro Storage Cluster while the SRDF/A/ACP array becomes a true failover site. Note that the Witness in the figure should be in a separate fault domain, whether array or virtual, and the third array could be thousands of miles away.

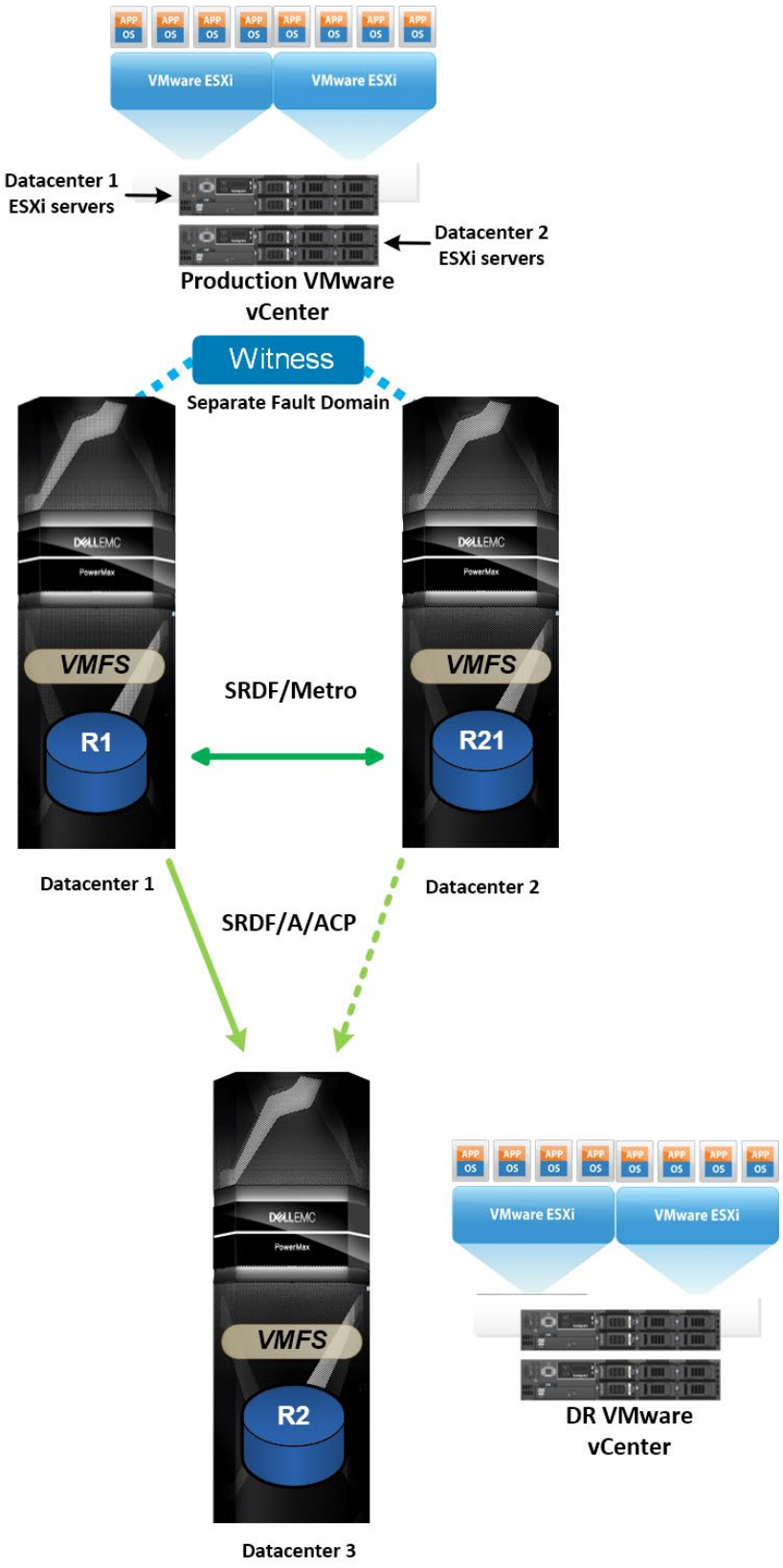

Figure 2. Smart DR

The use of Smart DR does not change the best practices for the vMSC configuration. As vMSC is an HA solution, Smart DR does, however, provide a superior DR solution over a single concurrent or cascaded leg since both the R1 and R2 can update the third site.

Note: An SRDF/Metro pair cannot be replicated to another SRDF/Metro pair, and thus it is not possible to have a vMSC configuration at the production and disaster recovery sites; however, if a DR event occurs that necessitates a failover, it is possible to remove the SRDF relationships between the production and disaster recovery sites and then create a new SRDF/Metro relationship and thus build a vMSC at the DR site.

2.2 Latency

Dell does not publish strict requirements for latency between SRDF/Metro arrays, however, keeping under 5ms round-trip time (RTT) would be an appropriate goal. VMware requires no more than 10ms RTT even if the solution could support it.

Dell recommends keeping the distance between arrays under 200 km for SRDF/Metro. What is most important, however, is what the application can sustain, rather than SRDF. SRDF/Metro can function at distances greater than 200 km; but it is unlikely applications running on the SRDF/Metro devices would tolerate the resulting response time. The key is to know the business requirements for application response time and design the solution around that. In the end, the lower the latency the better.

2.3 Bias and Witness

SRDF/Metro maintains consistency between the R1 and R2 during normal operation. If, however, a device or devices go not ready (NR) or connectivity is lost between the arrays, SRDF/Metro selects one side of the environment as the “winner” and makes the other side inaccessible to the host(s). There are two ways that SRDF/Metro can determine a winner: bias or SRDF/Metro Witness (array or virtual). The bias or witness prevents any data inconsistencies (e.g., split brain) which might result from the two arrays being unable to communicate.

Bias is a required component of SRDF/Metro, with or without a witness. Witness builds upon the bias functionality – in essence bias becomes the failsafe in case the witness is unavailable or fails. The initial `createpair` operation of SRDF/Metro will assign bias to the R1 site though it is possible to change it to the R2 after initial synchronization. Note changing the bias turns the R2 into the R1. In the event of a failure when using bias, SRDF/Metro makes the non-biased side inaccessible to the host(s) while the bias side (R1) survives. Bias is denoted by the state of “ActiveBias” on a device pair or SRDF group as in Figure 3. Note the SRDF Type is “R1”.

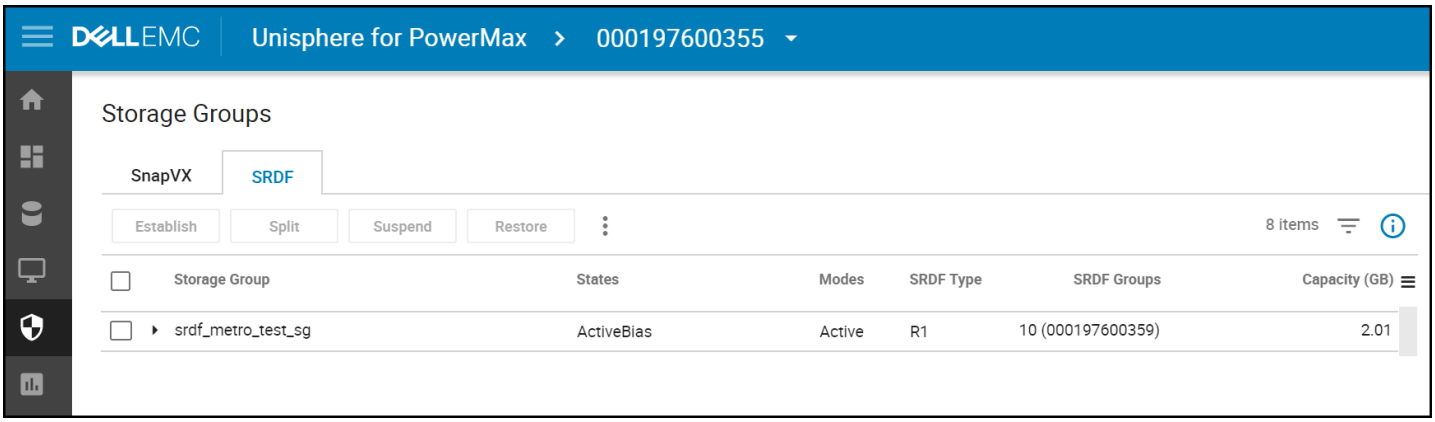

Figure 3. ActiveBias state for SRDF/Metro

If the bias side (R1) experiences the failure, then the entire SRDF/Metro cluster becomes unavailable to the hosts and will require user intervention to rectify. To avoid these types of failures, Dell offers a witness. The witness is an external arbiter that can be either array (physical) or virtual. An array witness runs on a separate VMAX/VMAX3/VMAX All Flash/PowerMax with the proper code. A separate SRDF group is created from each SRDF/Metro array (R1, R2) to the witness array and marked as witness or quorum group. In contrast, the virtual witness, or vWitness, is supplied as part of a VMware virtual appliance or as software running on a support operating system (physical or virtual). Unlike an array witness, the vWitness does not require an additional array, though it does require the eManagement Guest OS on the arrays as it communicates with a running daemon there. Both arrays in the SRDF/Metro pair must be able to communicate with the vWitness. Best practice is to place either type of witness at a third location (fault domain with separate power/network) so that it will not be subject to a failure impacting one or the other SRDF/Metro sites. Dell does not recommend using the arrays hosting the R1 or R2 of the SRDF/Metro relationship for the vWitness. It can cause undesirable failover results.

2.3.1 Priority

The use of the witness supersedes the bias functionality. If the SRDF witness groups are present or a virtual witness is configured before the `createpair` command is executed, the device pair(s) will automatically enter a “Witness Protected” state upon synchronization and the state will be “ActiveActive”. The ActiveActive state for SRDF/Metro for the same pair in Figure 3 can be seen in Unisphere in Figure 4 after a witness is added.

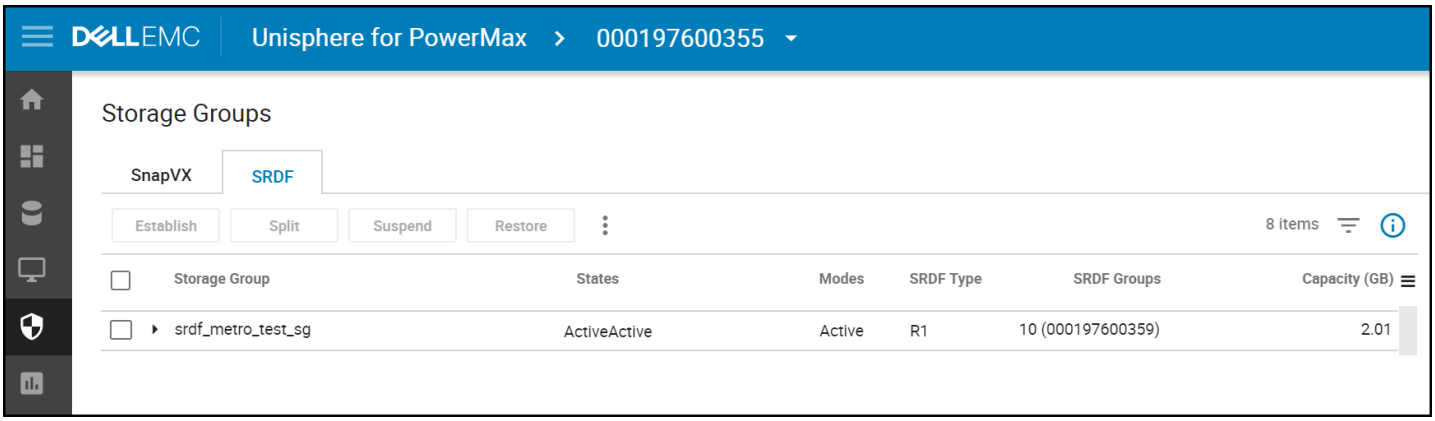

Figure 4. ActiveActive state for SRDF/Metro

Alternatively, if the witness is added after synchronization, at the next re-establish the witness will take effect. A pair, however, cannot be reconfigured to use a witness without suspending the SRDF/Metro pair first.

2.3.2 Witness latency

Each witness should be within 40 ms of the SRDF/Metro arrays. Use of a witness in a public cloud requires an RPQ.

2.3.3 Preferred winner

SRDF/Metro uses the concept of a preferred winner in the event of a failure. What this means is that one side is granted the privilege of requesting the lock from the witness instance first. When either array is running HYPERMAX OS 5977, the preferred winner is always the R1. When both arrays are running PowerMaxOS 5978 and higher, there are other factors taken into consideration. These factors include whether one of the legs has an SRDF relationship to another array, whether

the device is presented to a host, etc. Therefore, it is possible that what once was the R2 is changed to an R1 during creation or re-establish. For more information, please see the Dell SRDF/Metro vWitness Configuration Guide.

Note: It is not possible to choose a particular witness when configuring SRDF/Metro pairs. The arrays select the witness according to a set of conditions.

2.3.3.1 Cascaded and concurrent

As noted above, changes in the way the preferred winner is selected can impact which device of the pair is assigned the R1 and thus bias. When there is a third leg off of one of the devices of an SRDF/Metro pair, that device is likely to be assigned the preferred winner, or R1. Therefore, when using a witness, it is unlikely (there are other factors as explained) that a cascaded setup will remain that way, rather it will become concurrent. This can be particularly problematic when using VMware SRM, and therefore the SRDF SRA does not support cascaded with SRDF/Metro. If a cascaded setup is essential, bias is the only way to guarantee it, i.e., no witness. This, of course, is not recommended by Dell, but is an option for customers if their business demands it.

2.3.4 Multiple witnesses

Dell recommends configuring multiple array and/or virtual witnesses. A total of 32 vWitnesses are supported. In such cases SRDF/Metro handles the use of multiple witnesses so if the initial one fails, no user intervention is required to enable a secondary one, rather the code will pick another. Note that array witnesses will always take precedence over virtual ones; however, SRDF/Metro is able to use either in the event of a witness failure. Figure 5, Figure 6 and Figure 7 all show the same two vWitnesses configured on three different arrays. Both witnesses are in a good **State** and **Alive**, but only one virtual witness is **In Use**, while the other essentially in standby mode. This means there are SRDF/Metro pairs on each of these arrays, yet they all can use the same virtual witness.

Note: Do not configure a virtual witness on an SRDF/Metro pair that could use that very witness. The array has no way to check for this condition.

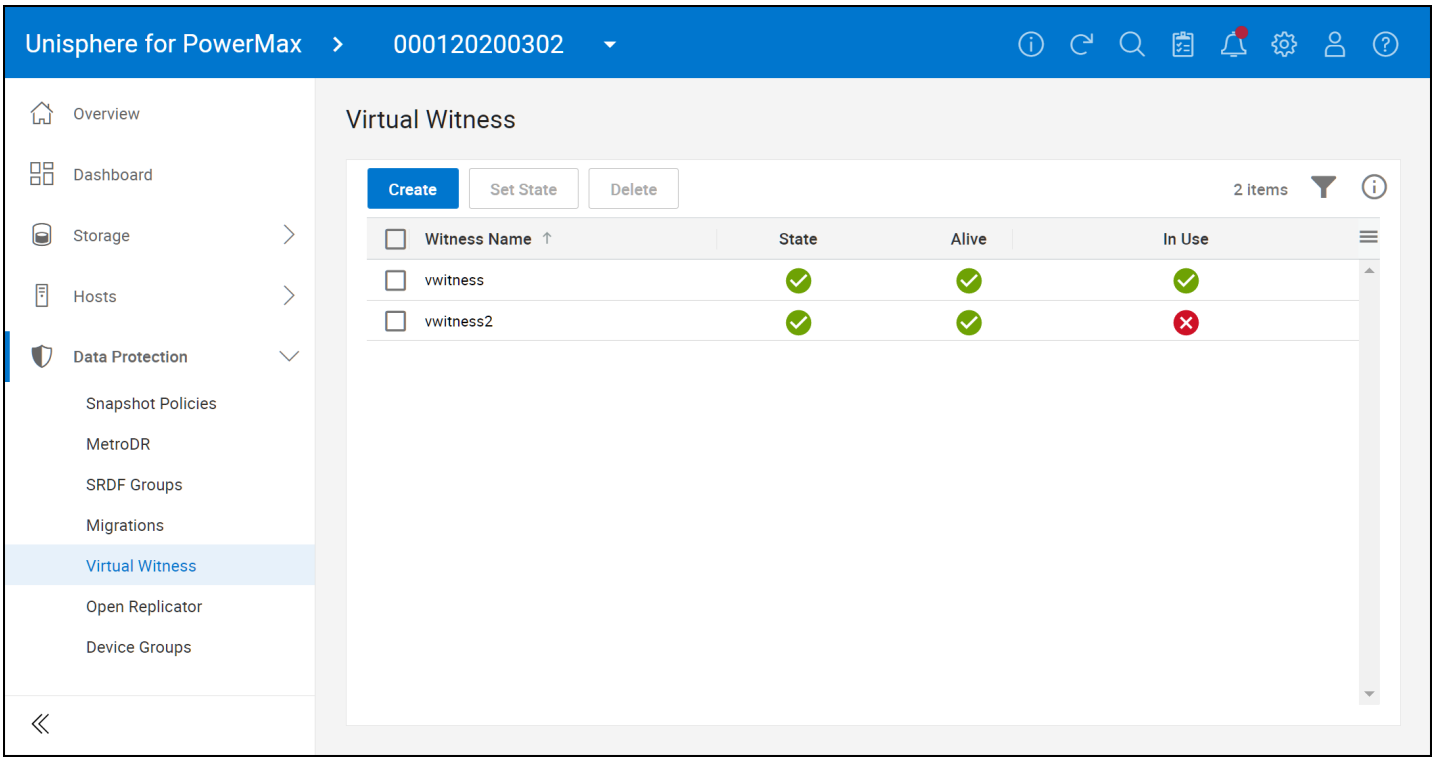

Figure 5. Virtual witnesses on array 000120200302

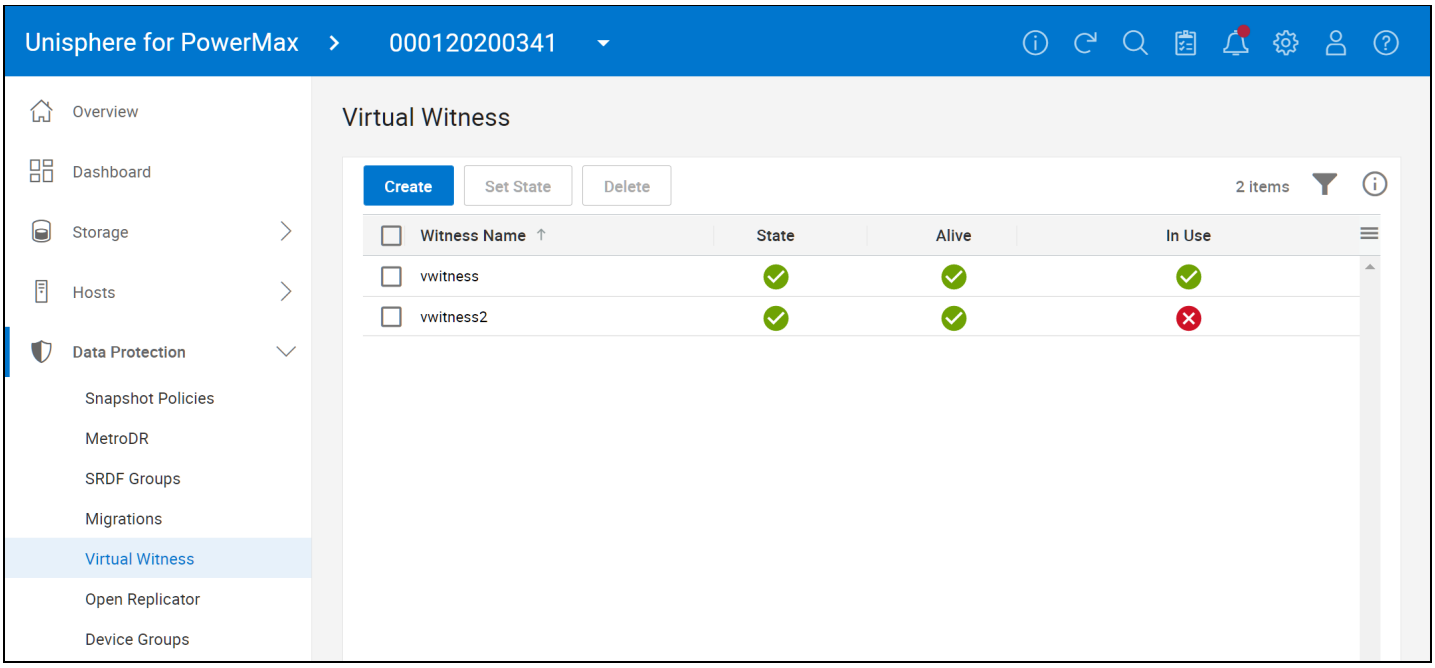

Figure 6. Virtual witnesses on array 000120200341

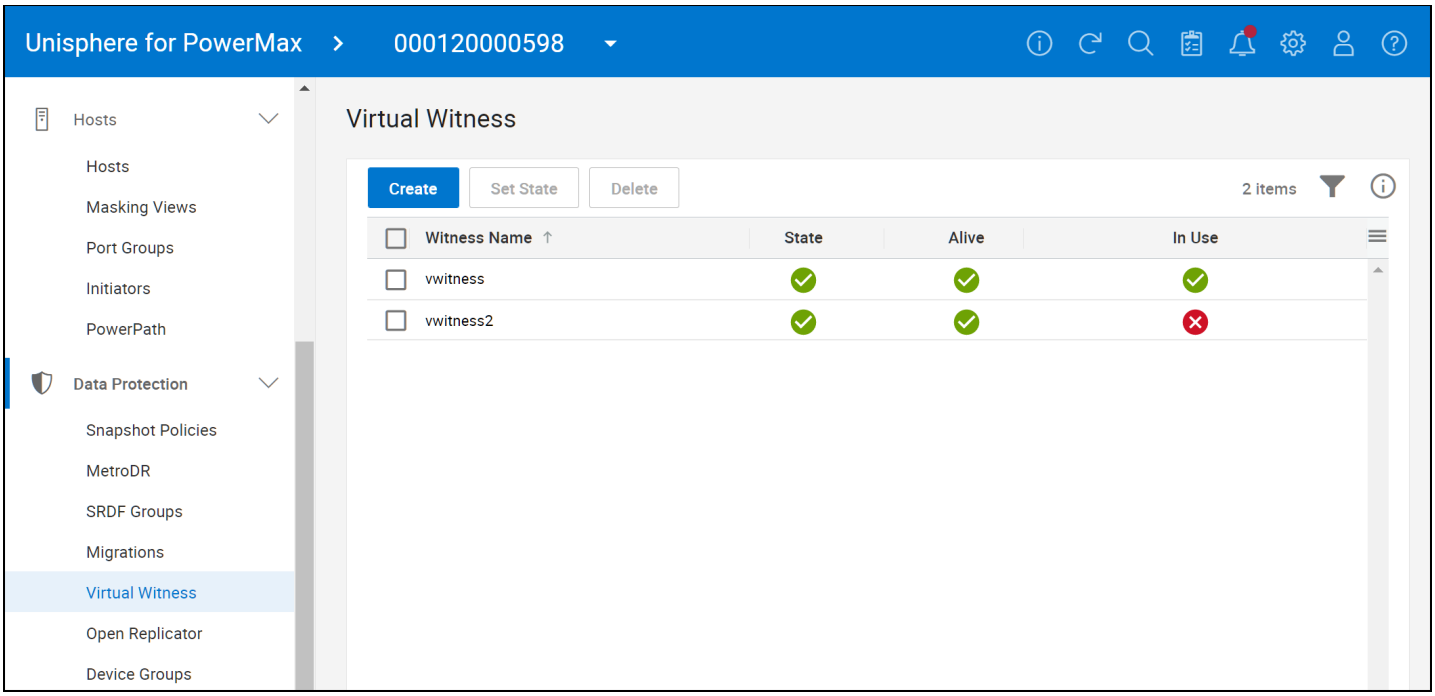

Figure 7. Virtual witnesses on array 000120000598

Note that only virtual witnesses appear in the above screen. To determine if there are array witnesses, navigate to the **SRDF Groups** screen and view existing groups where the **Type** will be labeled as **Witness**. In Figure 8, two array witnesses are identified.

Note: The array witnesses in Figure 8 are shown only as an example. The arrays in which the SRDF/Metro pairs are created should not be used as array witnesses for those very pairs. This is important as if the array witnesses exist, they will be used. The code will not prevent it.

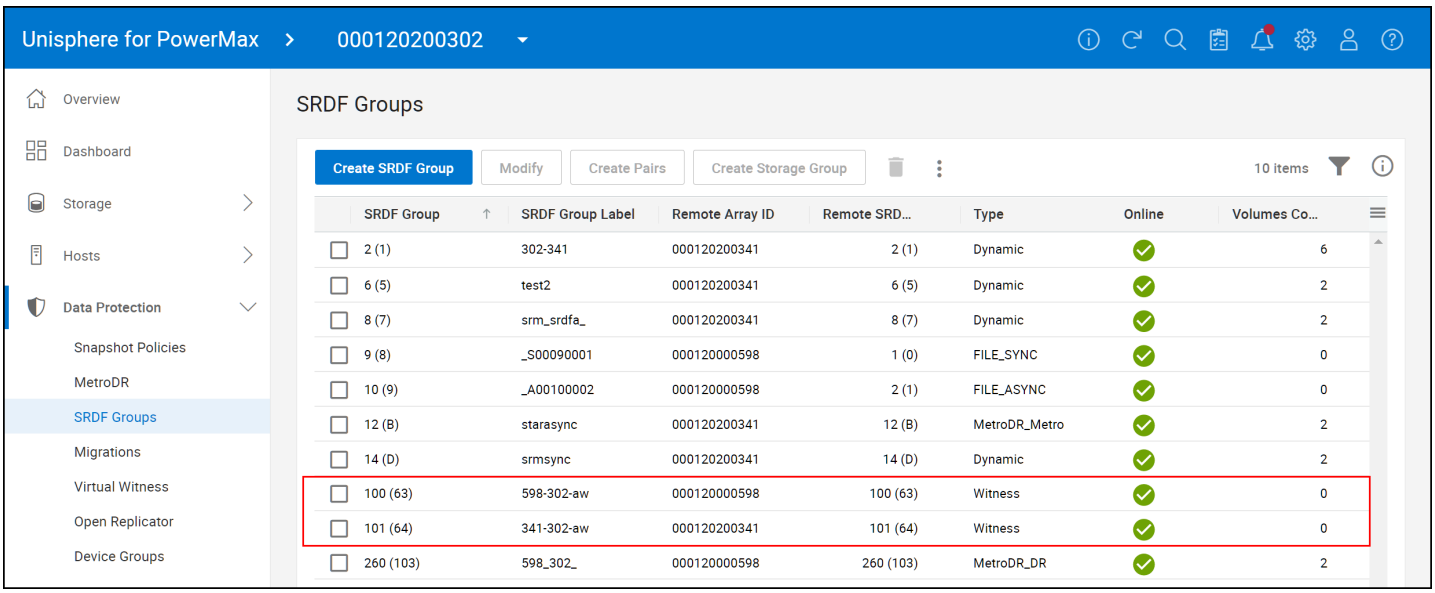

Figure 8. Array witnesses

The SRDF/Metro vWitness is available for VMAX3/VMAX All Flash storage arrays running HYPERMAX OS 5977 Q3 2016 Service Release and Solutions Enabler / Unisphere for VMAX 8.3 or later. Prior to PowerMaxOS 10, the vWitness is packaged as a VMware virtual appliance (vApp) for installation directly into the customer environment. This package is part of Unisphere for VMAX/PowerMax and Solutions Enabler vApps. Generally, the Solutions Enabler vApp is recommended due to its lower hardware requirements for those not requiring an external instance of Unisphere. Once installed, the vWitness will then utilize the local Embedded Element Manager (EEM) installed on each pair of VMAX3/VMAX All Flash/PowerMax arrays. The Witness Lock Service daemon is shown in Figure 9.

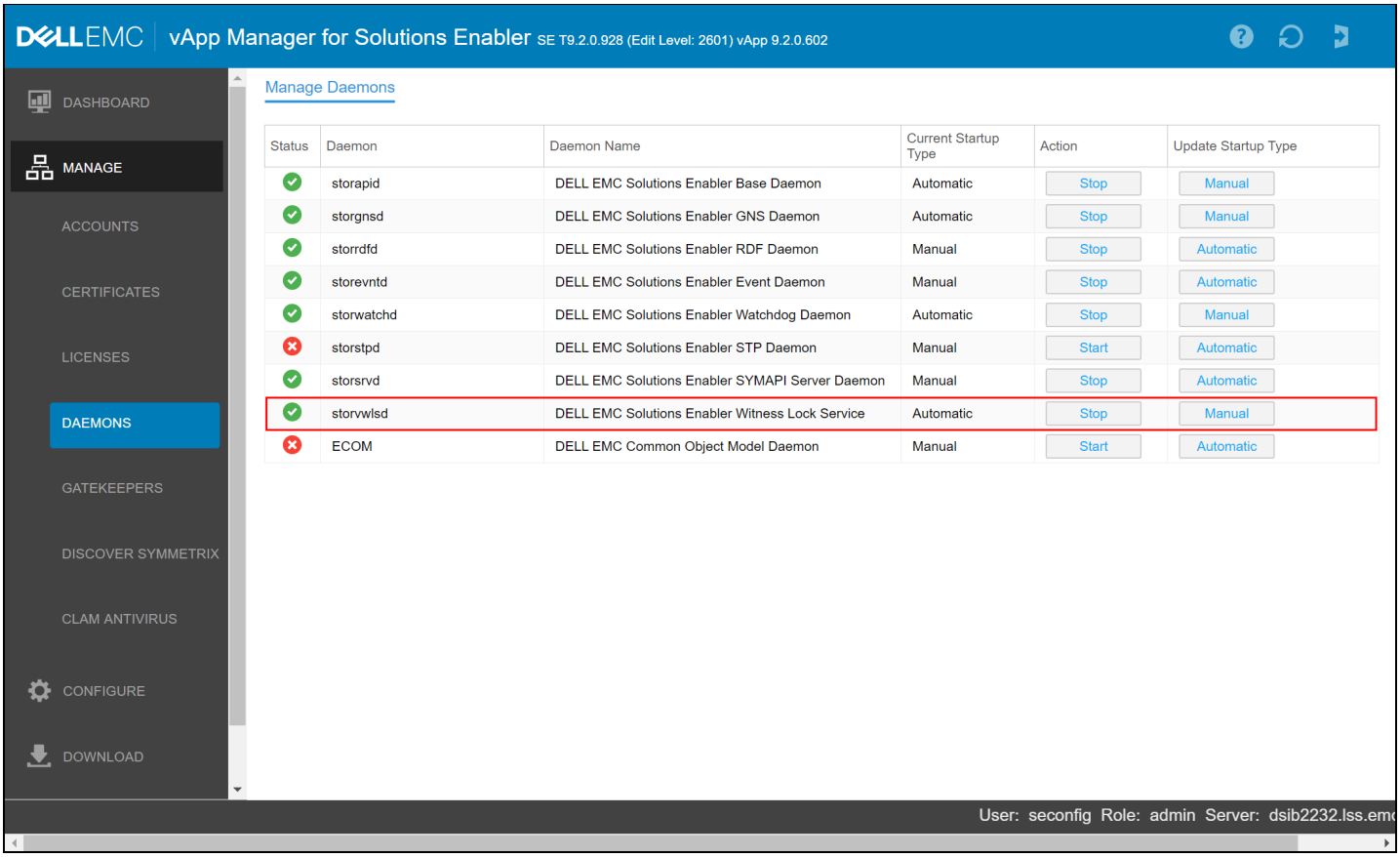

Figure 9. vWitness running on the vApp - Witness Lock Service

In PowerMaxOS 10, the vApp has been deprecated. Instead, the vWitness is delivered as a software package to be installed on a supported operating system. The Witness Lock Service Daemon is shown in Figure 10.

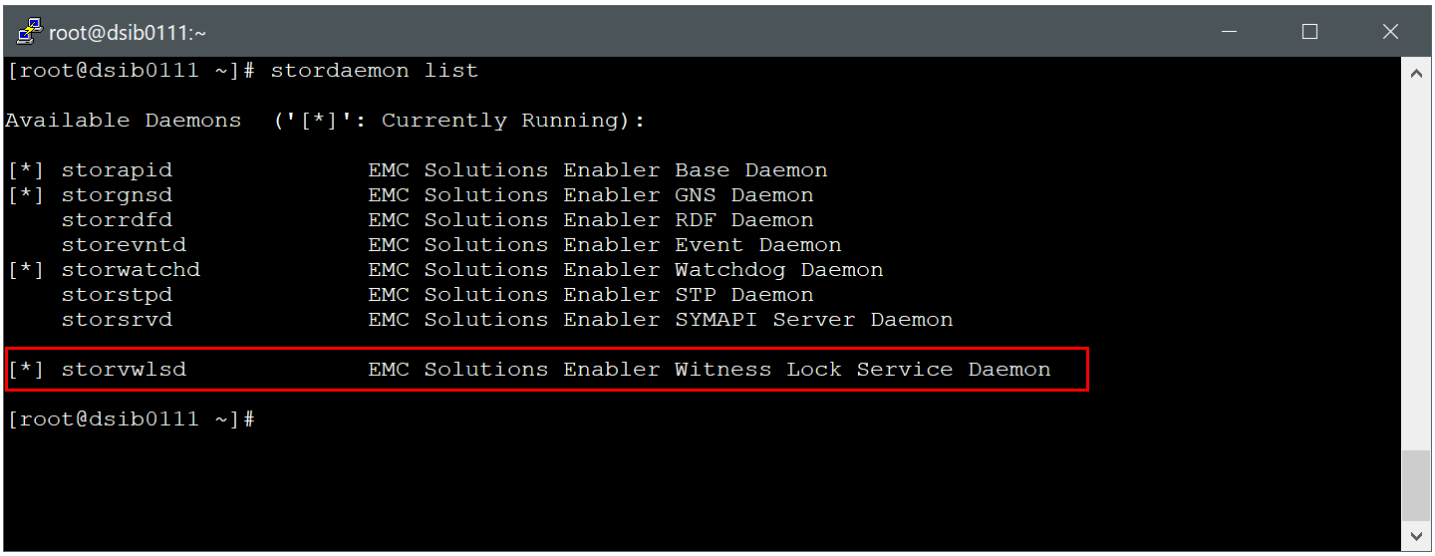

Figure 10. vWitness running on a support OS - Witness Lock Service

2.3.5 Witness failure

The screenshot in Figure 12 provides an example of a failure of the Witness where the two virtual witnesses in this configuration are down as seen in Figure 11.

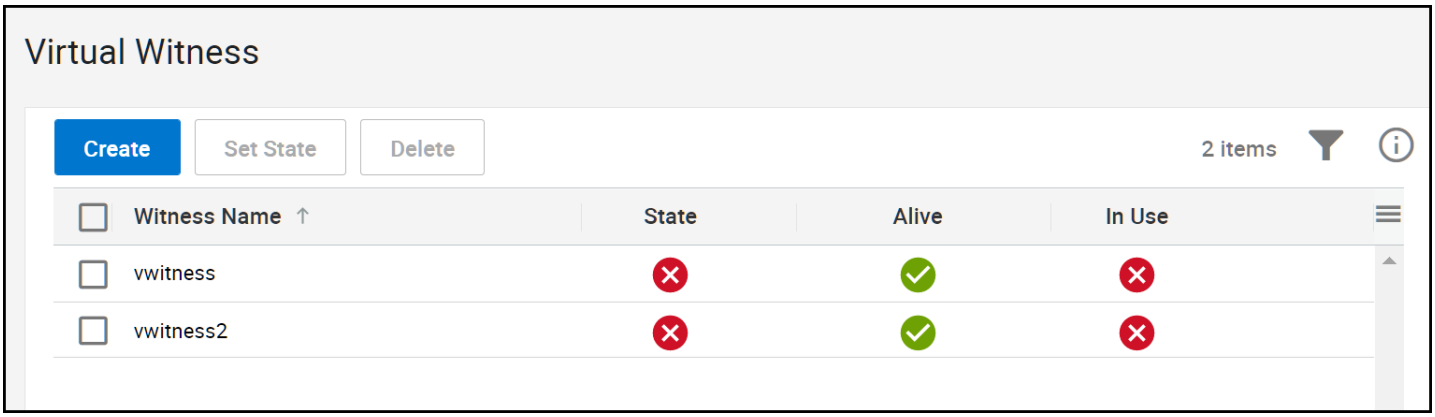

Figure 11. Failed virtual witnesses

Note that for the SRDF group in Figure 12 the configured type(C) is Witness, but the effective type(E) is Bias. This is due to the Witness status(S) being Failed.

Figure 12. An SRDF/Metro group with a failed Witness

It is also possible for the Witnesses to be impacted in some way, yet still reachable. In such cases there would be a “D” for the status or degraded. But once the issue with the Witness is resolved, be it failed or degraded, the group automatically is returned to an effective Witness state and a Normal status as in Figure 13. No suspension of the pairs is required.

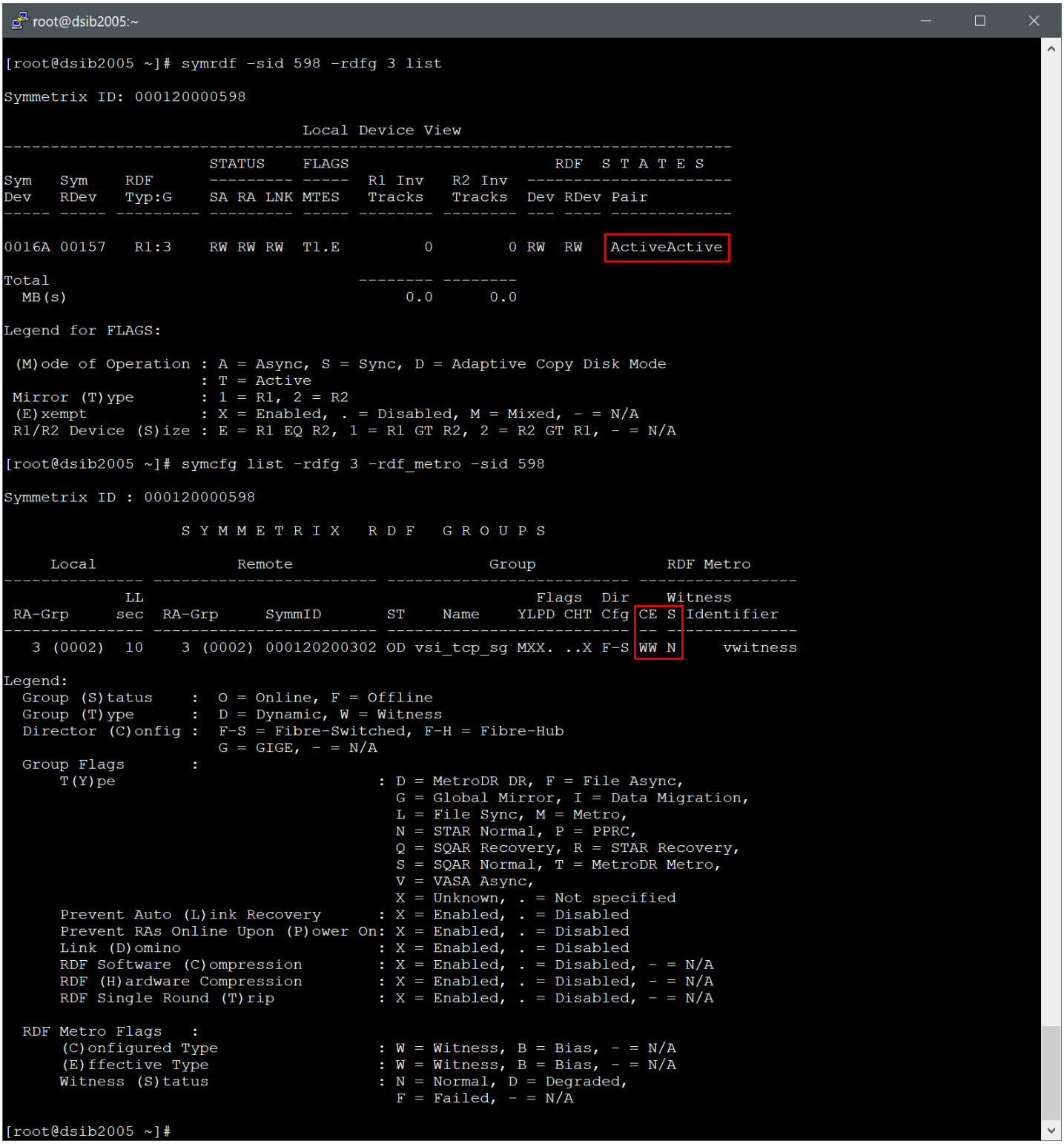

Figure 13. Restoring the Witness

The use of a witness instead of bias with SRDF/Metro and vMSC is strongly recommended by Dell.

2.4 SRDF/Metro limitations

As SRDF/Metro is an active/active mode of SRDF implementation rather than active/passive, there are some restrictions which do not exist for other SRDF modes. Some are included here for reference but for a complete list refer to *SRDF/Metro Overview and Best Practices Technical Notes* in the References section. Note that this list is based on PowerMaxOS 10 release of SRDF/Metro.

- Both the source (R1) and target (R2) arrays must be running HYPERMAX OS 5977.691.684 or higher for SRDF/Metro, or PowerMaxOS 5978 Q3 2020 or higher on all three arrays if running an MetroDR configuration.
- The R1 and R2 must be the same size.
- Devices cannot have Geometry Compatibility Mode (GCM) set prior to PowerMaxOS Q2 2018 SR.
- Devices cannot have User Geometry set.
- Concurrent and cascaded SRDF/A configurations are only supported with the HYPERMAX OS Q3 2016 SR and later.
- Controlling devices in an SRDF group that contain a mixture of source (R1) and target (R2) devices is not supported.
- The following operations must apply to all devices in the SRDF group: `createpair -establish, establish, restore, and suspend`
- vWitness configurations require Embedded Element Management (EEM or eMgmt) on each SRDF/Metro paired array.
- SRDF/Metro does not support Star configurations
- SRDF/Metro does not support NVMeoF
- MetroDR does not support online device expansion. MetroDR must be disabled first, and expansion done on the async R2 first, then R1, after which MetroDR can be re-enabled.
- Microsoft Cluster with SRDF/Metro requires a uniform configuration.

2.4.1 VAAI support

While SRDF/Metro supports all VAAI commands[2], the implementation of Full Copy or XCOPY, by necessity, was modified to accommodate the active/active nature of SRDF/Metro. These changes are noted below.

[2] Though technically not a VAAI command, SRDF/Metro does not support ODX.

2.4.1.1 SRDF/Metro XCOPY specifics

The following are caveats for SRDF/Metro when using Full Copy (XCOPY). In SRDF/Metro configurations the use of Full Copy entirely depends on whether the site is the one supplying the external WWN identity.

- Full Copy will not be used between a non-SRDF/Metro device and an SRDF/Metro device when the device WWN is not the same as the external device WWN. Typically, but not always, this means the non-biased site (recall even when using witness, there is a bias site). In such cases Full Copy is only supported when operations are between SRDF/Metro devices or within a single SRDF/Metro device; otherwise, software copy is used.

- As Full Copy is a synchronous process on SRDF/Metro devices, which ensures consistency, performance will be closer to software copy than asynchronous copy on SRDF/A devices.

- While an SRDF/Metro pair is in a suspended state, Full Copy reverts to asynchronous copy for the target R1. It is important to remember, however, that Full Copy is still bound by the restriction that the copy must take place on the same array. For example, assume SRDF/Metro pair AA, BB is on array 001, 002. In this configuration device AA is the R1 on 001 and its WWN is the external identity for device BB, the R2 on 002. If the pair is suspended and the bias is switched such that BB becomes the R1, the external identity still remains that of AA from array 001 (i.e., it appears as a device from 001). The device, however, is presented from array 002 and therefore Full Copy will only be used in operations within the device itself or between devices on array 002.

3 VMware vSphere Metro Storage Cluster (vMSC)

A VMware vSphere Metro Storage Cluster configuration is a VMware vSphere certified solution that combines synchronous replication with array-based clustering. These solutions typically are deployed in environments where the distance between datacenters is limited, often metropolitan or campus environments. Dell SRDF/Metro represents one of those certified solutions.

A VMware vMSC requires, what is in effect, a single storage subsystem that spans both sites. In this design, a given datastore must be accessible simultaneously from both sites. Furthermore, when problems occur, the surviving ESXi hosts must be able to continue to access datastores from the surviving array transparently and do so without impact to ongoing storage operations.

The storage subsystem for vMSC must be able to be read from and write to both locations. For active/active solutions, all disk writes are committed synchronously at the two locations to ensure that data is always consistent regardless of the location from which it is being read. This storage architecture requires significant bandwidth and very low latency between the sites involved in the cluster. Increased distances or latencies cause delays writing to disk, making performance suffer dramatically, and prevent vMotion activities between the cluster nodes that reside in separate locations.

Note: VMware (and therefore SRDF/Metro on VMware) supports 64 hosts in a cluster in vSphere 6 and 7.0, and 96 hosts in 7.0U1 and 8.

3.1 VMware vMSC with SRDF/Metro

Dell SRDF/Metro has the ability to concurrently access the same set of devices independent of the physical location and thus enables geographically stretched clusters based on VMware vSphere. This forms the basis of a vSphere Metro Storage Cluster. This allows for transparent load sharing between multiple sites while providing the flexibility of migrating workloads between sites in anticipation of planned events such as hardware maintenance. Furthermore, in case of an unplanned event that causes disruption of services at one of the data centers, the failed services can be quickly and easily restarted at the surviving site with minimal effort with the combined functionality of vSphere HA. Nevertheless, the design of the VMware environment has to account for a number of potential failure scenarios and mitigate the risk for services disruption.

3.1.1 Architecture

At a high conceptual level, a vMSC utilizing SRDF/Metro (witness is assumed) would appear similar to Figure 14.

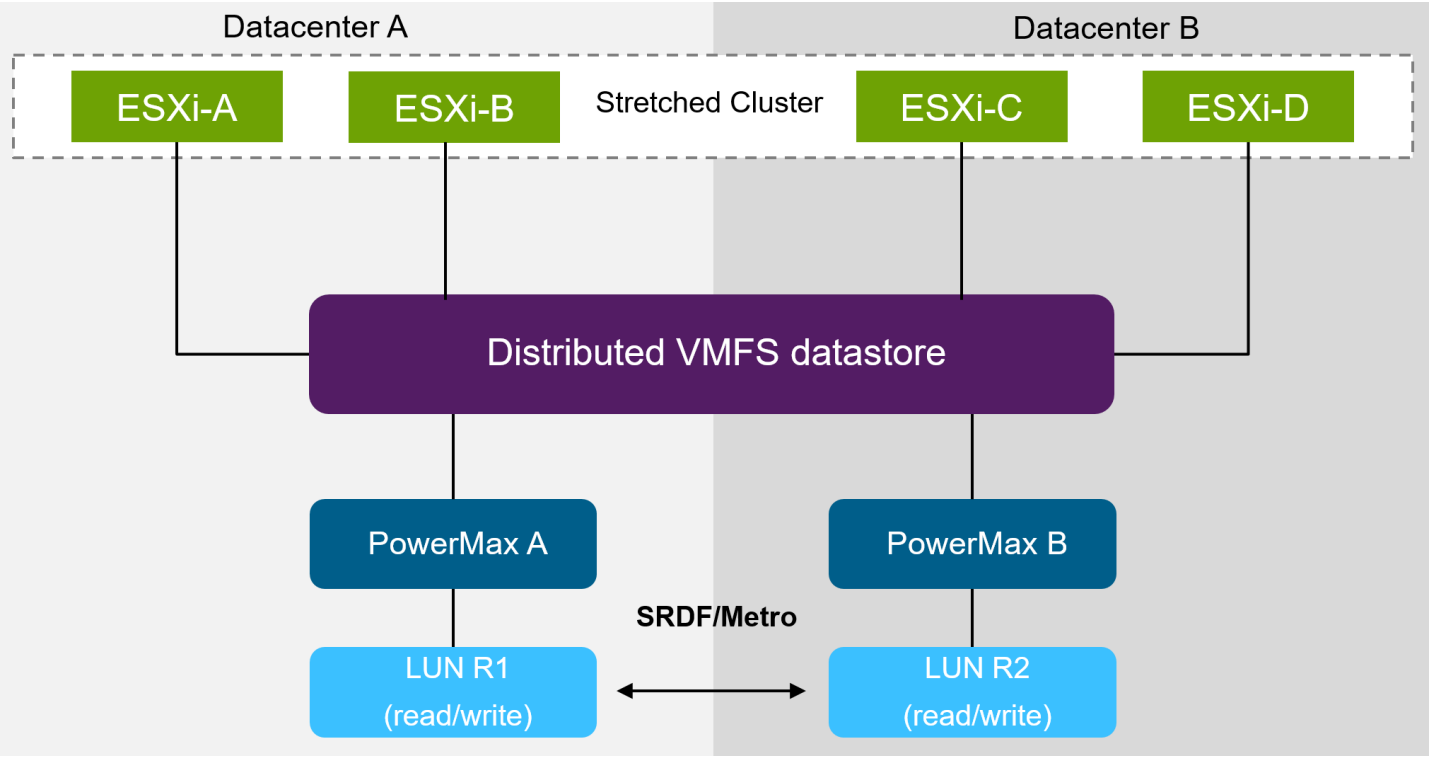

Figure 14. vMSC nonuniform configuration with SRDF/Metro

ESXi hosts A – D, despite being in separate datacenters, are in the same vCenter. Half the hosts see the R1 on PowerMax A, the other hosts see the R2 on PowerMax B; however, they all see the same VMFS datastore because from VMware's perspective, the R1 and R2 are the same device (i.e., they share an external WWN).

SRDF/Metro maintains the cache state on each array, so an ESXi host in either datacenter detects the "virtual" device as local. Even when two virtual machines reside on the same datastore but are located in different datacenters, they write locally without any performance impact on either of them. Each host accesses and is served by its own array with SRDF/Metro ensuring consistency. If a witness is not in use, there is site bias, meaning that one site will be the winner in the event of failure.

The zoning of the ESXi hosts in a vMSC can be handled in one of two ways, these are covered next.

3.2 Uniform and Nonuniform vMSC

There are two types of configurations available for vMSC: uniform and nonuniform. Dell supports either configuration. VMware defines them as:

- Uniform host access configuration – When ESXi hosts from both sites are all connected to a storage node in the storage cluster across all sites. Paths presented to ESXi hosts are stretched across distance.
- Nonuniform host access configuration – ESXi hosts in each site are connected only to storage node(s) in the same site. Paths presented to ESXi hosts from storage nodes are limited to the local site.

It should be noted that some uniform vMSC configurations are really active/passive configurations where only one device of the pair is active at one time. In essence only one array owns the device pair at one time. This is also referred to as site bias (unrelated to SRDF/Metro) or LUN locality. No matter which host the IO comes from, it is properly directed to the active side. In these configurations it is typical to split equally the device pair ownership across the two storage environments, but of course both datacenters see both storage arrays. SRDF/Metro does not have site bias since both sides are active.

3.2.1 General recommendation

While Dell supports both uniform and nonuniform configurations without restriction, Dell recommends nonuniform configurations. There are a number of reasons for this. Uniform or cross-connect configurations add to the complexity of the SAN design (bridging between sites) and can lead to zoning issues. Losing one of the datacenters will also generate SAN fabric events which will impact surviving hosts, and perhaps VMware HA taking place. From the hardware standpoint, and short of a complete disaster, it is also unlikely the array is the component that is going to fail in the datacenter, rather it will be the servers. Most arrays, but in particular the PowerMax, are incredibly resilient and in fact in the event of RAID failures in SRDF configurations, the PowerMax can even read from the remote array. Keeping the two environments isolated will reduce complexity and minimize any chance of one impacting the other. And finally, and most importantly, having servers in one datacenter accessing the array in the other datacenter will be adding latency to the application unless properly addressed in pathing. Pathing, therefore, is critical in a vMSC and will be discussed next.

Note: If the vMSC environment is part of SmartDR and utilizing VMware SRM, Dell recommends a uniform configuration to take advantage of the features of the SRDF SRA.

3.3 Pathing considerations

There are two available technologies for path management in an SRDF/Metro vMSC: PowerPath/VE or PP/VE, and Native Multipathing or NMP. Dell recommends PP/VE for all VMware environments, and in particular vMSC, because of its advanced algorithms for load balancing and failure detection, among others. Whichever pathing technology is utilized, however, there are considerations that should be addressed before implementation depending on whether a uniform or nonuniform vMSC is being used.

3.3.1 PowerPath/VE Autostandby

Although Dell recommends a nonuniform configuration, PP/VE offers a solution for uniform or cross-connect configurations that removes the risk of latency and dead path issues. This feature is known as Autostandby or ASB. In a uniform vMSC, when ASB is enabled (default) on PP/VE, ASB is able to determine which device paths have a higher latency to storage, automatically assigning them to a standby status.

The defaults of PP/VE 7.2[3] for PowerMax arrays are shown in Figure 15.

Note: The detail supplied for PowerPath/VE is based on version 7.2. Each version of PowerPath makes improvements to functionality and as such if using a different version than 7.2, consult the documentation. This does not change the general best practices, however.

[3] PP/VE 7.0 P01 is the minimum version required for vSphere 7.0, and PP/VE 8.0 for vSphere 8.0.

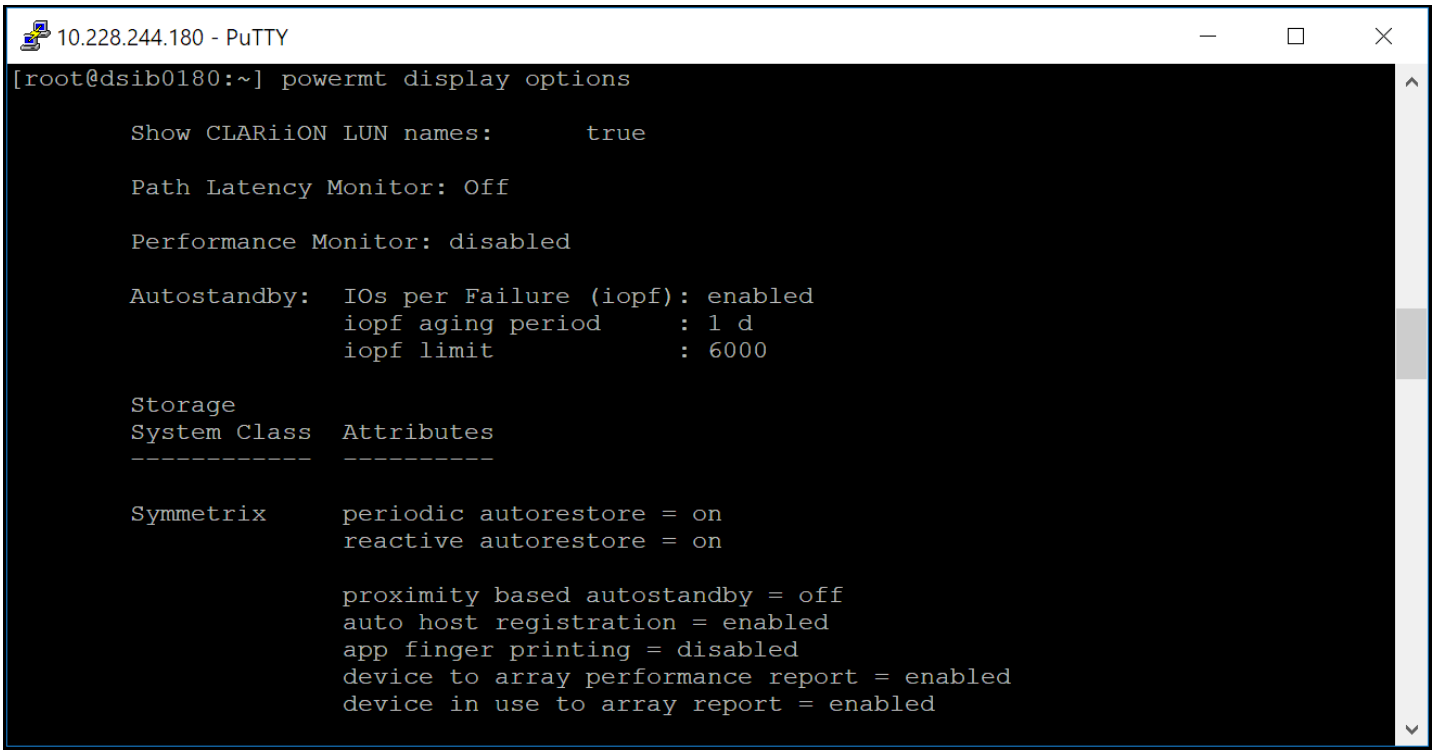

Figure 15. PP/VE defaults

Autostandby has two different modes it can use: proximity and IOs per failure. There is a third mode available known as Autostandby offline which can detect maintenance on ports but it does require a recent version of PowerMaxOS and as it is not pertinent to the vMSC discussion, it will not be covered. The other two modes, however, will be covered here. Note that the Autostandby feature has changed a number of times since it was introduced. Be sure to review the PowerPath Product Guide for the version used in the environment to understand the Autostandby behavior.

Note: Autostandby paths do not persist through reboot. After reboot PowerPath will re-test the paths for latency and assign Autostandby as appropriate.

3.3.1.1 IOs per failure

The first mode is IOs per failure which is enabled by default and labeled as ***asb:iopf***. IOs per failure is most useful in situations where there is the potential for "flaky" paths. When the path has 'x' number of failures, the path goes into standby for the aging period (which also can be adjusted). This mode is not recommended for vMSC configurations.

3.3.1.2 Proximity

The second is proximity based Autostandby which is disabled by default and labeled as ***asb:prox*** on a path. This is the mode most pertinent to a uniform vMSC setup as it can determine which paths are remote and which are local. In addition, proximity based Autostandby has a secondary option that can be supplied called threshold. The threshold is a latency value which is set and determines when a path should be set to either active or standby. By default, the threshold is zero(0), however it can be set to a value ranging from 0-5000 microseconds. With the threshold set, as long as the latency difference between two paths is above the value, PP/VE will set Autostandby, or asb:prox, on the path(s) with the higher latency. If all paths are below the threshold, then they will be set to active, and none to asb:prox. The idea behind the threshold setting is that if the environment can tolerate up to a particular latency, then it may be desirable to use a threshold value up to that latency to ensure that all paths are set to active. Note that the

default value of zero for the threshold guarantees that if you are running a uniform configuration that one of your array paths will be set to asb:prox. If it is essential that all paths are in an active state, either disable Autostandby (see below), or set the threshold to a high enough value to ensure the latency to the remote array in comparison to the local array will not exceed it.

3.3.1.3 Examples

Since Autostandby is recommended with PP/VE and uniform, the following example details how it might work in a real-world environment.

This VMware vSphere Metro Storage Cluster setup is comprised of only two hosts, each attached to its own array in different datacenters. The table in Figure 16 contains the pertinent information for the example. Note the FA ports have been configured with different numbers, so it is apparent which array is local and which is remote.

Host	Device	Site	FA Ports	Array
dsib0180.lss.emc.com	586	R1	1D:4, 2D:4	000197600357
dsib0182.lss.emc.com	FE	R2	1D:6, 2D:6	000197600358

Figure 16. Autostandby example environment

As array 357 is the R1, it will supply the external identity to device FE on array 358, the R2. In this example, then, both devices share WWN 600009700001976003575330303538 36. Each device is presented to both hosts on each FA port, resulting in 4 paths, 2 to the R1, 2 to the R2. Figure 17 shows the emcpower1 device on the R1 host prior to enabling the Autostandby feature. Note all paths are active, a true uniform configuration.

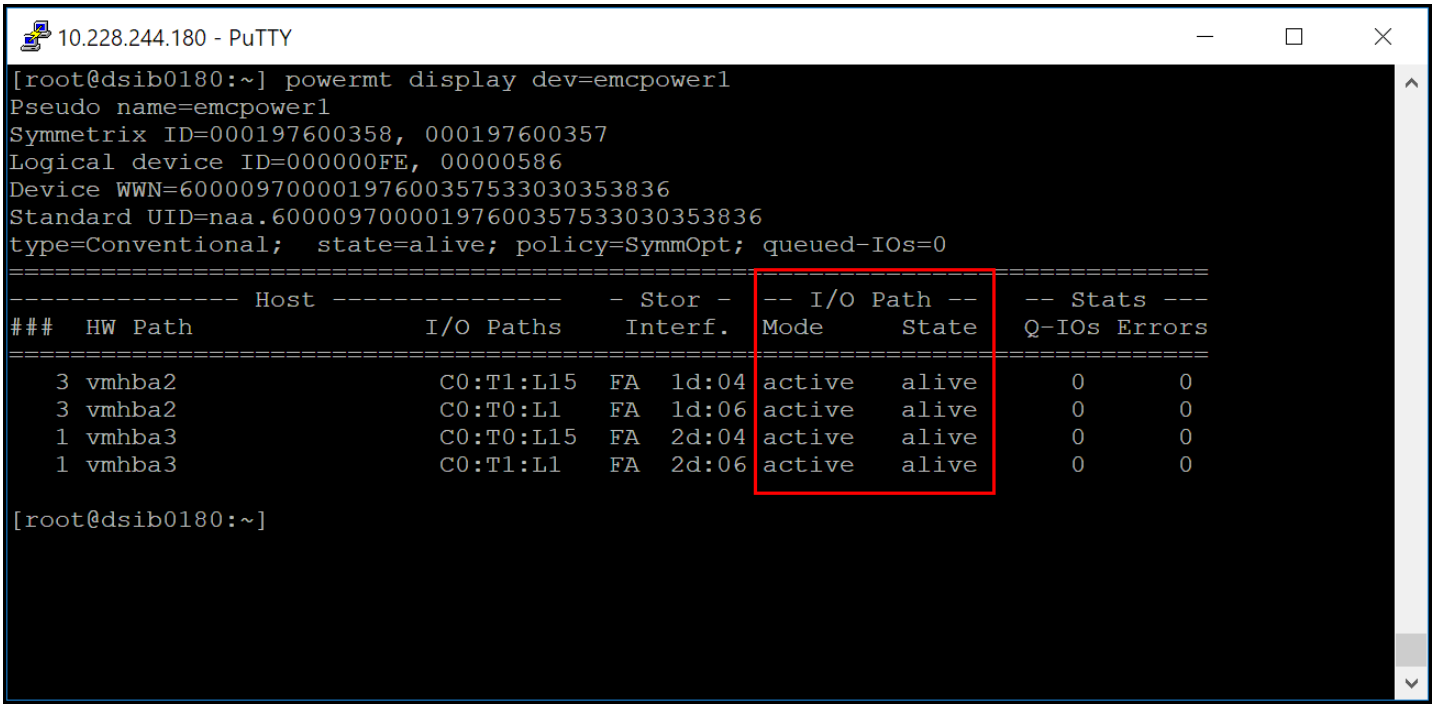

Figure 17. PP/VE path assignment prior to Autostandby activation

Now, enable Autostandby and set the trigger to “prox” or proximity. This will override the default of IOs per failure. PowerPath will immediately determine which paths are local to the host and which

are remote. Figure 18 again shows device emcpower1 on host dsib0180 after enabling Autostandby.

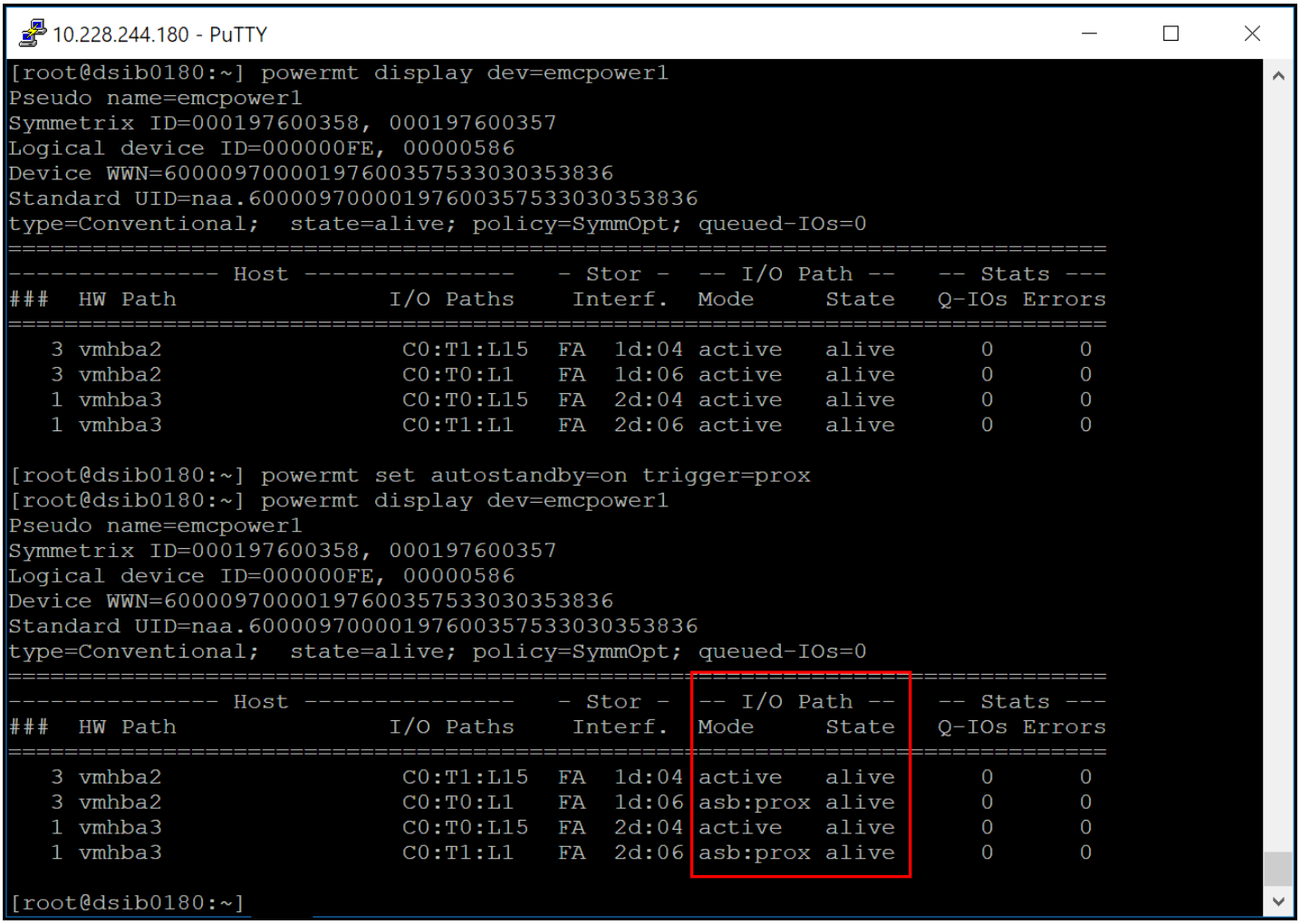

Figure 18. PP/VE path assignment in datacenter A

Notice that the two paths to the R2 device have been set to asb:prox, or Autostandby proximity, while the other two are active. This is expected. What this means is that PP/VE tested the paths (with default zero threshold) and determined that the paths going to device FE on array 358 had a higher latency than those going to device 586 on array 357 and set them accordingly for dsib0180. And on dsib0182 in Figure 19, PowerPath determines the opposite settings for emcpower2 (which represents the same devices) since array 357 is now the remote one.

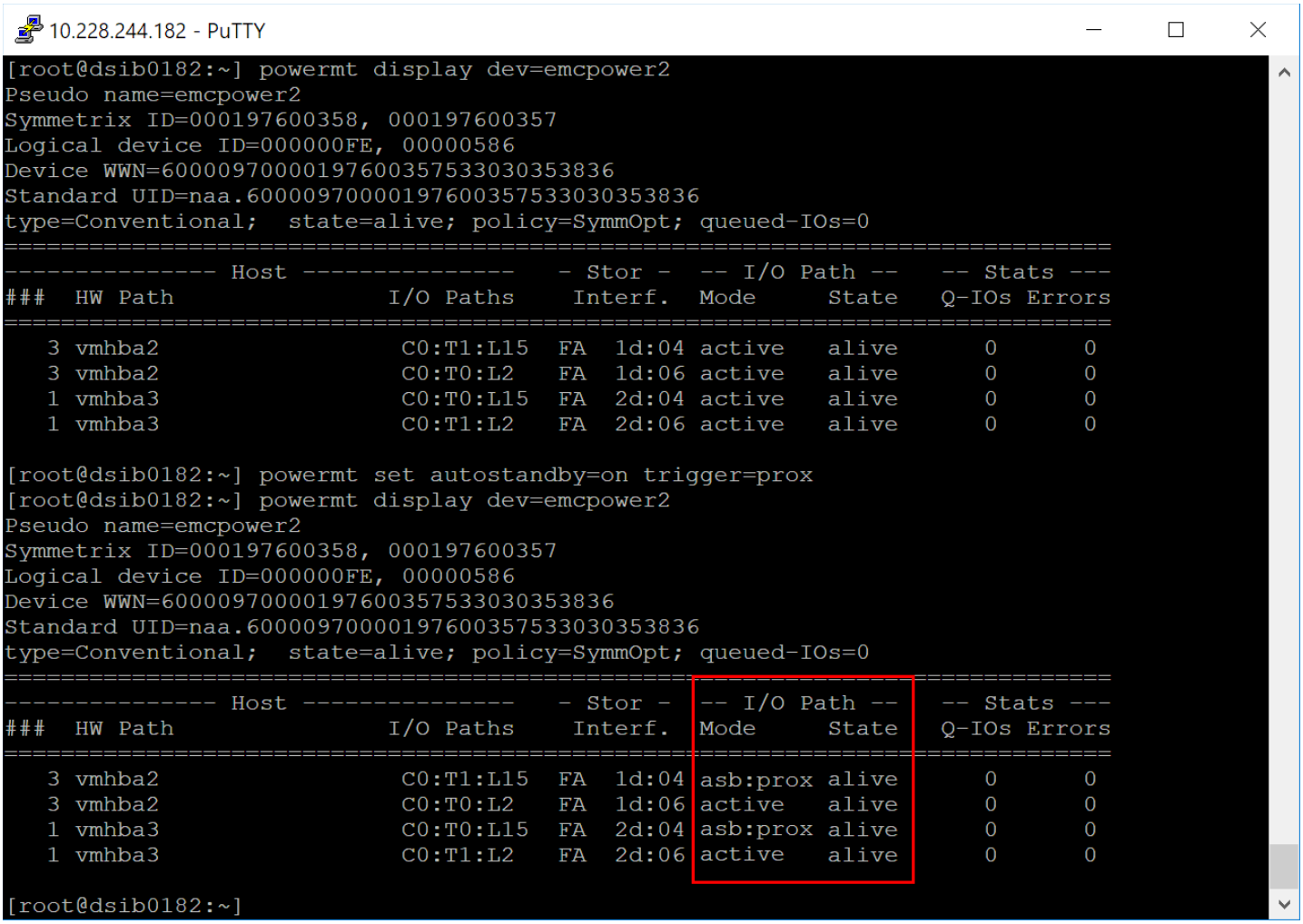

Figure 19. PP/VE path assignment in datacenter B

During normal operation, only the two active, local paths will be preferred on each ESXi host, but if there is a failure on one of the arrays, one of the ESXi hosts will have IO redirected to the asb:prox paths. Despite the fact the paths are in standby, PowerPath can still use them if it deems it necessary even without a failure. For instance, if one path to the active array becomes extremely slow, PowerPath might choose to send some IO to the standby path.

3.3.1.4 Threshold change

Some customers may wish to use a true uniform configuration while latency to the remote array in comparison to the local array remains below a certain threshold. If that is the case, the default threshold for Autostandby can be adjusted. Simply follow these steps. Note here that the optional parameter of “class” is supplied in case there are other arrays attached to the ESXi host which the customer does not wish to change.

- The threshold can be adjusted dynamically with the “reinitialize” value:

```
powermt set autostandby=reinitialize trigger=prox class=symm threshold=5000
```

After the change, PP/VE sets the new mode on the device in Figure 20:

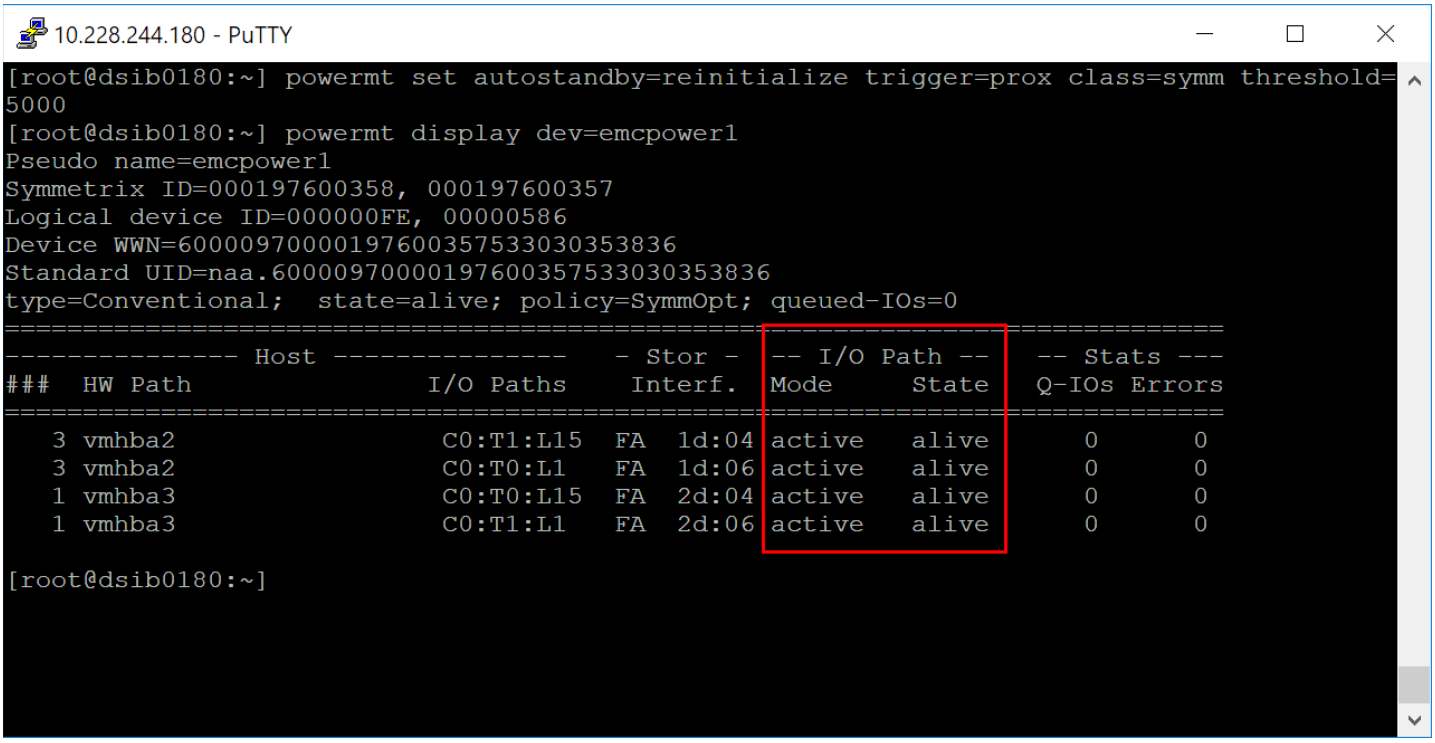

Figure 20. PP/VE settings after change in threshold

After the change, all the paths show as active since the latency difference between the arrays is lower than the threshold. If at some point during normal business operation it appears that one of the arrays is causing latency issues, issuing a reinitialize which will cause PP/VE to re-test the paths and set any paths to asb:prox that exceed the threshold:

```
powermt set autostandby=reinitialize trigger=prox class=symm
```

For customers who have a true stretched cluster where the arrays are co-located and already know they do not wish to use Autostandby, the PowerPath default configuration is appropriate.

3.3.1.5 Manual path setting

One final example of using Autostandby is a customer who has a stretched cluster (or co-located arrays) who nonetheless wishes to only use the secondary array as a standby. This environment is a bit tricky for Autostandby because as shown above, PP/VE will set paths to asb:prox in the default configuration and it may not choose the R2 device on the R1 side or the R1 device on the R2 side or if the latency is exactly the same it will set all paths to active. The best way to handle this is to not use Autostandby but rather set standby manually for the paths. The command to do that is:

```
powermt set mode=standby hba=<hba#> dev=all class=all
```

Here are a couple of examples:

```
rpowermt set port_mode=standby dev=vmhba2:C0:T0:L7

rpowermt set mode=standby hba=1 class=symm
```

If it is preferable to keep Autostandby active, changing the paths to active or standby requires the “force” option. Figure 21 demonstrates setting the mode from asb:prox to active.

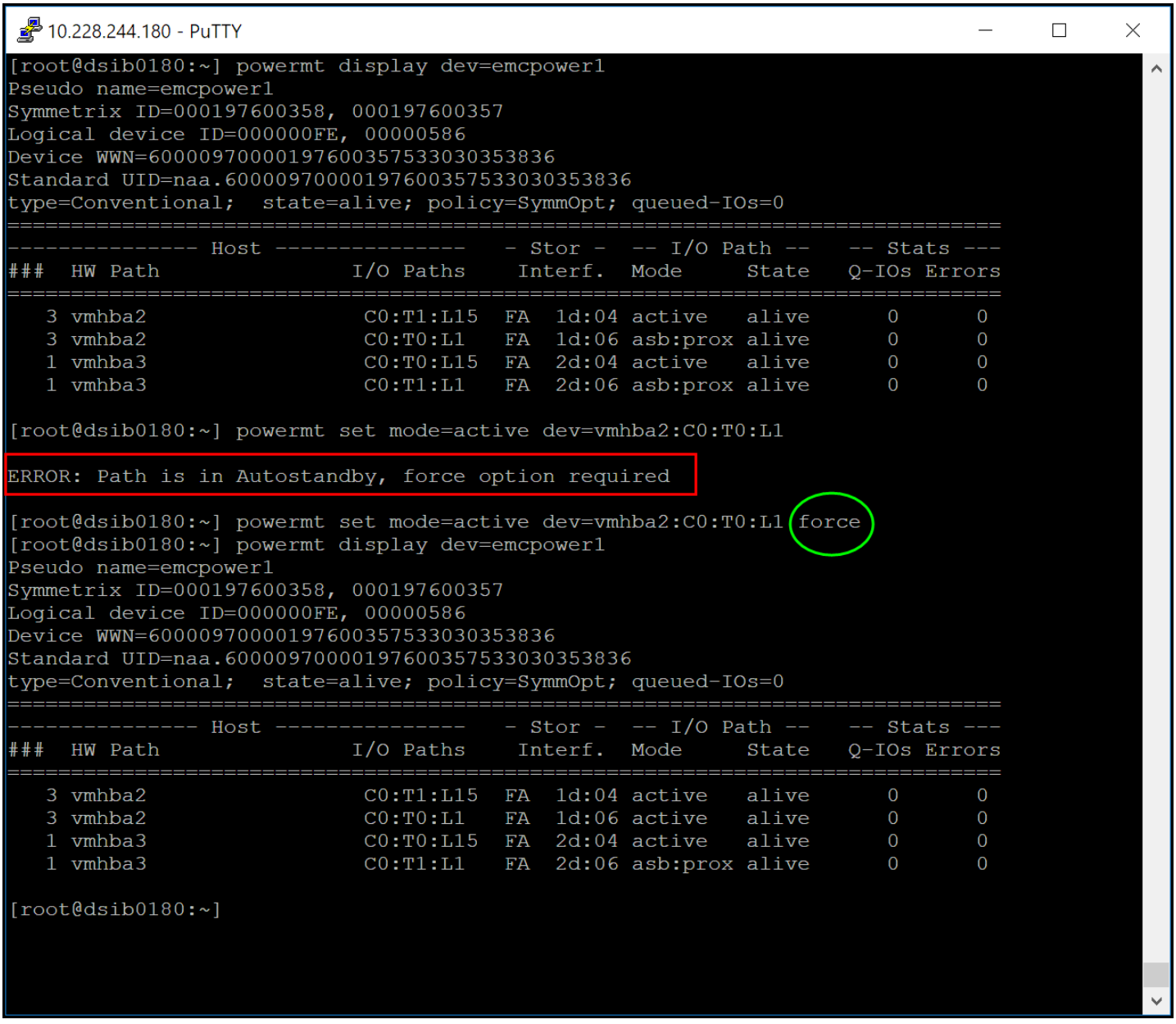

Figure 21. Forcing changes in PP/VE modes

To move a path from active to standby will not require force because the mode is not Autostandby. See this in Figure 22.

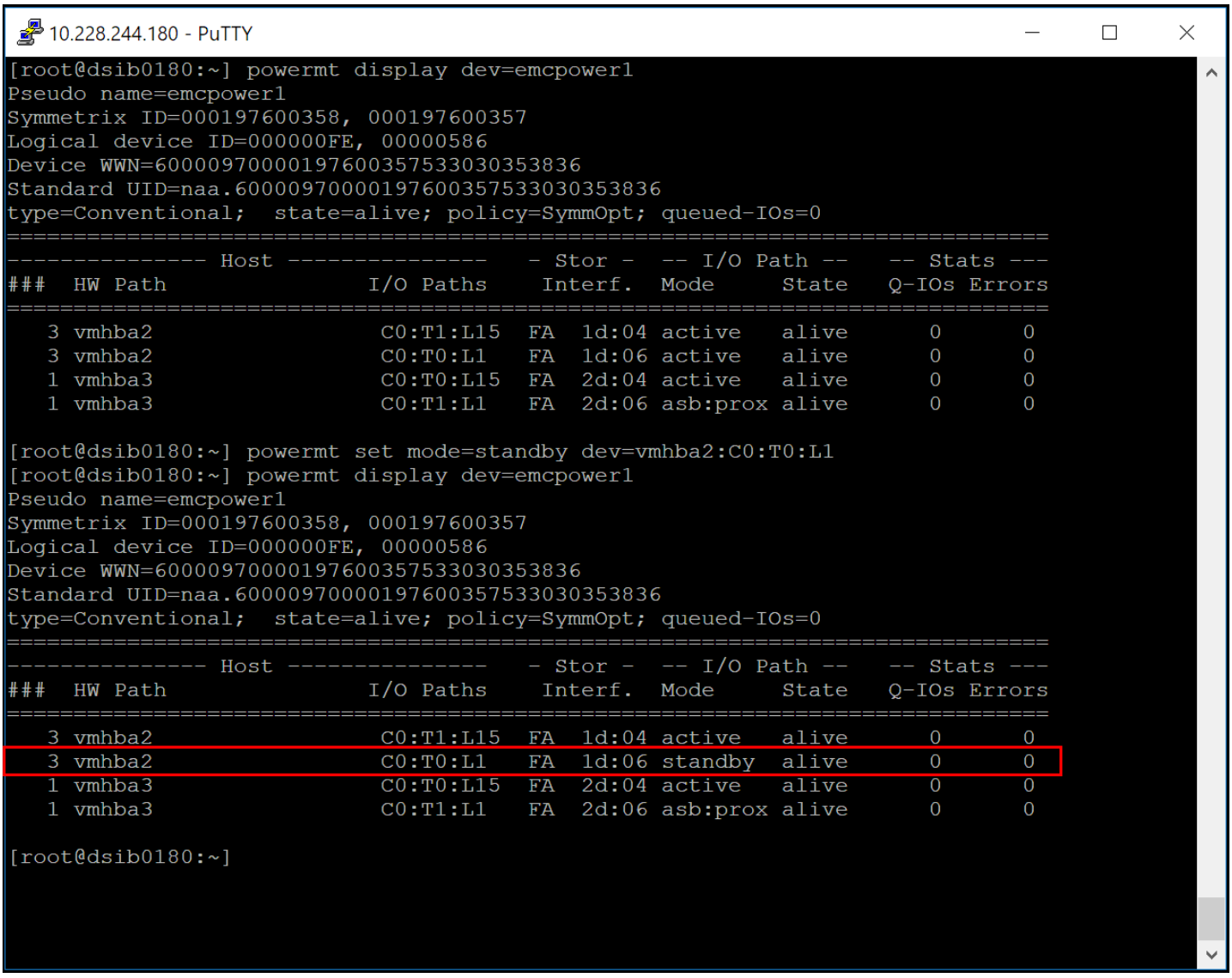

Figure 22. Changing from active to standby path in PP/VE

Note: Setting a path to standby does not mean it will only be used in the event of failure, rather that its use is heavily weighted against the active paths. If the active paths were heavily loaded, the standby path may be used.

One final caveat about Autostandby is if the user improperly assigns the R1 path to Autostandby, any non-Metro devices presented from the R1 array also will be set to Autostandby. Since all paths have the same value, PP/VE will still use them despite being asb:prox; however, the user should manually change them to active. To fix this, simply change the paths to active, remembering to use the force option. Path changes like those demonstrated above will not persist through reboot, unfortunately. It is possible to create a script to assign paths to standby upon reboot.

3.3.2 NMP for vSphere 6.7

When utilizing NMP pathing with a vMSC running SRDF/Metro, the PSP configuration will differ depending on the implementation. By default, Round Robin is the PSP for PowerMax arrays. Round Robin works just as the name suggests, VMware switches from one path to another, across

however many paths there are, sending IO. In releases prior to vSphere 6.7 U1, there are two methods to control when VMware switches paths – by the number of IOs (type=iops) or by the size of the data (type=bytes) being sent. By default, VMware uses type=iops and will send 1000 IOs down a path before switching to the next one.

3.3.2.1 Uniform

Uniform configurations present a challenge when using NMP as opposed to PP/VE because not only is NMP unable to distinguish between a local and a remote path but there is no intelligence to account for latency and congestion. As noted, the default behavior of the Round Robin PSP is to switch paths every 1000 iops. VMware and Dell best practice, however, is to set iops=1 to improve performance and error detection in VMware environments. Doing so in a uniform vMSC, though, can result in significant contention between the arrays. Recall that SRDF/Metro must always maintain consistency between the storage and if the same data set is being accessed/updated on each array in quick succession because every IO switches paths, this can cause unwanted latency. The greater the distance between the arrays, the greater the latency impact. For this reason, Dell recommends using VMware's default for iops of 1000 which has been shown in testing to reduce this latency. Note that no SATP rule is required as iops=1000 is what VMware sets out-of-the-box; however, if a rule already exists for iops=1 it should be removed and each device reset to 1000.

If the arrays are co-located in the same datacenter, if not the same physical location, it may be unnecessary to set iops as high as 1000 per switch. Some customers may choose to use iops=1 as a starting point and adjust higher if latency is experienced. Testing is always recommended by Dell.

3.3.2.2 Nonuniform

Because nonuniform vMSC configurations mean local hosts only see their local arrays, the default NMP configuration of iops=1 should be used.

3.3.3 NMP for vSphere 6.7 U1+ and Latency Round Robin

Beginning with vSphere 6.7 U1 VMware offers a new type of Round Robin NMP called "latency". The capability enables VMware to test the performance of the paths to a device and route IO appropriately. The feature is known as **Latency Round Robin.**

There are currently only two "types" of pathing available with NMP that are in use with PowerMax – Fixed and Round Robin. Dell no longer supports the third type, Most Recently Used (MRU). Since Fixed is path restrictive, Round Robin is the default policy (VMW_PSP_RR) for PowerMax devices presented to an ESXi host as seen in Figure 23. Note Dell still uses "EMC Symmetrix" as the description and part of the name (SYMM) despite no longer calling the arrays that name. The reason for this is it provides consistency across ESXi releases to leave it as that name. It is simply a moniker, however, and has no bearing on functionality.

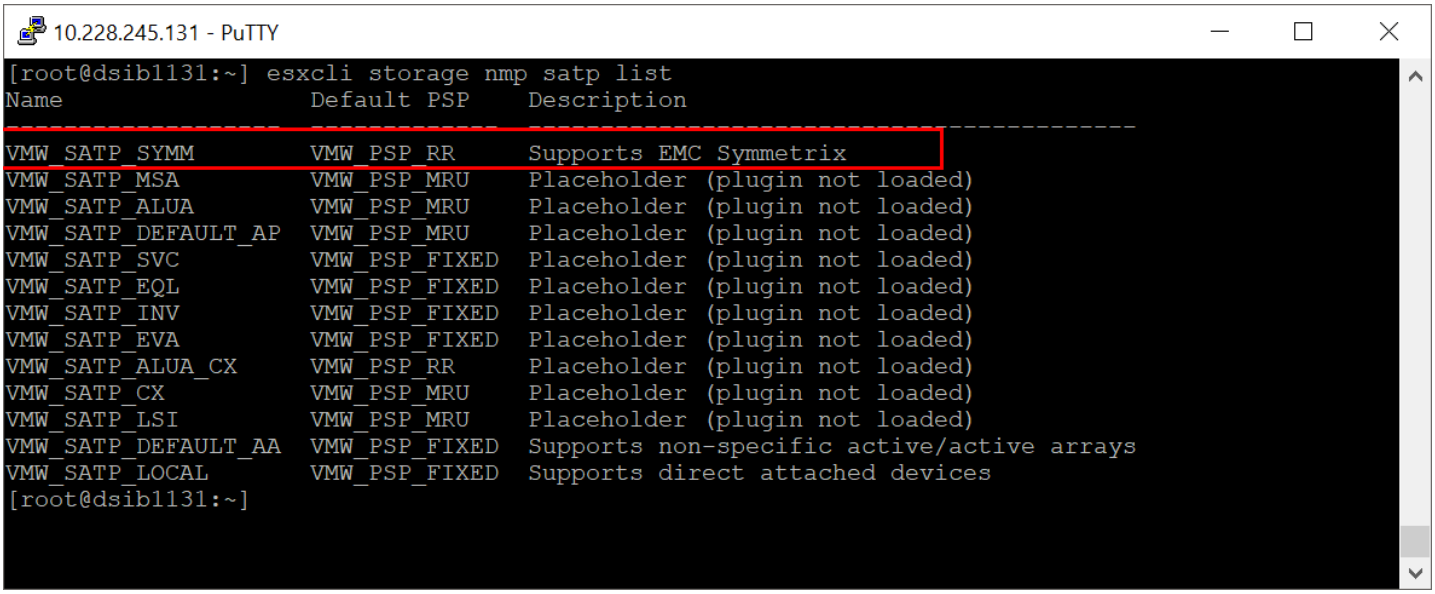

Figure 23. Default SATP for VMAX or PowerMax devices

As mentioned, Dell recommends using the default type=iops, with iops=1 for non-unform vMSC environments and iops=1000 for most uniform vMSC environments.

One thing that has always been lacking in VMware's NMP implementation is the ability to consider the response time/latency of the path in use. This is exactly what the latency setting for Round Robin is designed to do. When type=latency is set on a device, both the latency and pending IOs of a path are used to determine whether an IO should be sent down a particular path.

By default, the latency option of Round Robin is enabled in vSphere 6.7 U1 and higher whether a new install or upgrade. The parameter that controls latency is known as EnablePSPLatencyPolicy. It is available both in the CLI:

```
[root@dsib0142:~] esxcfg-advcfg -g /Misc/EnablePSPLatencyPolicy
Value of EnablePSPLatencyPolicy is 0
```

and in the GUI in Figure 24.

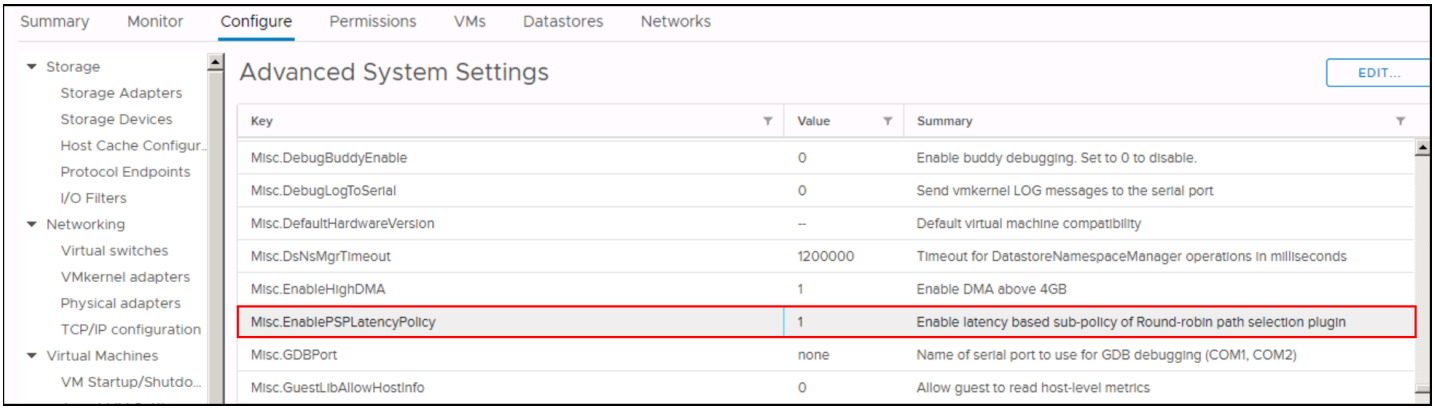

Figure 24. Setting NMP latency in the vSphere Client

If enabled (default), the user can dynamically apply the new type to a device(s):

```
esxcli storage nmp psp roundrobin deviceconfig set -d <Device_ID> --
type=latency
```

Here in Figure 25 is an example of setting latency for a device followed by the command to list the configuration:

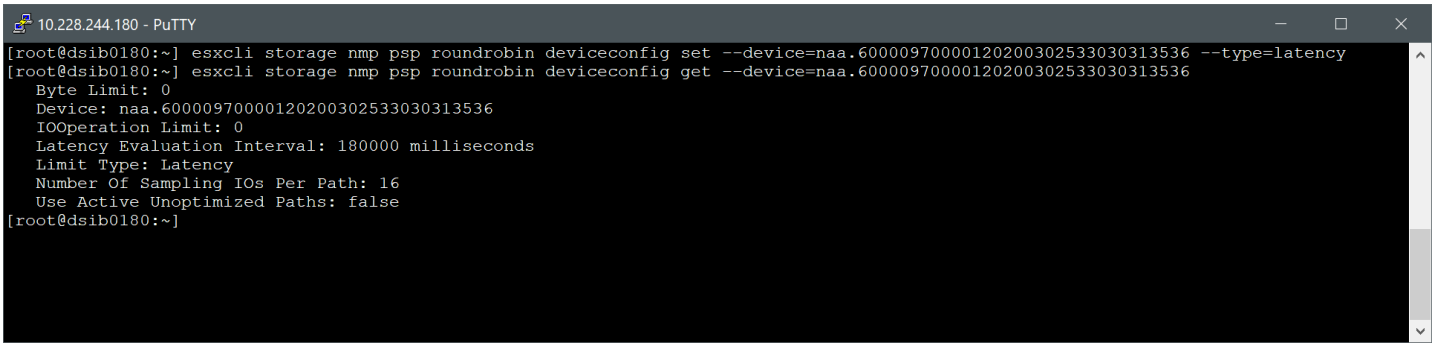

Figure 25. Setting type=latency on a PowerMax device

It is also possible to create an SATP rule that will ensure future devices are assigned the type of latency. The command is below:

```
esxcli storage nmp satp rule add -s "VMW_SATP_SYMM" -V "EMC" -M "SYMMETRIX" -
P "VMW_PSP_RR" -O "policy=latency"
```

Be sure that no other user rules exist, or VMware will return an error. For instance, if the iops=1 rule is already present it will need to be removed first by running:

```
esxcli storage nmp satp rule remove -s "VMW_SATP_SYMM" -V "EMC" -M
"SYMMETRIX" -P "VMW_PSP_RR" -O "iops=1"
```

In deciding which path to use for a particular IO, VMware monitors all the active paths and calculates the average latency based on either time or number of IOs. Once latency is set on a device, the first 16 IOs per active path are used to calculate the latency. Subsequent IOs will be directed to the path with the lowest latency. After the initial sampling window, VMware will re-test every 3 minutes by default. The sampling time and amount are present in a device interrogation in the parameters **Latency Evaluation Interval** and **Number of Sampling IOs Per Path** seen in Figure 25.

The time between re-assessing the paths and the number of IOs to use for the sample can be changed, though VMware recommends against it as testing has shown those values to be effective.

3.3.3.1 Uniform

Uniform vSphere Metro Storage Clusters using SRDF/Metro are a perfect use case for employing the latency type in vSphere 6.7 U1, vSphere 7 or vSphere 8. Rather than blindly switching paths every 1000 IOs, VMware will instead test the paths, and if the arrays are any distance apart, VMware will send most of the IO down the local paths to the local array. In practice, it turns a uniform configuration into a mostly nonuniform configuration, while still maintaining the remote paths in the case of failure.[4] It is most akin, therefore, to PP/VE with Autostandby. In environments with significant distance between datacenters, this mechanism can provide significantly better results than the current recommendation of switching paths every 1000 IOs for a uniform vMSC.

[4] When the arrays are co-located, VMware will tend to treat all paths equally.

3.3.3.2 Nonuniform

The use of the latency type in nonuniform settings has not shown to be more effective than iops=1 and therefore currently Dell does not recommend using it in these configurations; however, in environments that nonetheless experience significant congestion on the single array it can be beneficial.

These recommendations are being made after extensive testing. In the next section some of that is detailed for completeness.

3.3.3.3 Testing Results

As part of the pathing best practices, testing was undertaken to demonstrate the differences between iops=1000 and the new latency type in a uniform vMSC. The uniform setup utilized a distance simulator between the arrays to mimic a campus cluster and load was generated using the IOMETER software.

Figure 26 is a graph of one particular test of the uniform vMSC using IOMETER. It details the following relative results for iops=1000 (blue) and type=latency (orange):

- Average write response
- Average read response
- Average response

The graph also includes the total IOPS in the legend.

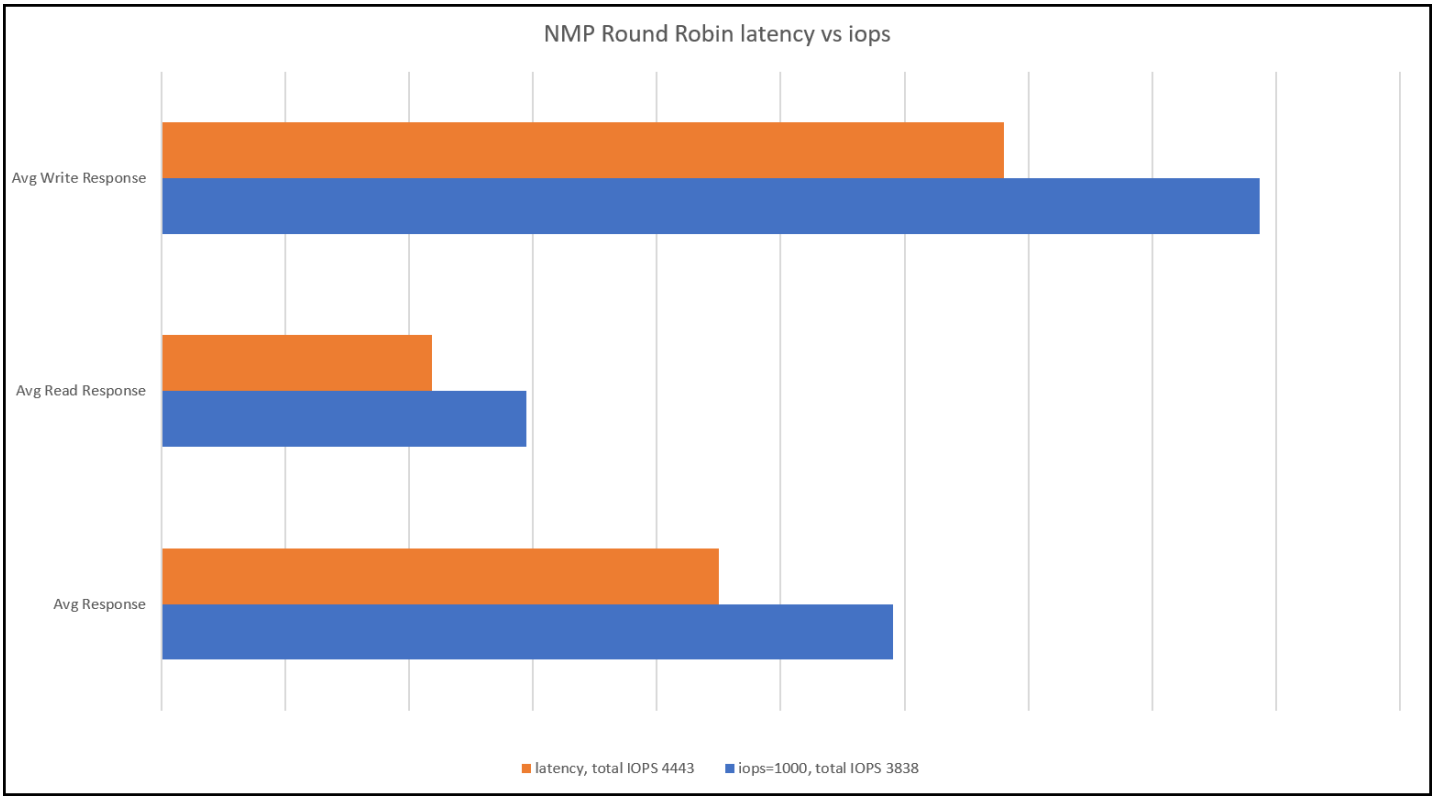

Figure 26. Round Robin type comparison

The results clearly show that the latency type has both superior response times as well as overall IOPS. Detailed analysis determined that while iops=1000 used all paths, regardless of distance,

latency used the local paths the majority of the time, producing the better performance. These results confirm Dell's recommendation of type=latency when running a uniform vMSC with vSphere 6.7 U1, vSphere 7 or vSphere 8.

3.3.4 Polling time for datastore paths

By default, VMware polls for new datastore paths every 300 seconds. In an SRDF/Metro environment, Dell recommends changing this to 30 seconds to avoid the need to manually rescan after presenting the R2 devices. The value that requires changing is ***Disk.PathEvalTime*** and it must be changed on each ESXi host as demonstrated in Figure 27.

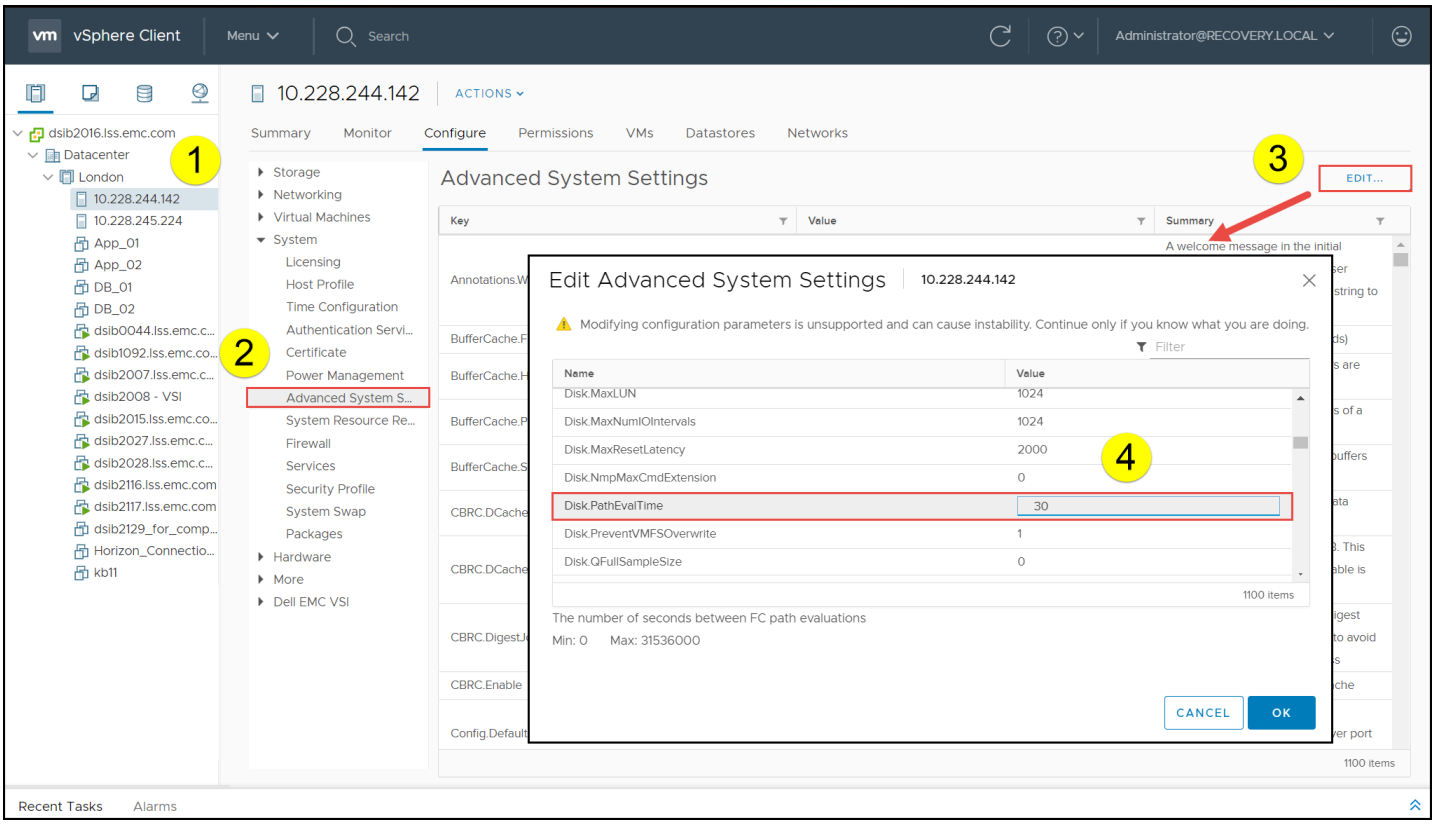

Figure 27. Changing path evaluation time

3.3.5 ALUA and Mobility ID

A brief mention of ALUA, or Asymmetrical Logical Unit Access, here is warranted as many array vendors use this policy. While Dell supports ALUA with SRDF/Metro and the mobility ID, it is unable to determine which device is local to the array and therefore does not add any benefit in a vMSC configuration. It is therefore not a recommended pathing option; however, it is the only pathing option available for use with the Mobility ID and SRDF/Metro. Dell does not currently support NMP RR with Mobility ID and SRDF/Metro.

4 vSphere Cluster Configuration with SRDF/Metro

This section will specifically focus on the best practices when running vMSC with SRDF/Metro.

4.1 vSphere DRS

vSphere Distributed Resource Scheduler or DRS is a VMware utility that balances load across all ESXi hosts in a cluster. Therefore, to properly load balance across the stretched SRDF/Metro cluster, Dell recommends enabling vSphere DRS. **vSphere DRS** is located under the **Cluster -> Configure -> Services** menu in the vSphere Client. Select **vSphere DRS**, then **EDIT**, and move the switch to the right to enable, leaving the defaults. The five steps are outlined in Figure 28.

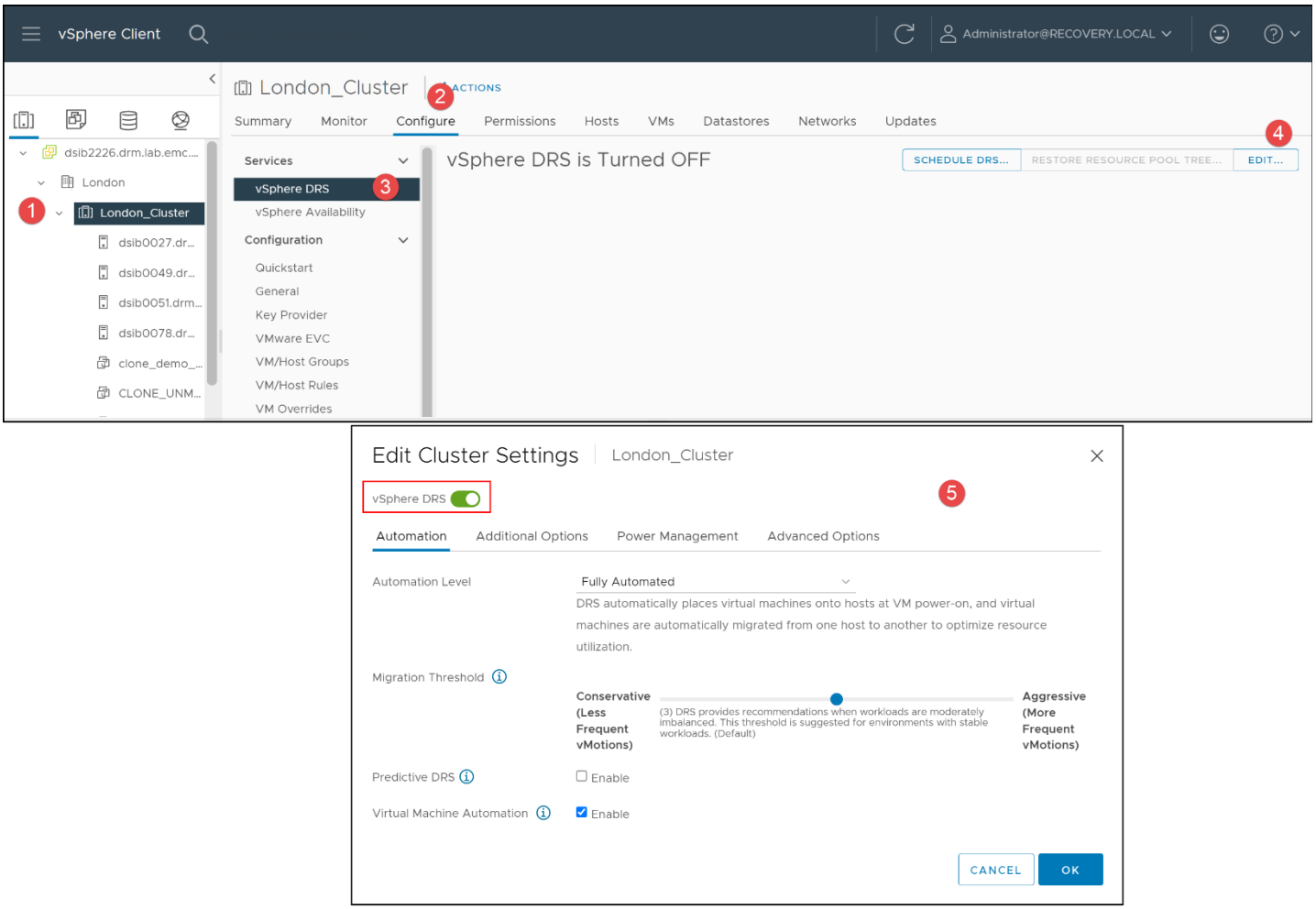

Figure 28. Enabling vSphere DRS in vSphere 7 U3

vSphere DRS uses host CPU and memory resources to determine how best to load balance across the hosts of a cluster. Depending on the automation level, in order to maintain the CPU/memory equity, VMware will move VMs between ESXi hosts using vMotion. For most environments, however, the default is most appropriate where VMware will make recommendations rather than actually issue the vMotion.

Regardless of load distribution, some customer environments require that particular VMs run at one site (ESXi hosts in one datacenter) except in the event of failure. DRS must comply with these rules. The next section will discuss how this is accomplished.

4.2 Site Preference

Many vMSC stretched clusters are backed by a storage system that employs an active/passive solution, wherein only one device of a pair is read/write at any time. This means the datastore residing on that device can only be accessed from the hosts at one datacenter at any one time. In such configurations, therefore, it is essential that the VMs on that datastore “prefer” to run at the read/write site unless there is a failure. This might be referred to as site affinity. VMware offers the opportunity to create rules that govern this site affinity. Fortunately, the Dell solution of SRDF/Metro is an active/active solution permitting read/write on both devices. This means that a VM can run on any host in the cluster, at any datacenter and does not require these site affinity rules for that purpose; however, that does not mean the rules cannot be useful in an SRDF/Metro vMSC. For instance, take a common multi-tier application like Oracle ERP which requires two application (e.g., Oracle Apps) and two database (e.g., Oracle RAC) VMs. In this environment to provide the best availability, ensuring there is an application and database VM running on each datacenter is preferable. If DRS is only making decisions based on CPU and memory, it is very likely the tiers will not be evenly separated, particularly in an environment with many other VMs. Fortunately, through VM/Group and VM/Host rules, users can pre-empt DRS by telling VMware that certain VMs should be associated with certain hosts if possible. DRS will work to enforce these rules when necessary.

4.2.1 VMware VM/Host Groups and VM/Host Rules

To ensure that VMs run on the desired hosts and datacenter while the cluster is healthy, it is necessary to create VM/Host Groups and VM/Host Rules. Setting these up is a fairly simple procedure in the vSphere Client. These rules will set forth which VMs “belong” to which host, and thus which datacenter.

To access the wizards to create Host Groups and Rules, navigate to **Cluster -> Configure -> Configuration**, expanding the group. Here in Figure 29 the two categories are shown: **VM/Host Groups** and **VM/Host Rules**. Note that the DRS automation level must be set to **Fully automated** (default) to permit VMware to move the virtual machines as necessary as previously seen in Figure 28.

Note: The screenshots herein are from an early version of vSphere 7. Newer releases have subtle differences in the icons.

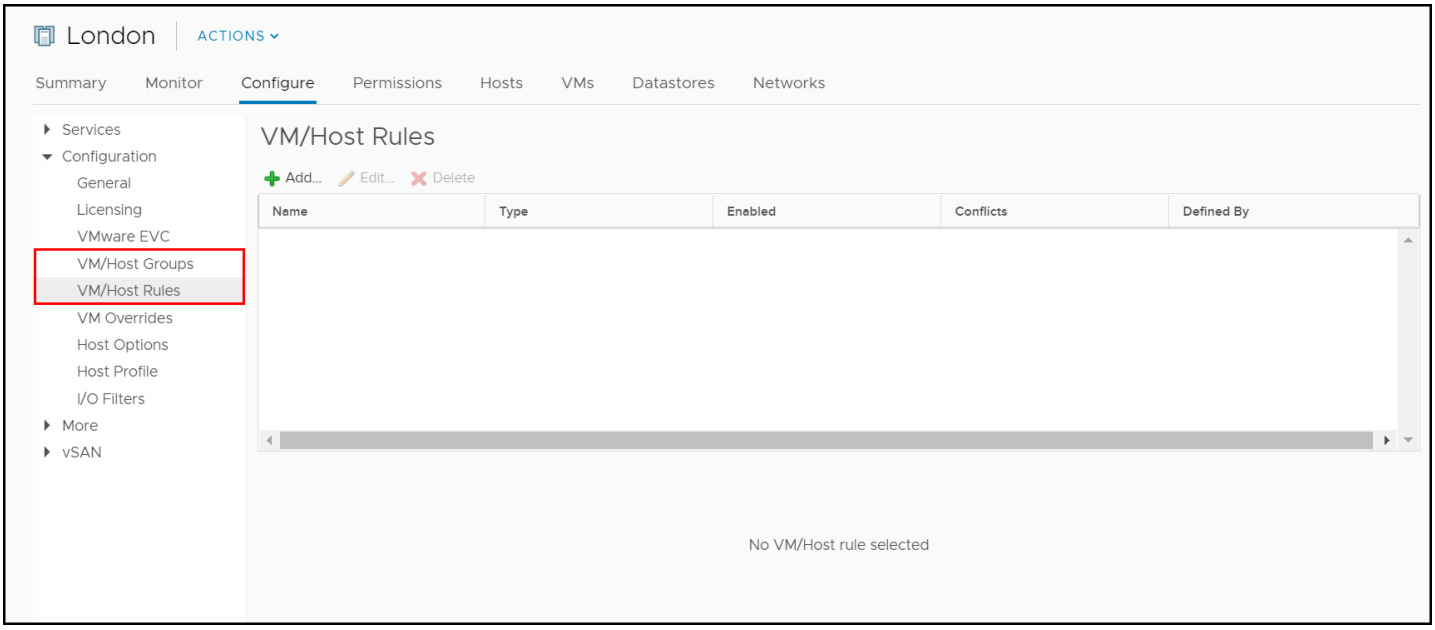

Figure 29. Locating DRS Groups and Rules

For VM/Host Groups, create two VM Groups and two Host Groups, representing the two datacenters. In this example, cluster London consists of an East side and West side datacenter. The East side is represented by **East_Side_Host_Group**, containing one ESXi host[5], and the West side is represented by **West_Side_Host_Group** containing the other ESXi host in the cluster. Similarly, each datacenter has a VM Group containing the VMs that should be associated with their respective datacenters. Figure 30 shows the detail of the four groups while Figure 31 is the completed summary.

[5] A production environment would contain many more ESXi hosts than this lab example. All ESXi hosts present in each respective datacenter should be added to the Host Group. Upon power on, vSphere DRS will place the VM on one of the ESXi servers automatically.

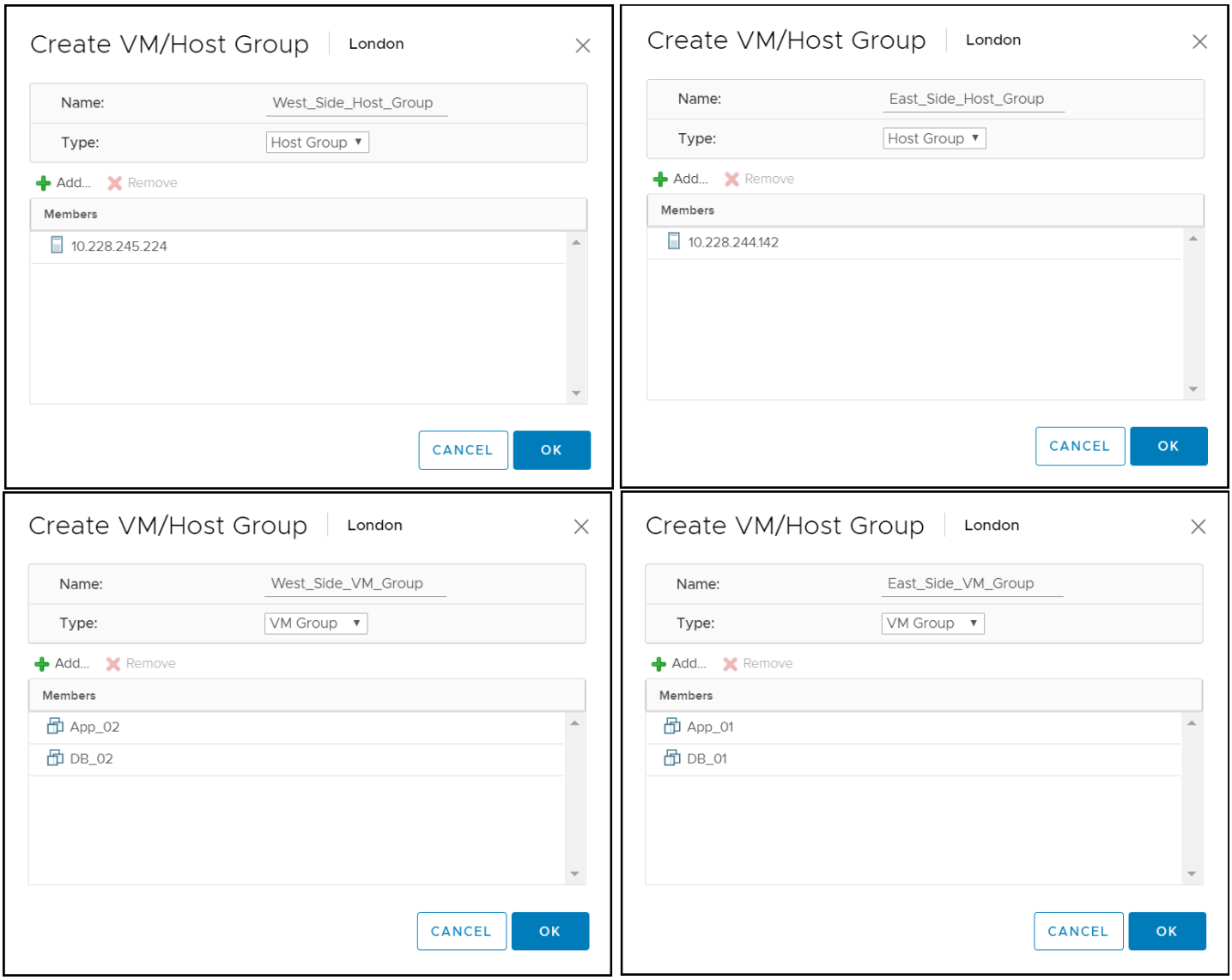

Figure 30. Creation of groups with their hosts and virtual machines

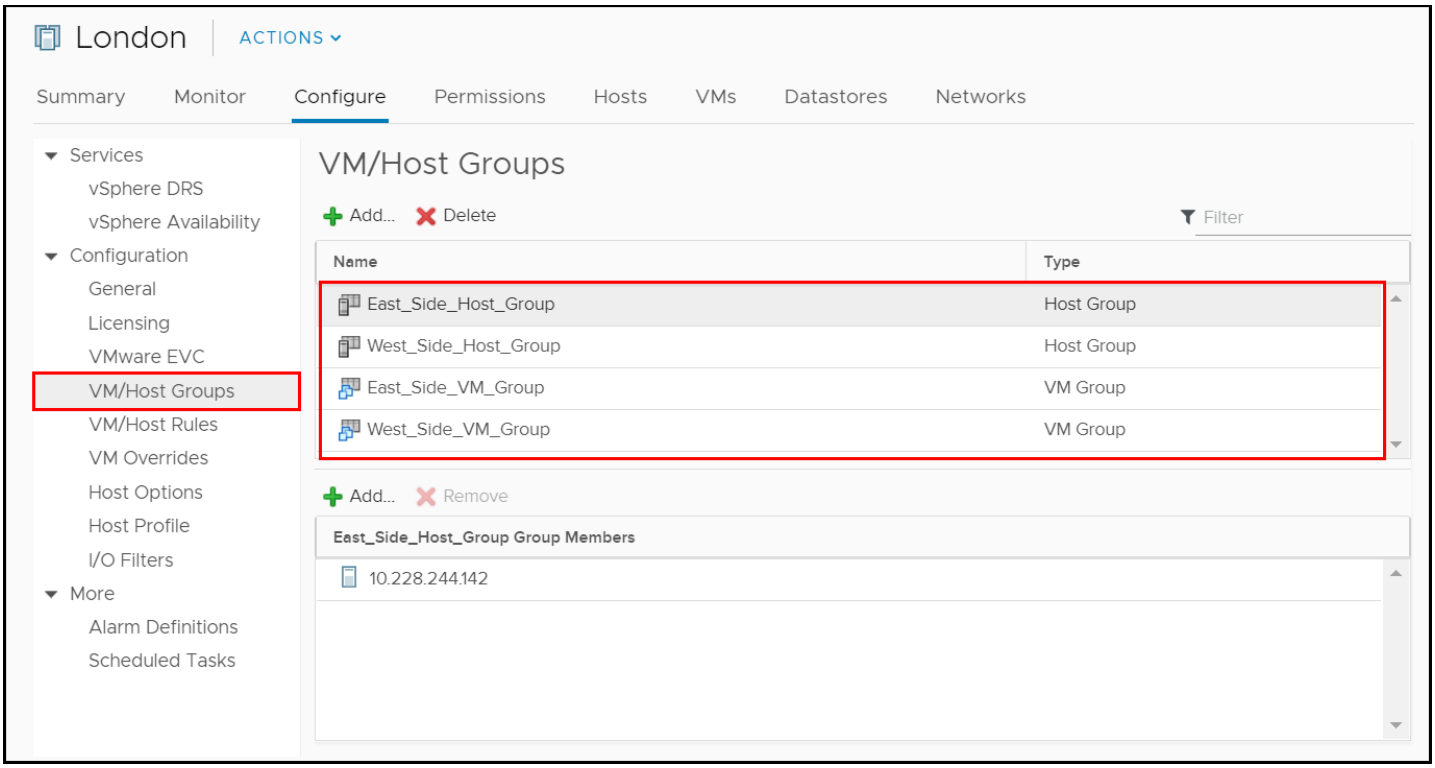

Figure 31. Summary of groups

Now that the groups are in place, rules need to be created to govern the site affinity (or host affinity) of the VMs. There are two rules, one that applies to the East side and one that applies to the West side. When setting up the rules, there are the following options in Figure 32 for how the VM Group should work with Host Group:

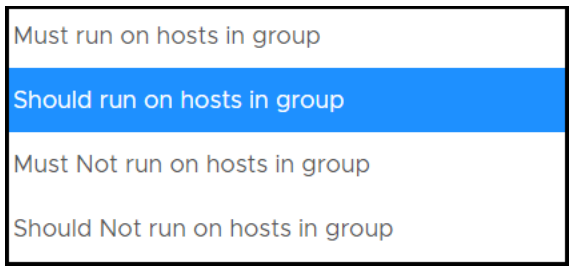

Figure 32. Rule options

It is essential that **Should run on hosts in group** is selected when configuring the rules and not **Must run on hosts in group**. There must be flexibility for the VMs to start-up on the other host(s) that are not part of the Host Group if needed. If **Must run on hosts in group** is chosen, this would mean in the event of a datacenter failure, the VMs running on the ESXi hosts in that datacenter could not be restarted in the other datacenter. The way the rules are configured here permit the VMs associated with the failing datacenter to be brought up on the other host(s) that are part of the site that did not fail. In addition, if the failed hosts are returned to the cluster at a later time, DRS will automatically migrate the VMs back to their original hosts due to the rule.

The newly created rules **East_Side_Rule** and **West_Side_Rule**, are seen in Figure 33. Each rule directs that the VMs in the VM Group **Should run on hosts in group** in the Host Group. In practice it means the VMs have affinity to the hosts in one or the other datacenter.

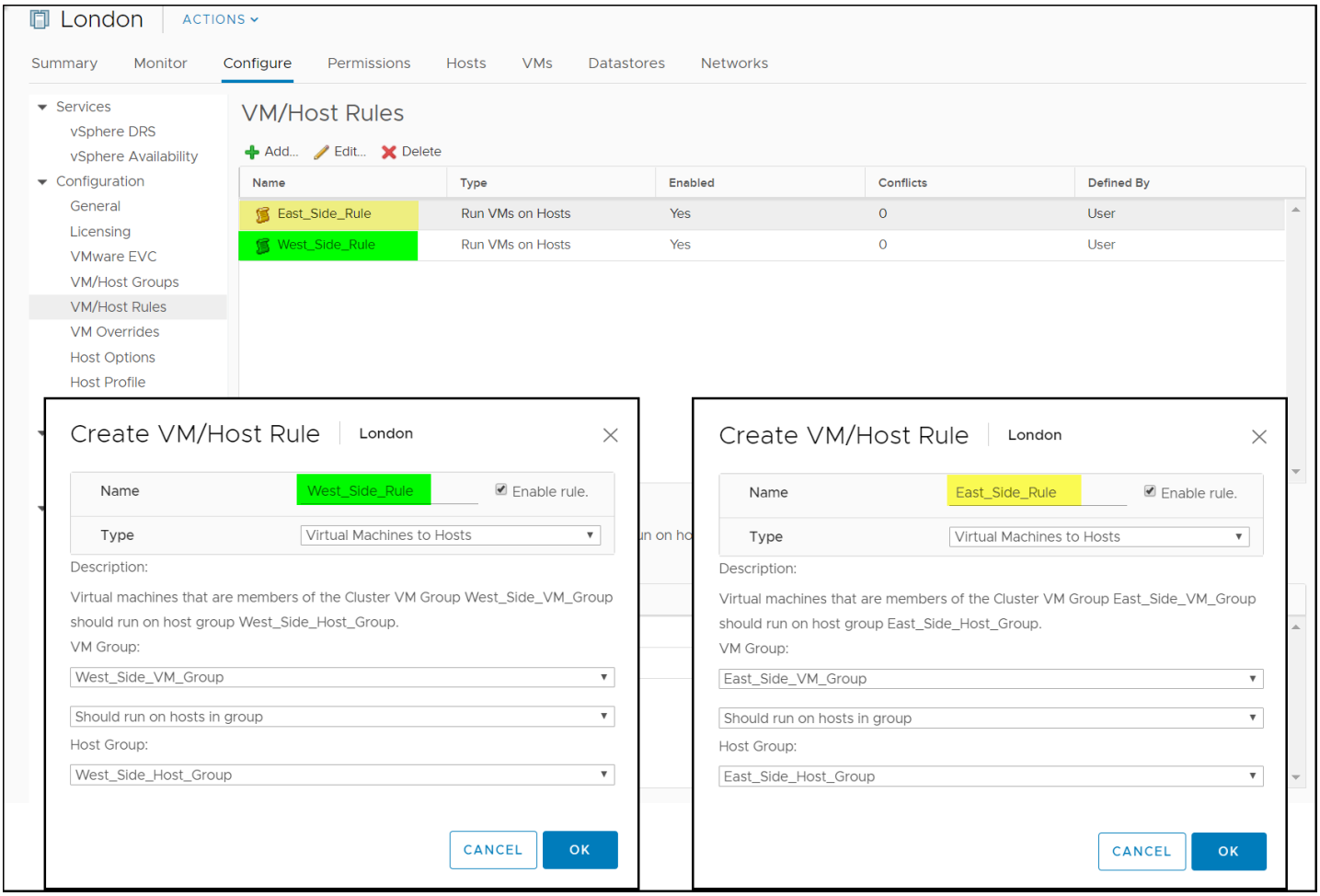

Figure 33. Creation of VM/Host Rules for groups

4.3 vSphere HA

In a vMSC environment utilizing SRDF/Metro, vSphere HA is critical to provide the needed availability by guarding against not just individual host, network, and storage failures, but also complete site failures. SRDF/Metro allows both sides to provide coherent read/write access to the same virtual volume. That means that on the remote site, the paths are up, and the storage is available even before any failover happens. When this is combined with host failover clustering technologies such as VMware HA, it creates a fully automatic application restart for any site-level disaster. The system rides through component failures within a site, including the failure of an entire array.

In this configuration, a virtual machine can write to the same virtual device from either cluster. In other words, if the customer is using vSphere DRS, which allows the automatic load distribution on virtual machines across multiple ESXi servers, a virtual machine can be moved from an ESXi server attached to the R1 array to an ESXi server attached to R2 array without losing access to the underlying storage. This configuration allows virtual machines to move between two geographically disparate locations with up to 150ms of latency, the limit to which VMware vMotion is supported across hosts.

Note: Prior to vSphere 7, VMs with very large memory requirements take much longer to vMotion. In vSphere 7 VMware made significant changes to the vMotion engine and even VMs with large memory footprints transfer efficiently and with far less impact to the running environment.

4.3.1 Site failure example

In the event of a complete site failure, shown in Figure 34, SRDF/Metro Witness automatically assigns the R2 array as the winner, rather than following the R1 site bias. vSphere HA detects the failure of the virtual machines and restarts the virtual machines automatically at the surviving site with no user intervention. It is therefore critical to properly configure vSphere HA to aid in this automated recovery.

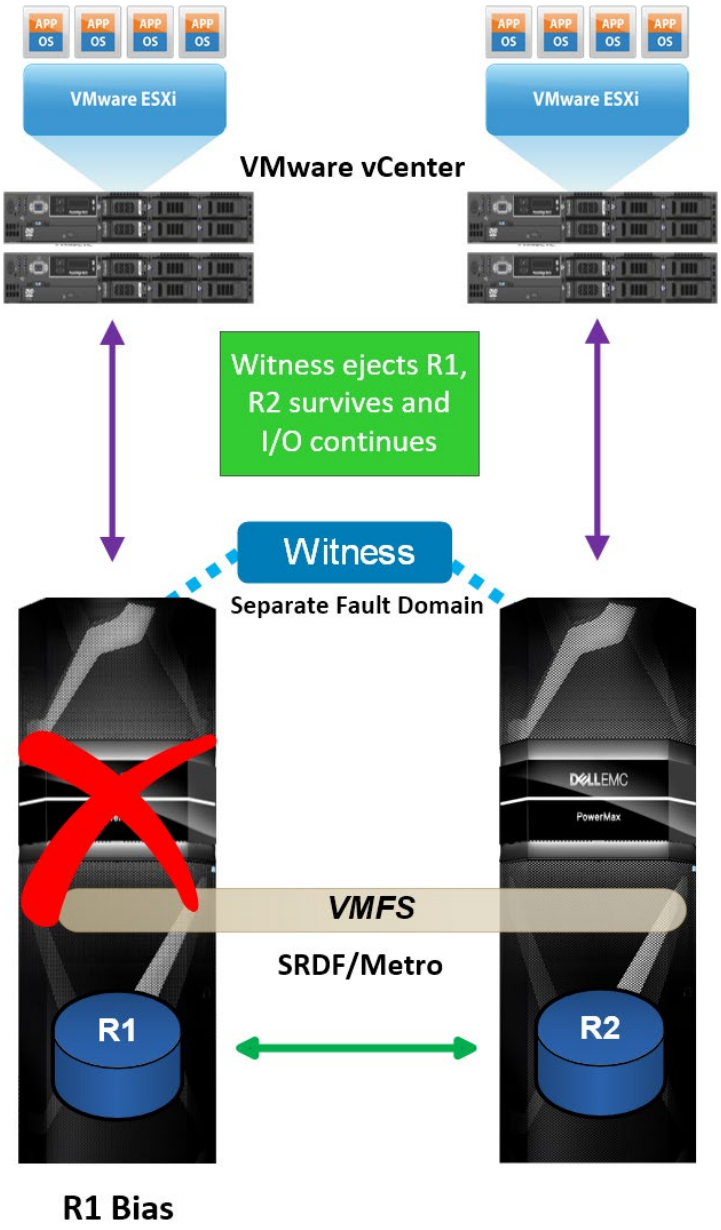

Figure 34. vMSC with SRDF/Metro Witness

Note: As this paper will not detail all failure scenarios of an SRDF/Metro configuration, please see the Technical Note *SRDF/Metro Overview and Best Practices* in the References section if more information is desired.

4.4 Enabling vSphere HA

To enable vSphere HA, navigate to **Cluster -> Configure -> vSphere Availability** and then **EDIT vSphere HA**. To activate, slide the bar to the right and green in the steps outlined in Figure 35.

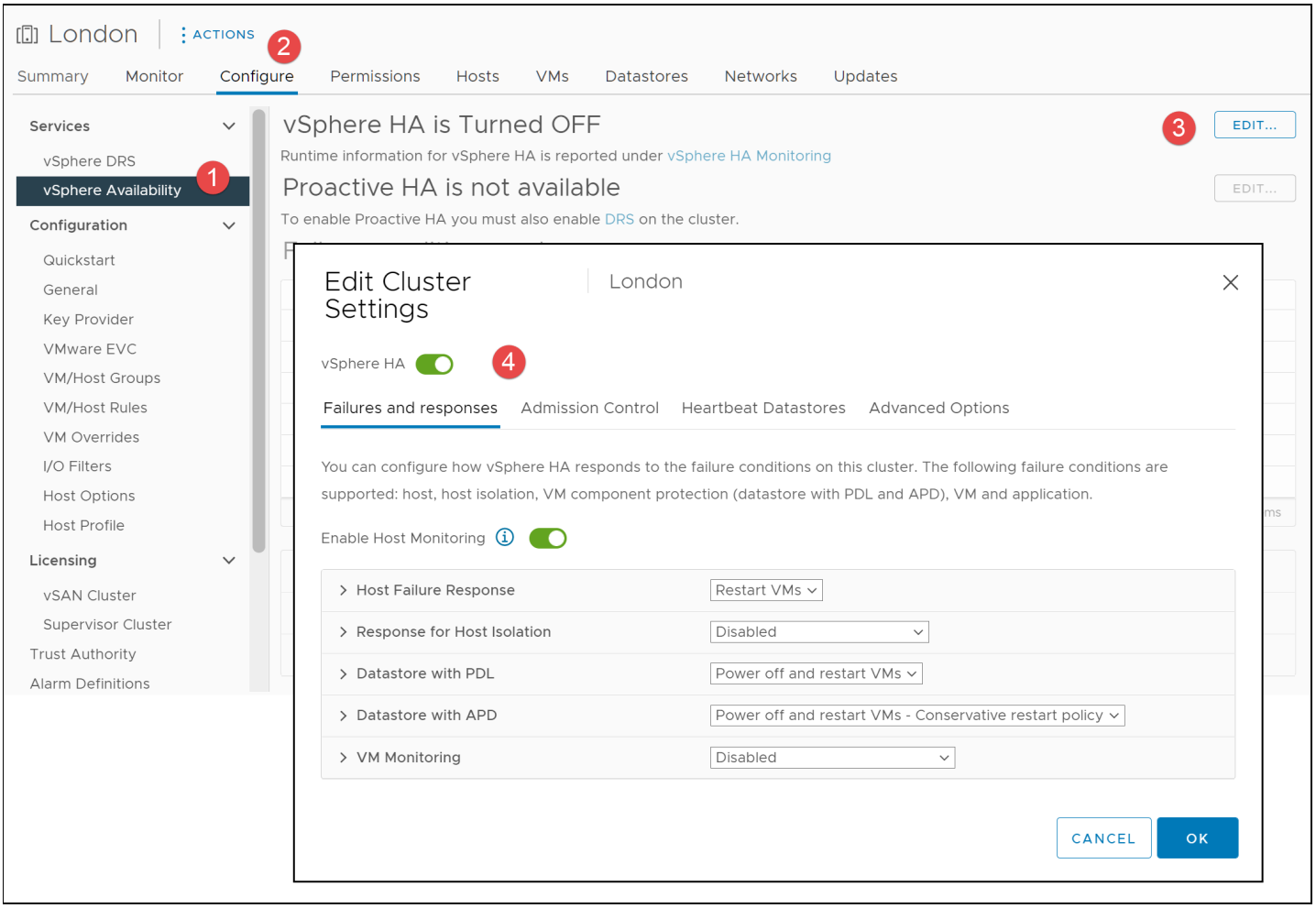

Figure 35. Setting vSphere HA services in vSphere 7 U3

The subsequent sections detail the configuration of vSphere HA. The settings are modified while in the edit screen seen in Figure 35.

4.4.1 Admission Control

As the main business driver of SRDF/Metro and vSphere Metro Storage Cluster is high availability, it is important to ensure that server resources exist to failover to a single site. Both VMware and Dell recommend using a percentage-based policy for Admission Control with HA as it is flexible and does not require changes when additional hosts are added to the cluster. Therefore, the Admission Control policy of vSphere HA should be configured for 50% CPU and 50% memory. This is demonstrated in Figure 36.

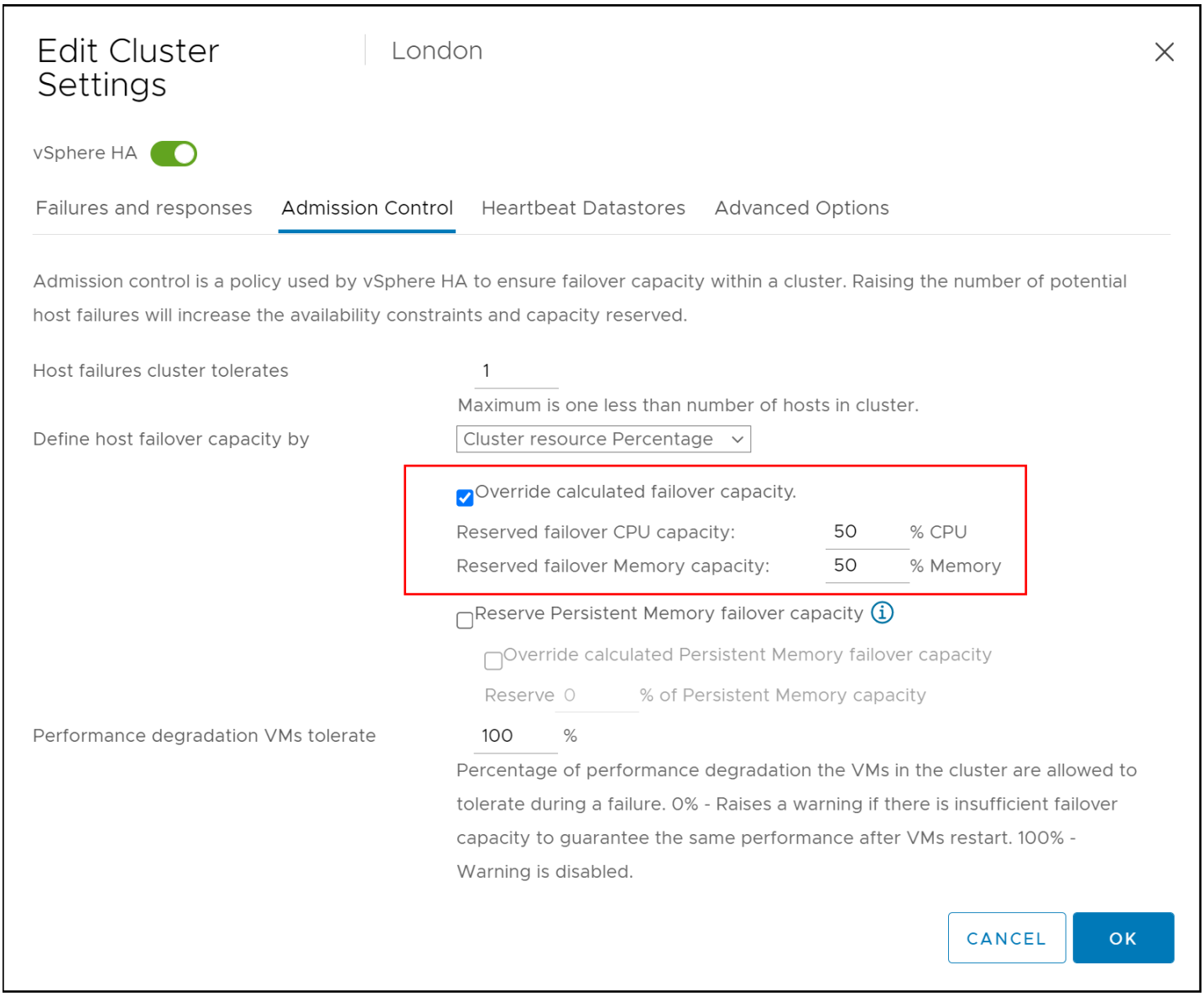

Figure 36. vSphere HA Admission Control

4.4.2 Heartbeating

vSphere HA uses heartbeat mechanisms to validate the state of a host. There are two different types of heartbeating:

- network (primary)
- datastore (secondary)

If vSphere HA fails to determine the state of the host with network heartbeating, it will then use datastore heartbeating. If a host is not receiving any heartbeats, it uses a fail-safe mechanism to detect if it is merely isolated from its master node or completely isolated from the network. It does this by pinging the default gateway. It is possible to configure additional isolation addresses in case the gateway is down. VMware recommends specifying a minimum of two additional isolation addresses. Each address should be local to one datacenter. This is configured in the Advanced

Options of vSphere HA using the option name ***das.isolationAddress.x*** ('x' is incremented for each address starting with zero):

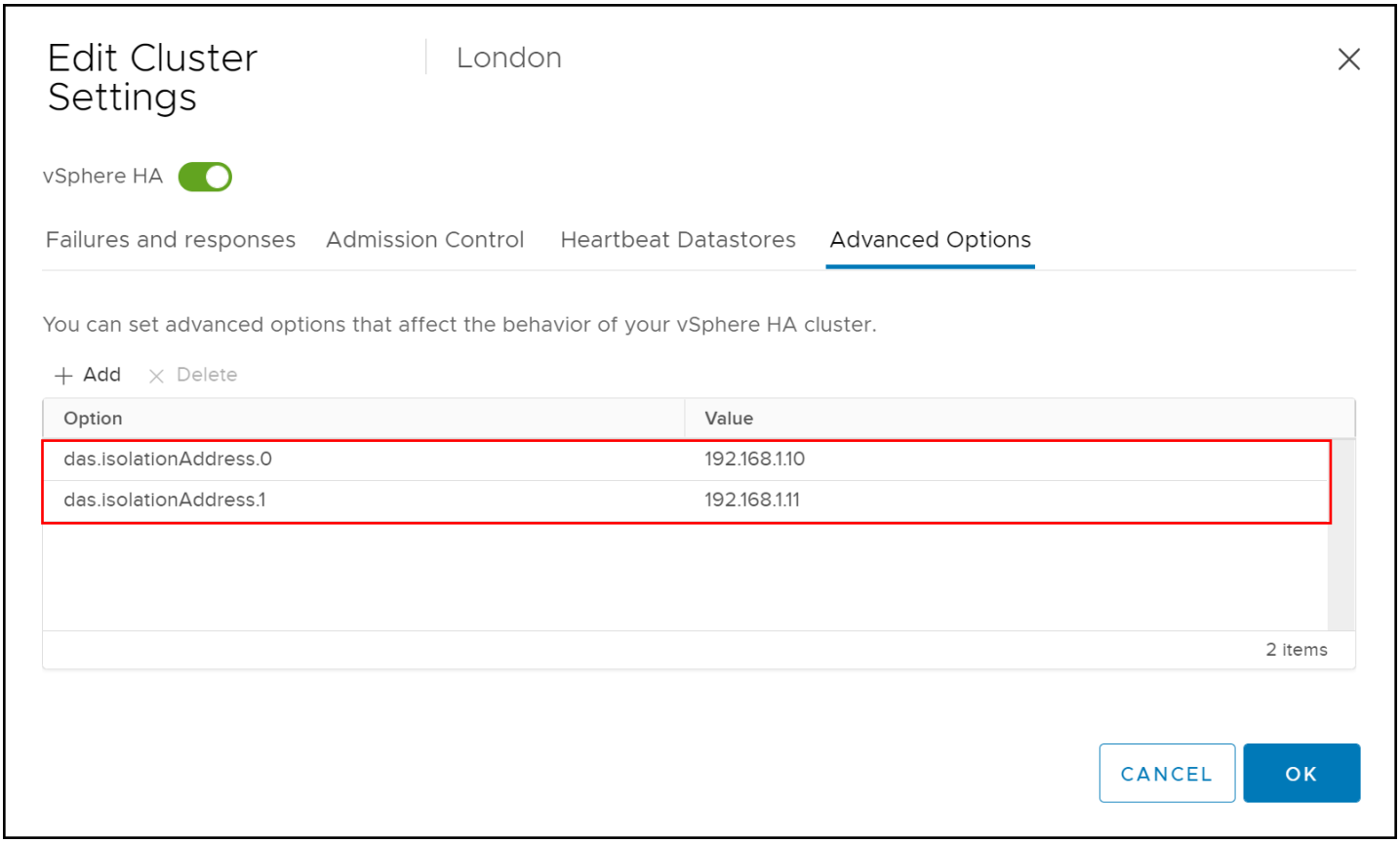

Figure 37. Adding network isolation addresses for vSphere HA

For the heartbeat mechanism, the minimum number of heartbeat datastores is two and the maximum is five. VMware recommends increasing the number of heartbeat datastores from two to four in a stretched cluster environment. This provides full redundancy for both datacenter locations. Defining four specific datastores as preferred heartbeat datastores is also recommended. To increase the minimum heartbeat datastores, in the Advanced Options of vSphere HA add a new option named ***das.heartbeatDsPerHost*** and set the value to 4, depicted in Figure 38.

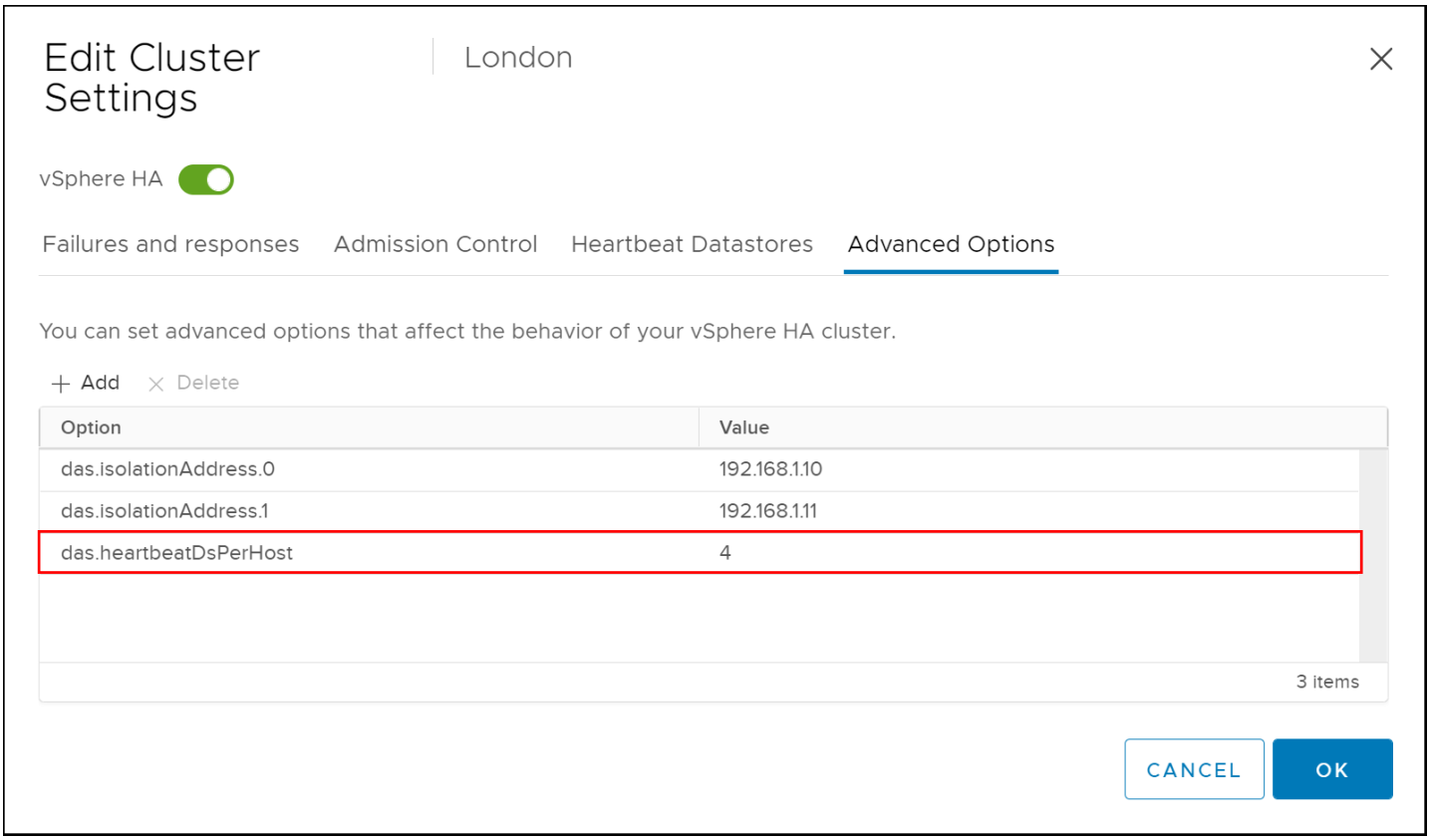

Figure 38. Change minimum heartbeating datastores

Once the minimum heartbeat datastores are configured, under the tab **Heartbeat Datastores** select the radio button **Use datastores from the specified list and complement automatically if needed**. Then select the four specific datastores to use in the configuration. Since SRDF/Metro is an active/active solution, it is perfectly acceptable to select datastores that are visible on all hosts. The recommended setup is shown in Figure 39.

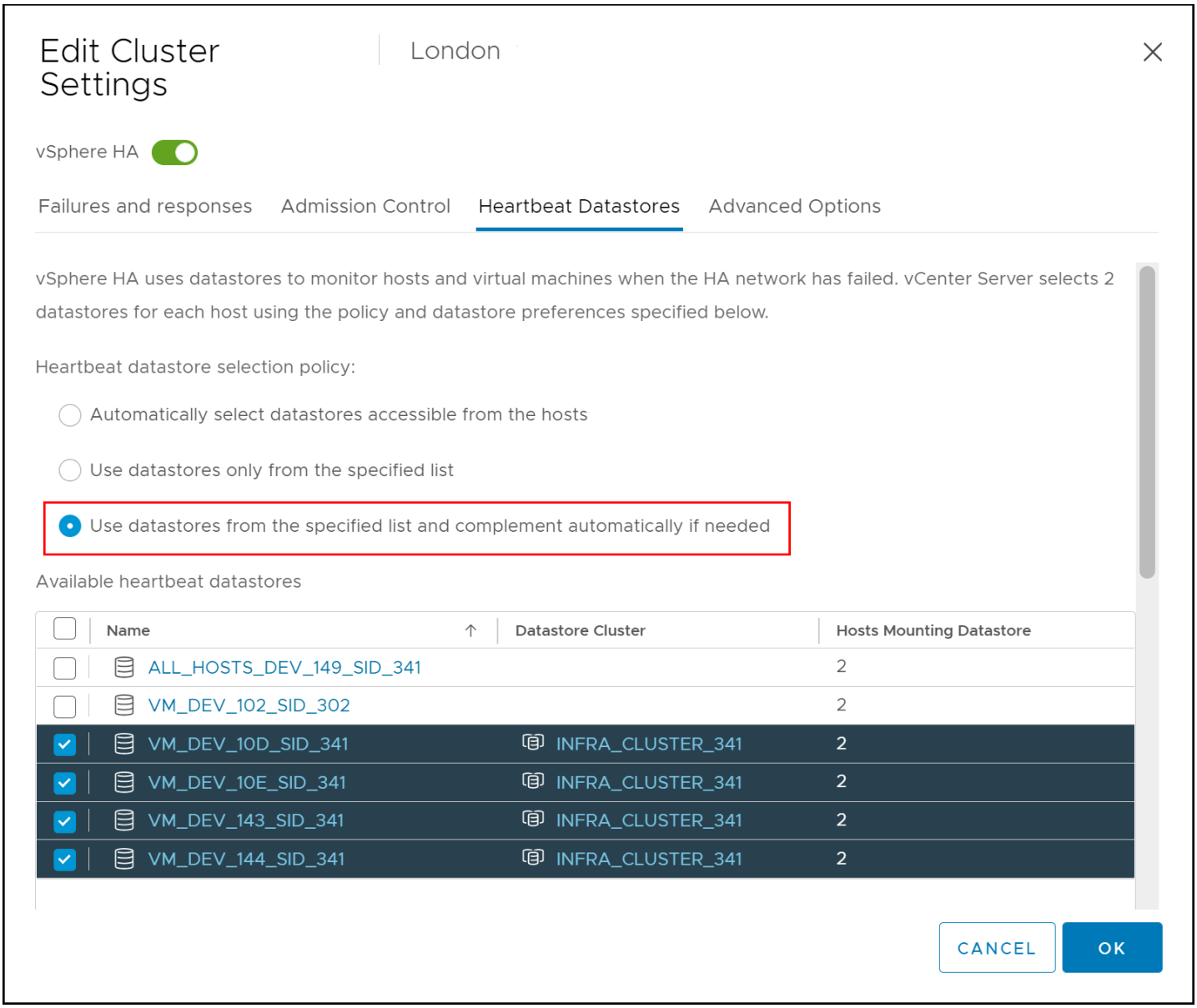

Figure 39. Modify heartbeat datastore selection policy

4.4.3 All Paths Down (APD) and Permanent Data Loss (PDL)

An additional, important feature that should be addressed when enabling vSphere HA is Host Hardware Monitoring or VM Component Protection. This relates to the conditions All Paths Down and Permanent Data Loss.

4.4.3.1 APD

All paths down or APD, occurs on an ESXi host when a storage device is removed in an uncontrolled manner from the host (or the device fails), and the VMkernel core storage stack does not know how long the loss of device access will last. VMware, however, assumes the condition is temporary. A typical way of getting into APD would be if the zoning was removed.

4.4.3.2 PDL

Permanent device loss or PDL, is similar to APD (and hence why initially VMware could not distinguish between the two) except it represents an unrecoverable loss of access to the storage. VMware assumes the storage is never coming back. Removing the device from the storage group that underlies the datastore from would produce the error.

Note: VMware relies on SCSI sense codes to detect PDL. If a device fails in a manner that does not return the proper sense codes, VMware will default to APD behavior.

4.4.3.3 PDL AutoRemove

PDL AutoRemove is a feature that automatically removes a device from a host when it enters a PDL state. Because vSphere hosts have a limit of 255 disk devices per host in vSphere 6.0, a device that is in a PDL state can no longer accept IO but can still occupy one of the available disk device spaces. Therefore, it is better to remove the device from the host. This is less of a concern in vSphere 6.5 and 6.7+ where device limits are 512 and 1024 respectively, however it is an important concept to understand, nonetheless.

PDL AutoRemove occurs only if there are no open handles left on the device. The auto-remove takes place when the last handle on the device closes. If the device recovers, or if it is re-added after having been inadvertently removed, it will be treated as a new device. In such cases VMware does not guarantee consistency for VMs on that datastore.

In a vMSC environment, such as with SRDF/Metro, VMware recommends that AutoRemove be left in the default state, enabled. In an SRDF/Metro environment this is particularly important because if it is disabled, and a suspend action is taken on a SRDF/Metro pair(s), the non-biased side will experience a loss of communication to the devices that can only be resolved by a host reboot.

For more detail on AutoRemove refer to VMware KB 2059622.

4.4.4 VMCP

vSphere offers some capabilities around APD and PDL for the vSphere HA cluster which allow automated recovery of VMs. The capabilities are enabled through a feature in vSphere called VM Component Protection or VMCP. When VMCP is enabled, vSphere can detect datastore accessibility failures, APD or PDL, and then recover affected virtual machines. VMCP allows the user to determine the response that vSphere HA will make, ranging from the creation of event alarms to virtual machine restarts on other hosts.

VMCP is part of vSphere HA and enabled by default in vSphere when HA is enabled. It is recognized by the label **Enable Host Monitoring** which is seen in Figure 40. Note how the switch is already enabled even before vSphere HA is enabled. In versions prior to 6.7, VMCP had to be manually enabled by checking a box labeled **Protect against Storage Connectivity Loss**.

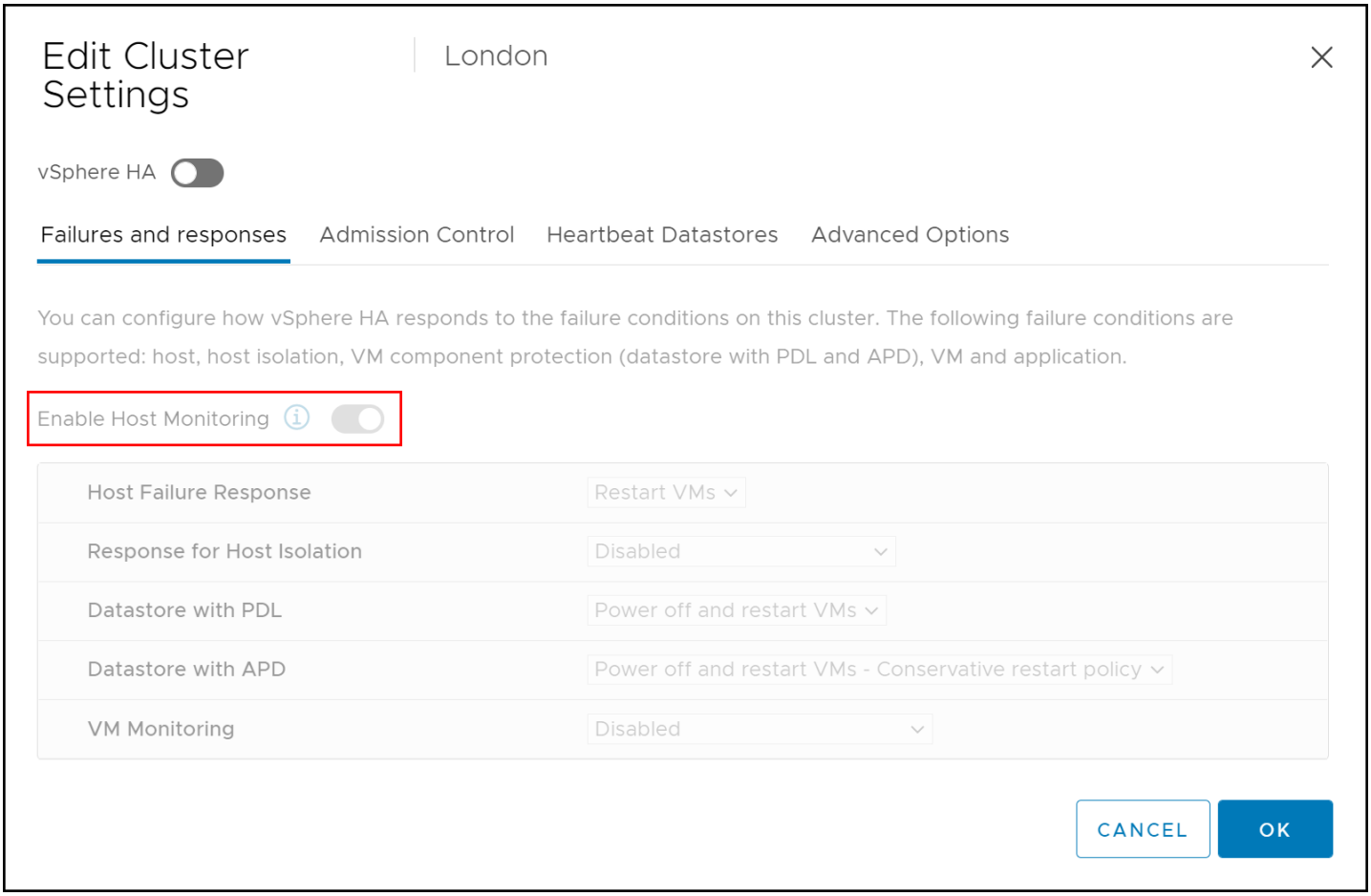

Figure 40. VMCP

Once vSphere HA is enabled, therefore, storage protection levels and virtual machine remediation can be chosen for APD and PDL conditions as shown in Figure 41. By default, VMware sets the most common options for each category. For example, **Host Failure Response** is set to **Restart VMs**, as otherwise HA would leave the VMs in a failed state. Each category will be addressed in the next sections. Note that in some versions of vSphere prior to 7 U3, the defaults are disabled for all but **Host Failure Response** and some of the options have changed.

Note: It is essential VMware Tools is installed on the VMs participating in the vSphere HA cluster for some of the VMCP actions to function.

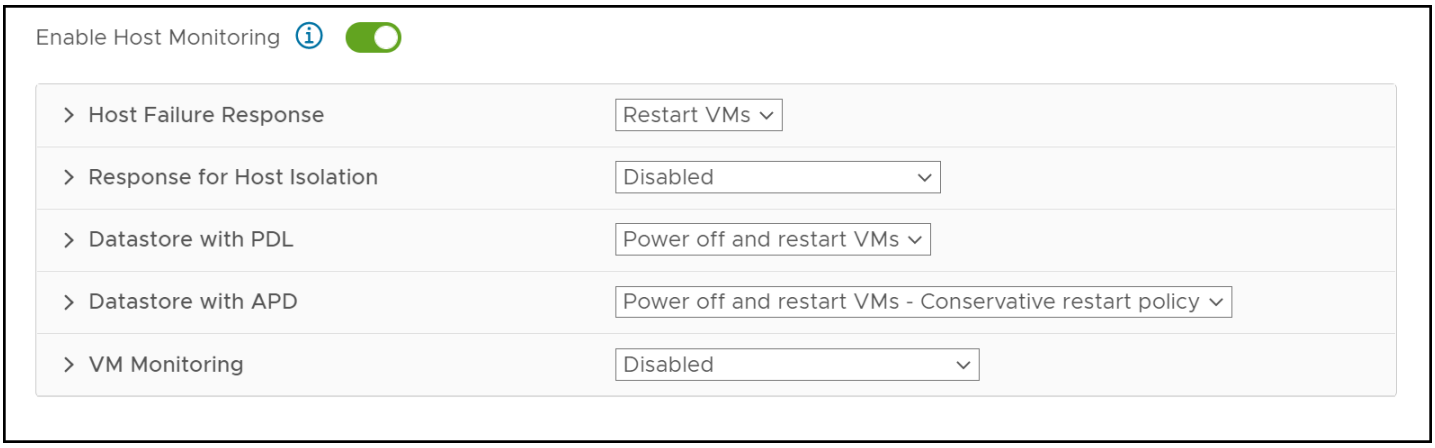

Figure 41. Storage and VM settings for VMCP

4.4.4.1 Response for Host Isolation

Host isolation is when a particular host in the cluster cannot communicate with other hosts in the cluster. At that point the cluster isolates the host, so it is no longer part of the cluster. The options for handling this event are in Figure 42, with **Disabled** being the default.

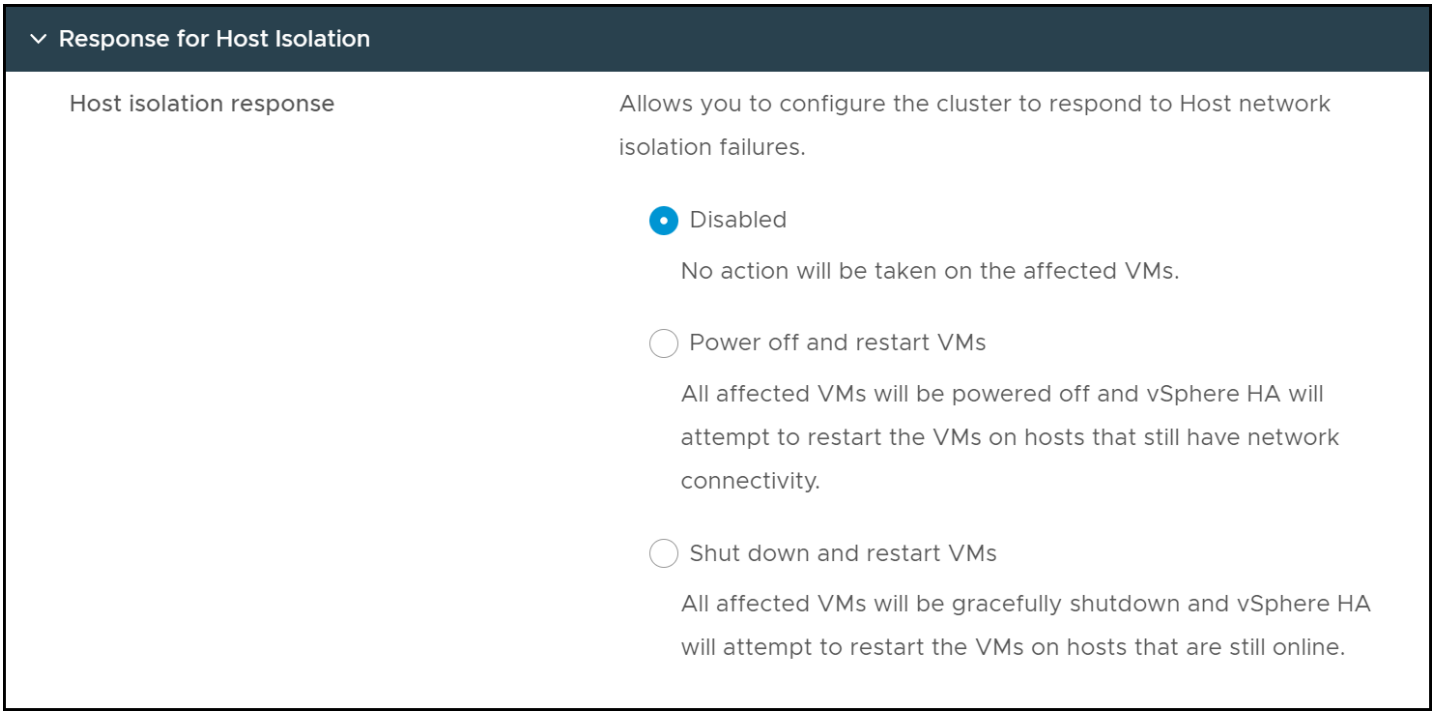

Figure 42. VMCP - Response for Host Isolation

Host isolation may be a transient condition due to an inability for the host to receive a heartbeat, or it may be a more serious condition. It is perfectly acceptable, therefore, to leave the setting as Disabled, i.e., VMware will do nothing, or to set it to shut down and restart the VMs (preferable to forced power off). Previously Dell and VMware recommended leaving this setting as default, but more recent documentation appears suggests setting it to the options **Shut down and restart VMs**, though without detailed explanation. Therefore, Dell makes no definitive recommendation as either option is acceptable.

4.4.4.2 PDL VMCP settings

The PDL settings are the simpler of the two failure conditions between APD and PDL to configure. This is because there are only two choices other than leaving it disabled:

- Issue events
- Power off and restart VMs

As the purpose of HA is to keep the VMs running, the recommendation is to **Power off and restart the VMs**. As this is the default, leave the radio button next to this option set as in Figure 43.

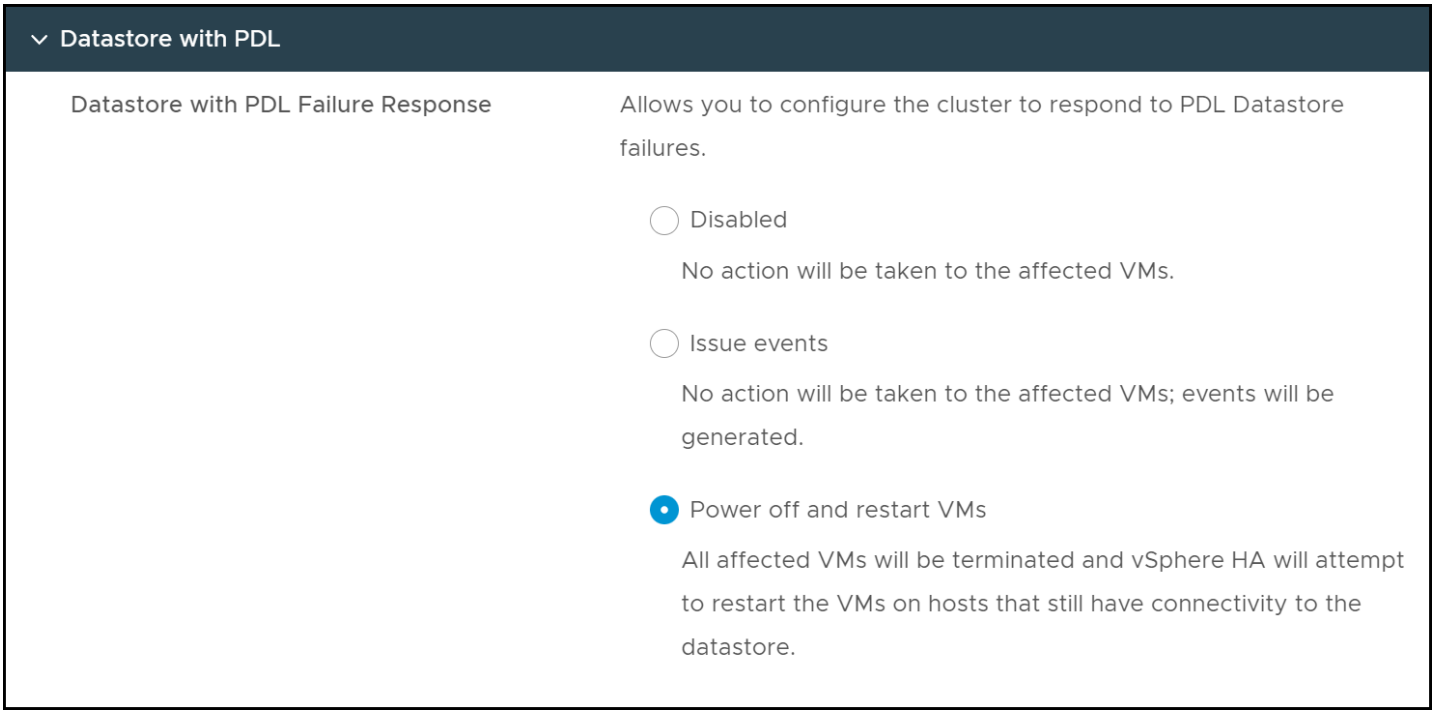

Figure 43. VMCP - PDL recommended setting

4.4.4.3 APD VMCP settings

As APD events are by nature transient, and not a permanent condition like PDL, VMware provides a more nuanced ability to control the behavior within VMCP. Essentially, however, there are still two options to choose from:

- vSphere can issue events
- vSphere can initiate power off of the VMs and restart them on the surviving host(s) (aggressively or conservatively)

In general, when vSphere recognizes an APD condition it starts a 140 second countdown timer. When the timer ends, the device is officially marked as APD timeout. At this point, vSphere HA starts a 180 second countdown timer before it acts, however using VMCP this behavior can be enhanced by telling vSphere to power off and restart the VMs either conservatively or aggressively instead of waiting. The difference between how vSphere behaves in a conservative or aggressive configuration has to do with the other hosts in the cluster. If conservative is selected, VMware will attempt to power off and restart the VM only if it is able to determine there is another host in the cluster with access to the datastore where the VM resides. In the aggressive setting, VMware powers off the VM and then determines if a host is available for restart. This may mean the VM will not be able to restart. Note that if the cluster does not have sufficient resources, neither approach will terminate the VM. Dell recommends a conservative approach which is the default. This is the setting seen in Figure 44.

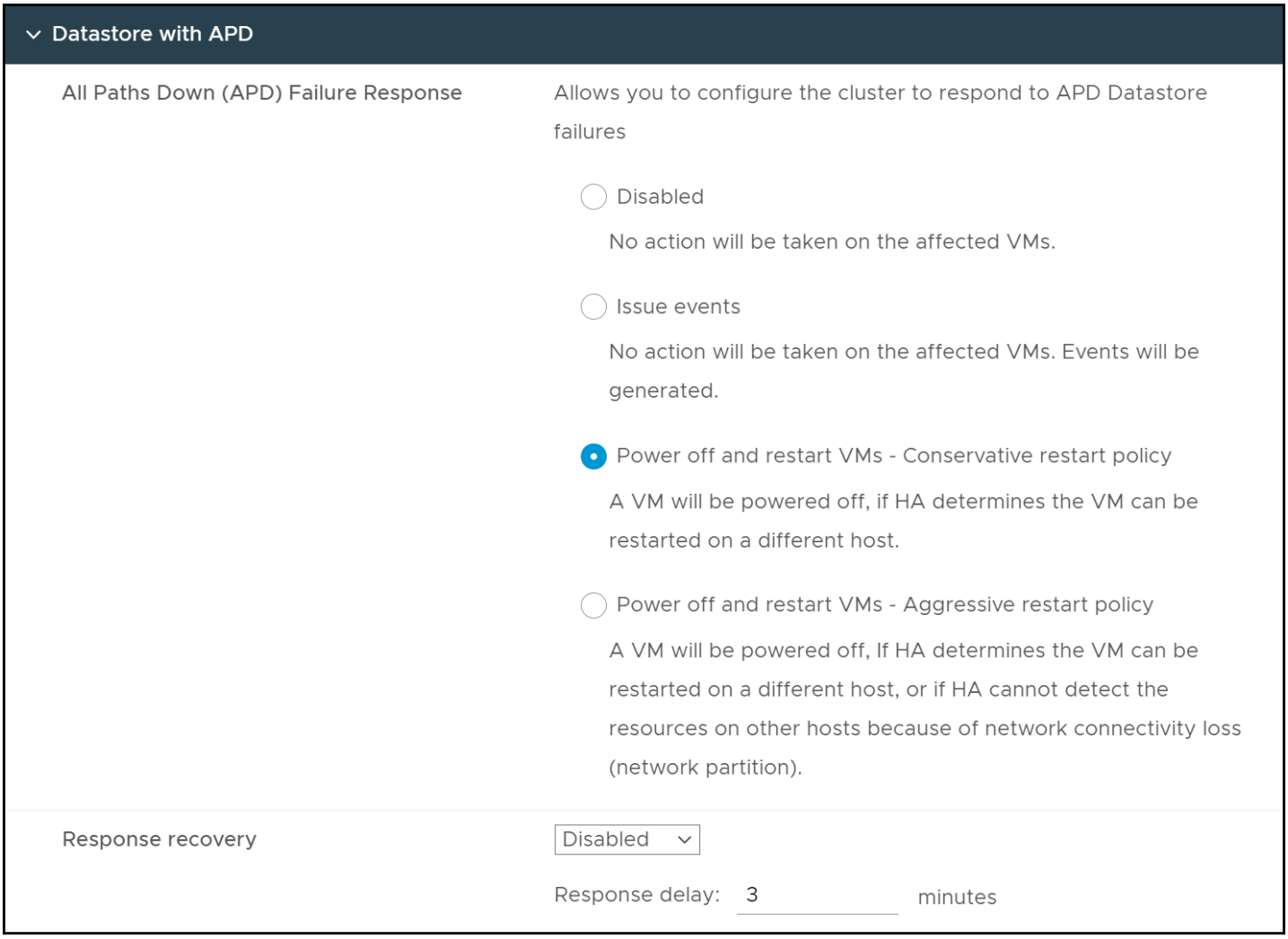

Figure 44. VMCP - APD recommended setting

If **Issue events** is selected, vSphere will do nothing more than notify the user through events when an APD event occurs. As such, no further configuration is necessary. If, however, either aggressive or conservative restart of the VMs is chosen, an additional option may be selected to further define how vSphere is to behave.

4.4.4.4 Response recovery

In Figure 44 the bottom box contains a policy setting called **Response recovery**. In addition to setting the delay for the restart of the VMs, the user can choose whether vSphere should act if the APD condition resolves before the user-configured delay period is reached. If the setting **Response recovery** is set to **Reset VMs**, and APD recovers before the delay is complete, the affected VMs will be reset which will recover the applications that were impacted by the IO failures. This setting does not have any impact if vSphere is only configured to issue events in the case of APD. VMware and Dell recommend leaving this set to **Disabled** (the default setting) so as to not unnecessarily disrupt the VMs.

Note: If either the Host Monitoring or VM Restart Priority settings are disabled, VMCP cannot perform virtual machine restarts. Storage health can still be monitored, and events can be issued, however.

4.4.4.5 VM Monitoring

The purpose of VM Monitoring is to restart VMs if the VMware Tools heartbeat is not responding and there is no storage or network IO over a specified period – 120 seconds in the recommended preset configuration. Dell recommends activating **VM Monitoring Only** as seen in Figure 45, accepting the default values.

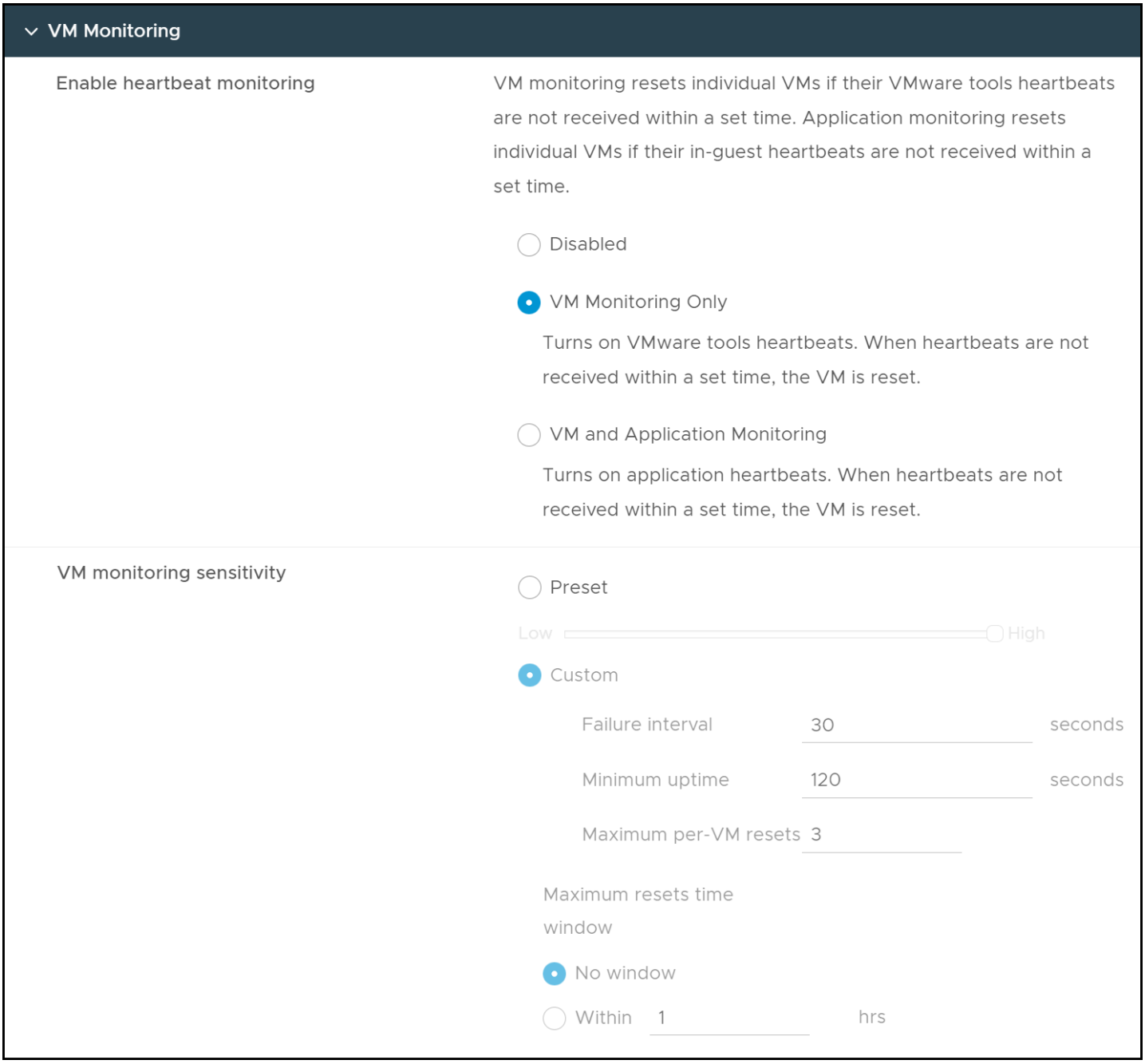

Figure 45. VMCP - VM Monitoring

The completed recommendations for all VMCP are shown in Figure 46.

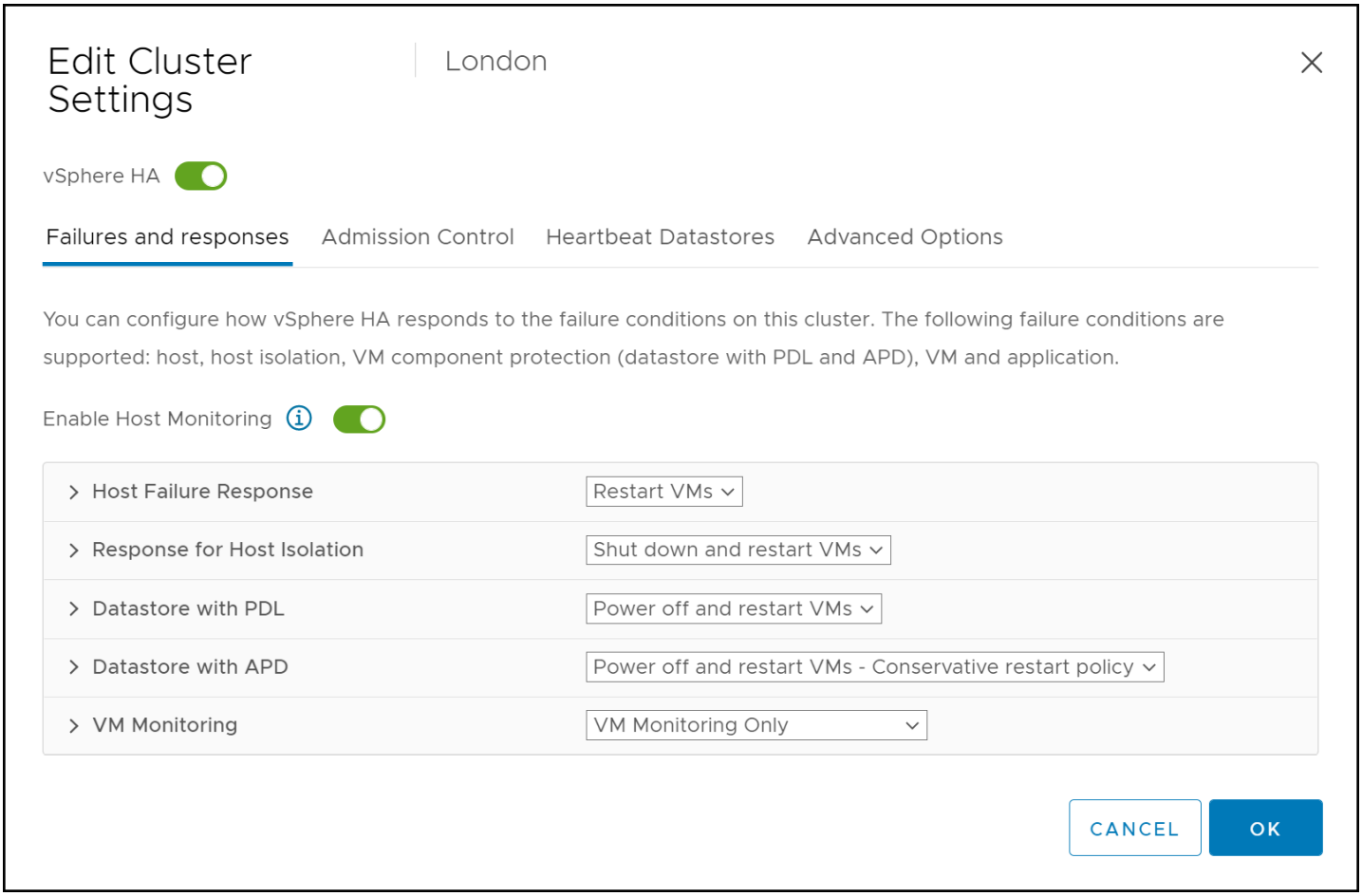

Figure 46. VMCP summary

4.5 VM restart order/priority

When a failure occurs, and VMs need to be restarted, the order in which they come up can be critical. Fortunately, VMware provides a mechanism to set both the order and priority. The VM restart priority consists of the following categories:

- Lowest
- Low
- Medium
- High
- Highest

By default, all VMs have a priority of Medium.

Order priority can be essential for VMs that provide the infrastructure of an environment such as DNS or single sign-on. In addition, multi-tier applications like Oracle Apps that consist of a web, application and database tier, require that the database is running before the application can start, and in turn the application must be running before the web tier can start. VMware offer VM Overrides for these situations.

4.5.1 VM Overrides

The **VM Overrides** is accessible from the **Cluster -> Configure -> Configuration -> VM Overrides** menu here in Figure 47.

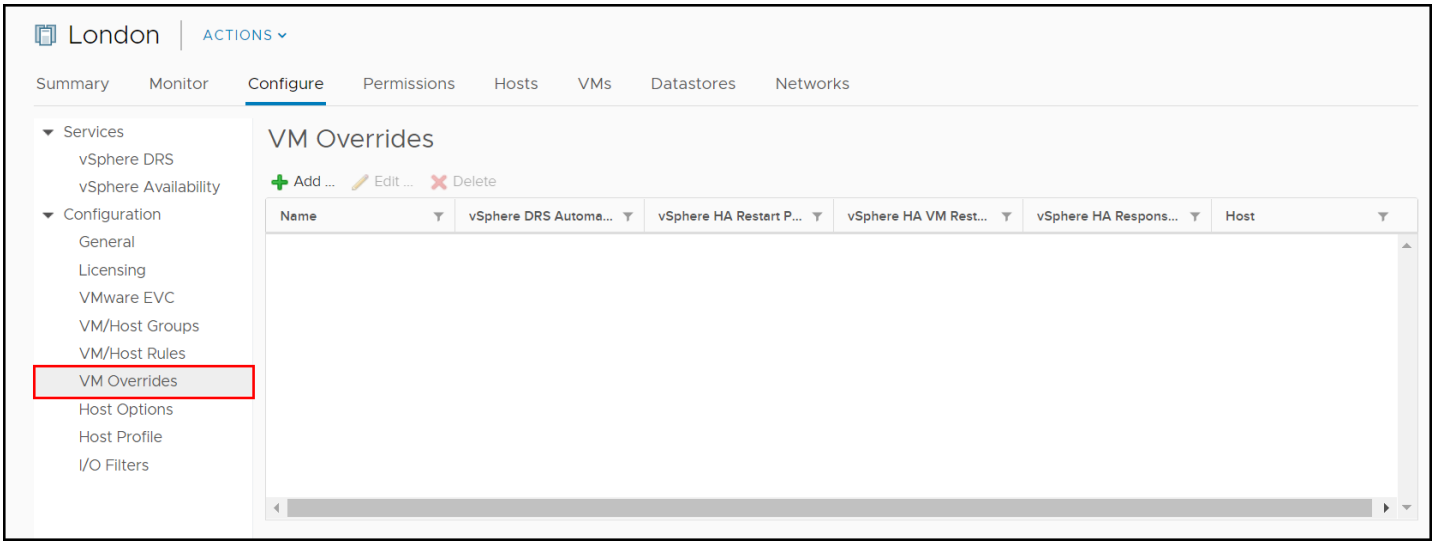

Figure 47. VM Overrides

The wizard has two steps. In the first step in Figure 48, select the VMs that require special priority.

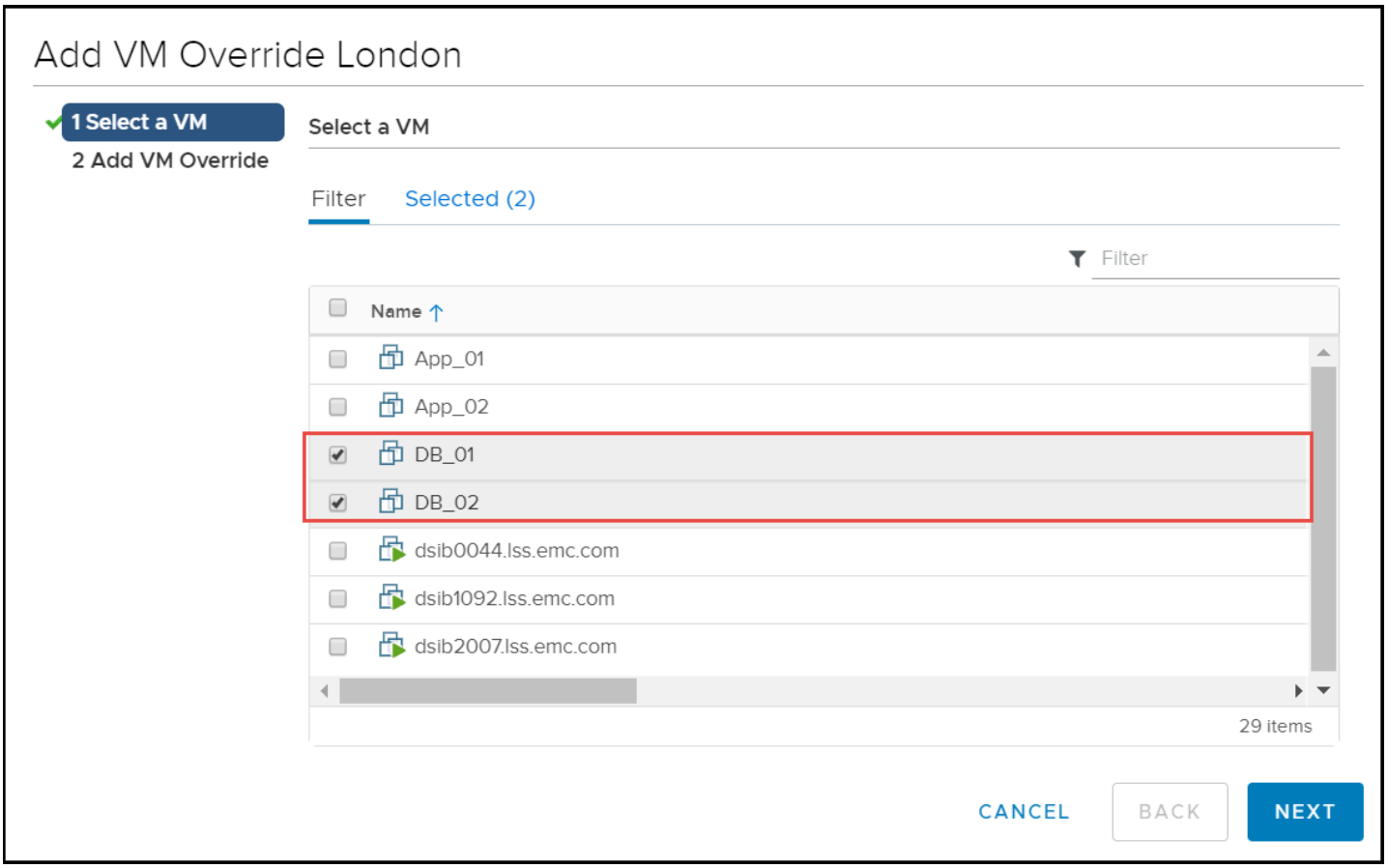

Figure 48. VM Overrides - Step 1

In the second step, override the restart priority and set it to the highest, to ensure those VMs, in this case database VMs, start first. Also, select the next box to override when to start the next VMs. In this example in Figure 49, it will use the guest heartbeat to determine when to start the next highest priority VMs.

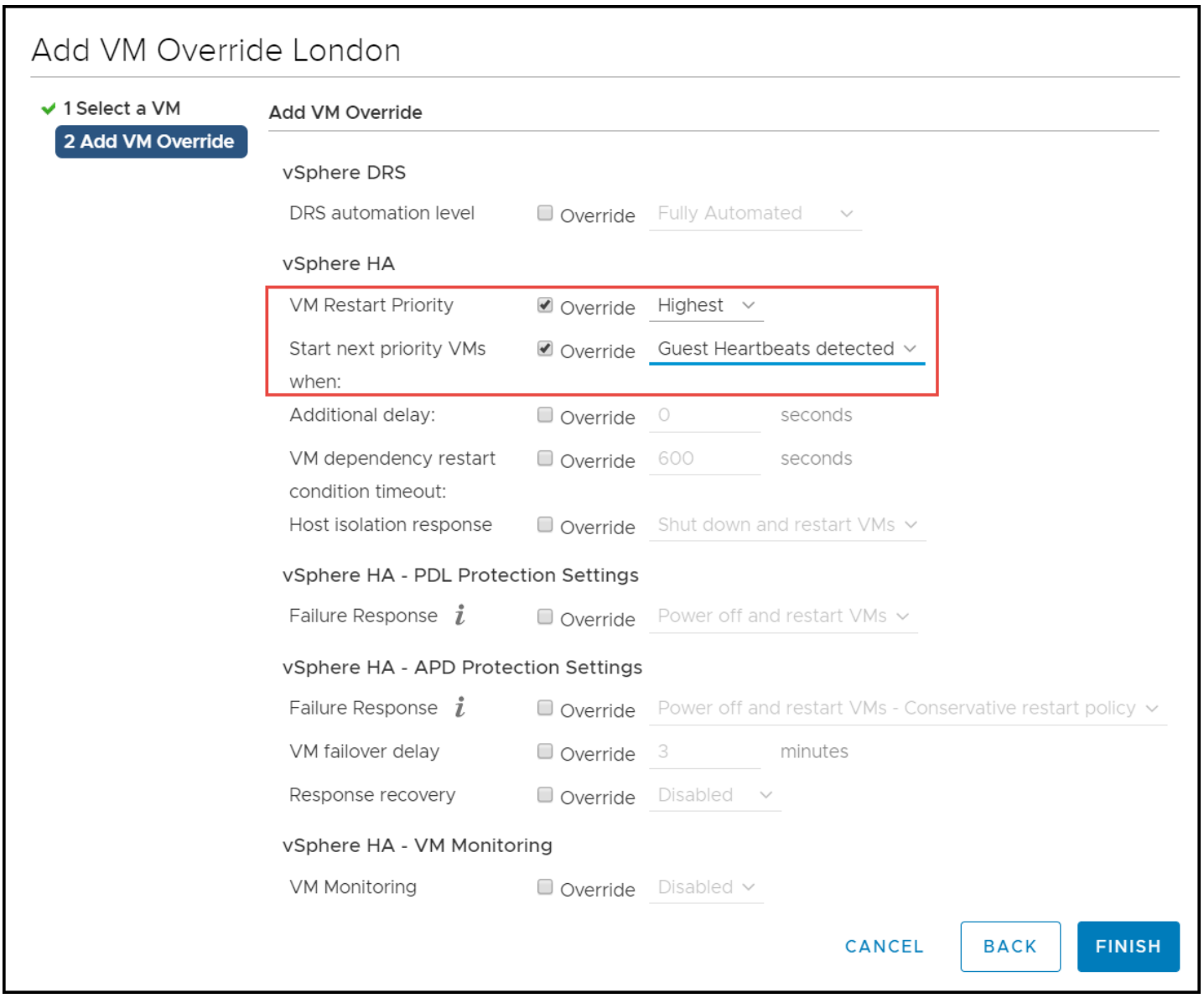

Figure 49. VM Overrides - Step 2

By default, vSphere will start the next VMs after 600 seconds even if the heartbeat is not detected. VMware does not recommend adjusting this value. In this example, a second VM Override would be created for the application tier as High, then the web tier as Medium.

4.5.1.1 VM startup dependency

A second method can be employed to determine startup order. This involves setting up the previously discussed VMware VM/Host Groups and VM/Host Rules. Rather than configuring site affinity, VM Groups can be setup with dependencies upon one another. For example, using the database and application tiers, setup a VM Group for each as in Figure 50.

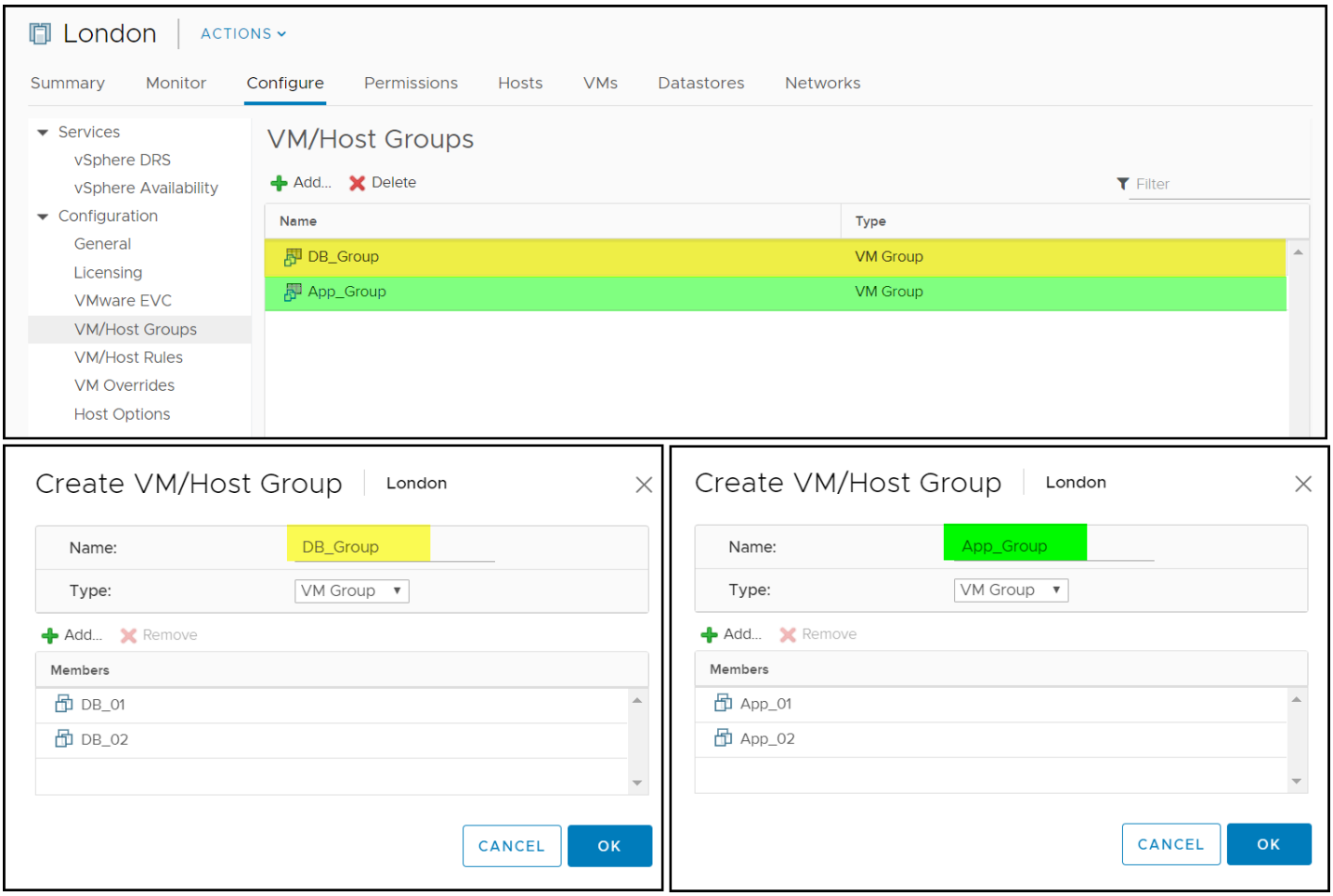

Figure 50. VM Groups for startup dependency

Now, setup a VM Rule which governs how the two VM Groups should interact. In Figure 51 the **Oracle_Startup** rule says that before the **App_Group**, containing the application VMs, can startup, the **DB_Group**, containing the database VMs, must first start. One significant difference between this startup method and the previous one using VM Overrides, is VM Rules cannot be broken. In other words, if the **DB_Group** fails to start, vSphere will not start the **App_Group**. Recall that after 600 seconds the startup priority moves forward, regardless of whether the higher priority VM startups have succeeded.

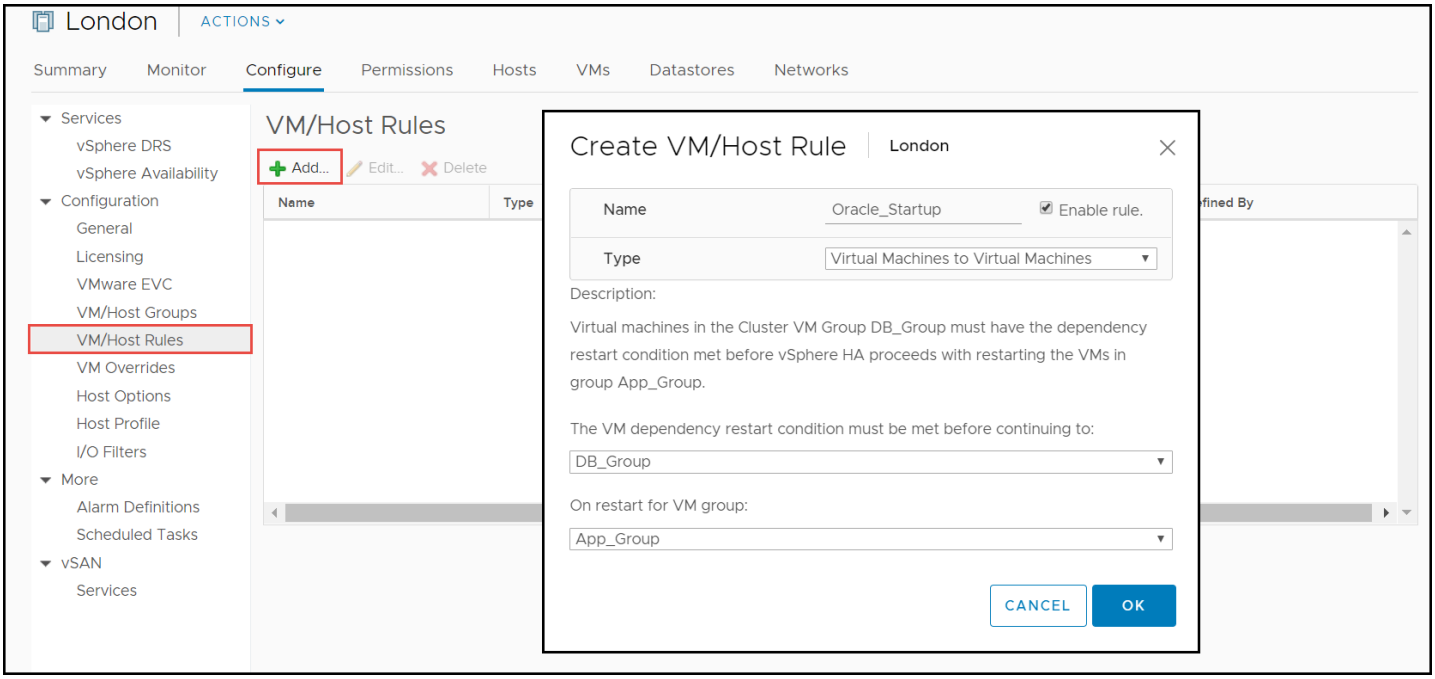

Figure 51. VM Rule for startup dependency

5 Conclusion

Dell SRDF/Metro is an enterprise-class technology that dissolves distance by providing active/active access to dispersed PowerMax arrays enhancing availability and mobility. Using SRDF/Metro in conjunction with a VMware vSphere Metro Storage Cluster (vMSC), including vSphere DRS and HA, provides new levels of availability suitable for the most mission critical environments without compromise.

These technologies provide the basis by which a customer can ensure high availability at both the hardware and software level – through the nature of SRDF/Metro and vMSC.

6 References

6.1 Dell

- The Dell PowerMax and VMware vSphere Configuration Guide
- Dell SRDF/Metro Overview and Best Practices Technical Notes
- Dell SRDF/Metro vWitness Configuration Guide
- Using VMware Storage APIs for Array Integration with Dell PowerMax
- Unisphere for PowerMax Product Guide

6.2 VMware

- vSphere General Documentation (https://docs.vmware.com/en/VMware-vSphere/index.html)
- SRDF/Metro vMSC support VMware KB article (http://kb.vmware.com/selfservice/microsites/search.do?cmd=displayKC&docType=kc&externalId=2134684)
- VMware vSphere Metro Storage Cluster Recommended Practices (https://core.vmware.com/resource/vmware-vsphere-metro-storage-cluster-vmsc)

Appendix

This appendix covers the steps required to setup SRDF/Metro for use in a vMSC environment. Unisphere for PowerMax 10.0.1.0 is utilized, though earlier versions of Unisphere have similar interfaces.

1.1 Setting up SRDF/Metro

SRDF/Metro can be configured with Solutions Enabler (CLI) or with Unisphere for PowerMax. Dell recommends using Unisphere to configure SRDF/Metro to reduce complexity. This section details the setup and includes those tasks directly related to SRDF/Metro setup. In this example, the following objects are assumed to already exist as their creation is independent of SRDF/Metro:

- Initiator groups for each site
- Port groups for each site
- Devices on each site created and placed in a single storage group on each array
- A masking view exists for the R1 array, but NOT for the R2

In the included example, a vWitness is used as that is the most common implementation, but for completeness the process to create an array witness is covered here.

1.1.1 Array witness group creation

To create an array witness, within Unisphere for PowerMax navigate to the **Data Protection -> SRDF Groups** and select **Create SRDF Group** as in Figure 52.

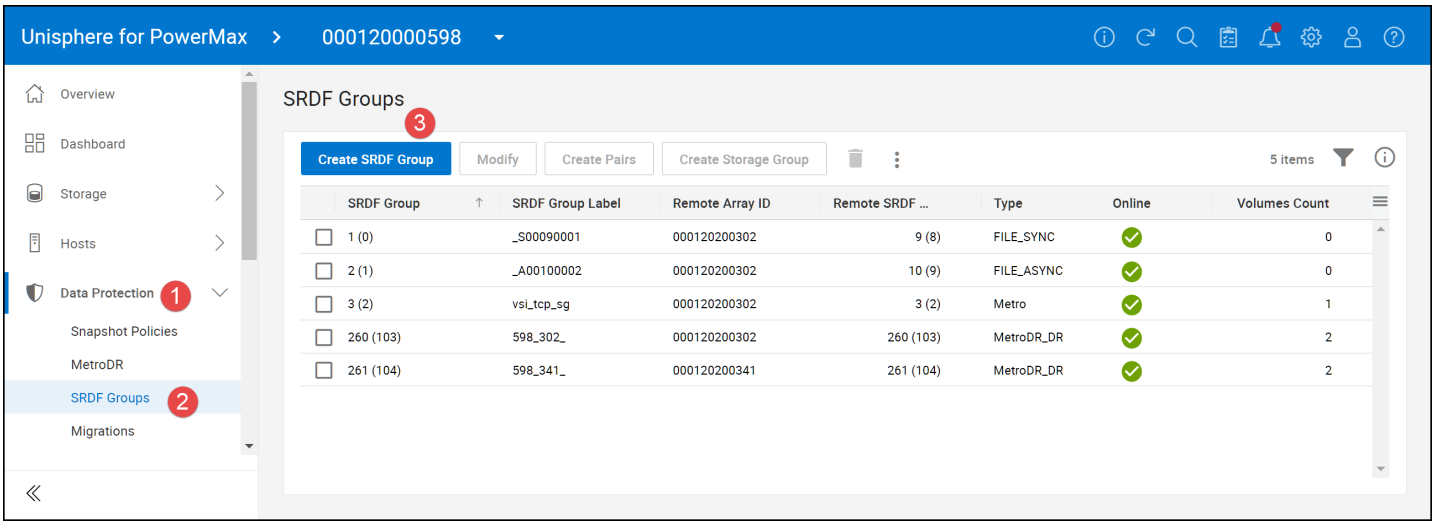

Figure 52. SRDF group creation in Unisphere for PowerMax

Provide an SRDF Group Label to designate the function of the group. In this example in Figure 53, the label “aw-302598” is used to indicate this is a witness group between the R1 302 array and the witness array, 598. Note that the group here is created from the witness array but it could be created from the R1 as well. Next, check the box **SRDF/Metro Witness Group**.

If an array witness group already exists to the remote array, the setting for SRDF/Metro Witness Group will be grayed out.

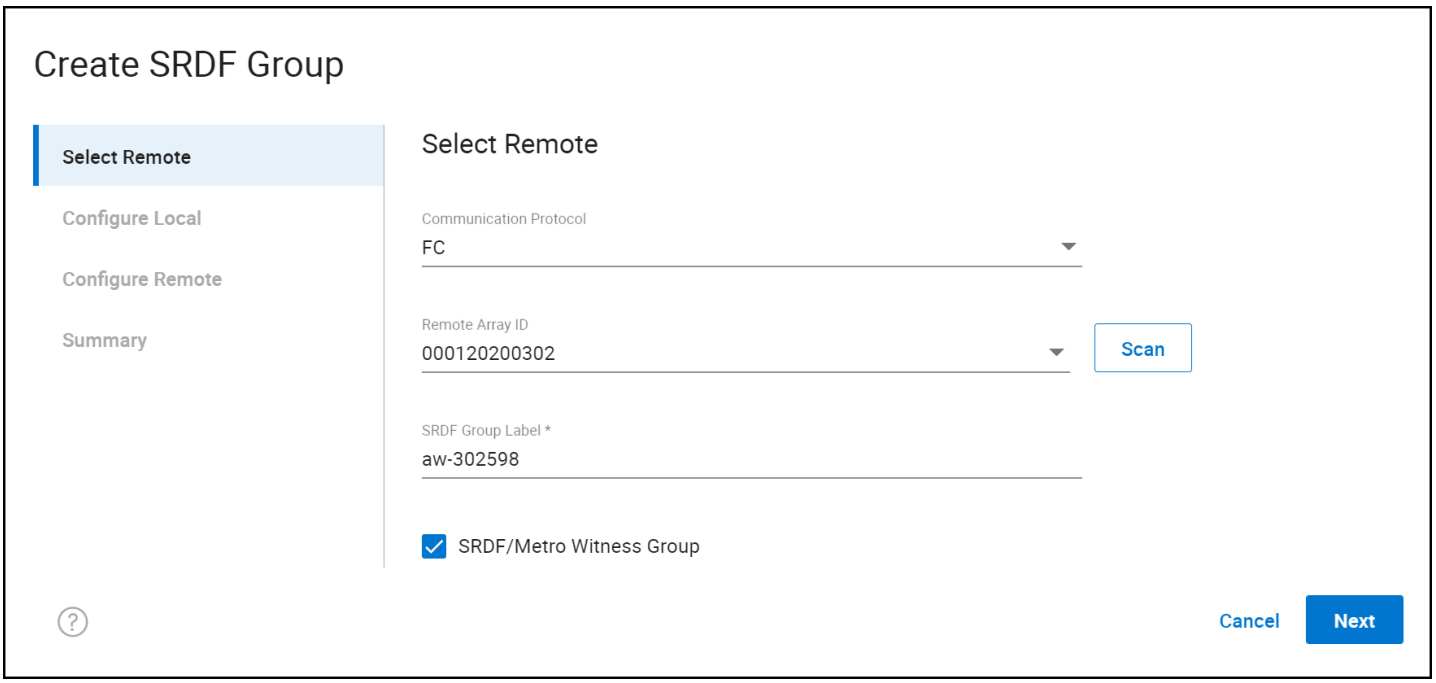

Figure 53. Creating the SRDF/Metro Witness Group from R1 to Witness array -

In steps 2 and 3, select the RDF ports and assign an RDF group. It is best to use the same number on each array for consistency. In this example, group 65 is used. Figure 54 is the summary in step 4 before creation.

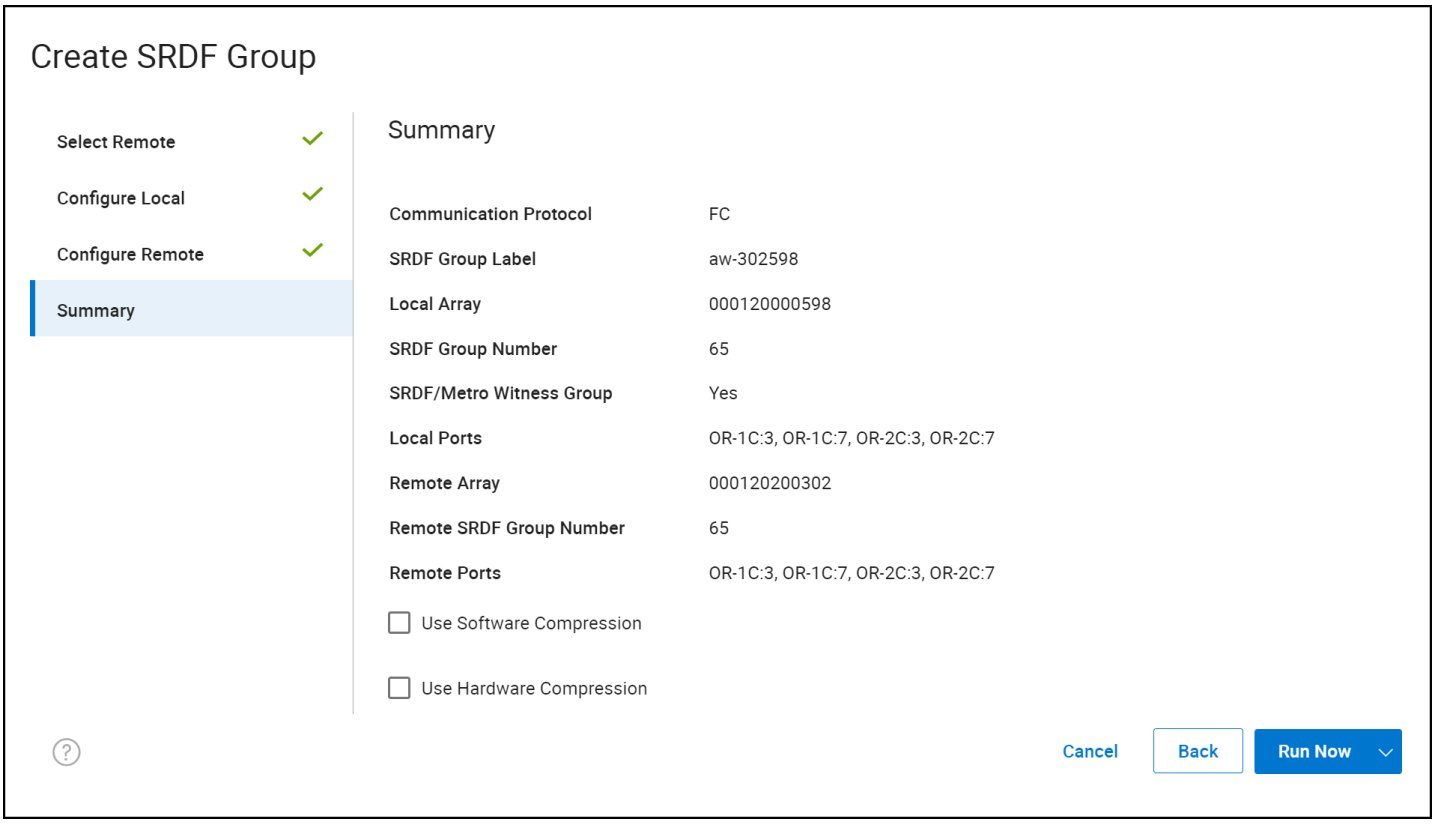

Figure 54. Creating the SRDF/Metro Witness Group from R1 to Witness array - Summary

The same steps then need to be run between the R2 array, 341, and the witness array 598. A new SRDF group number will be required as 65 was used previously on the R1. When complete, compare the two groups in Figure 55 and note that the row **SRDF Group Type** is designated as **Witness**.

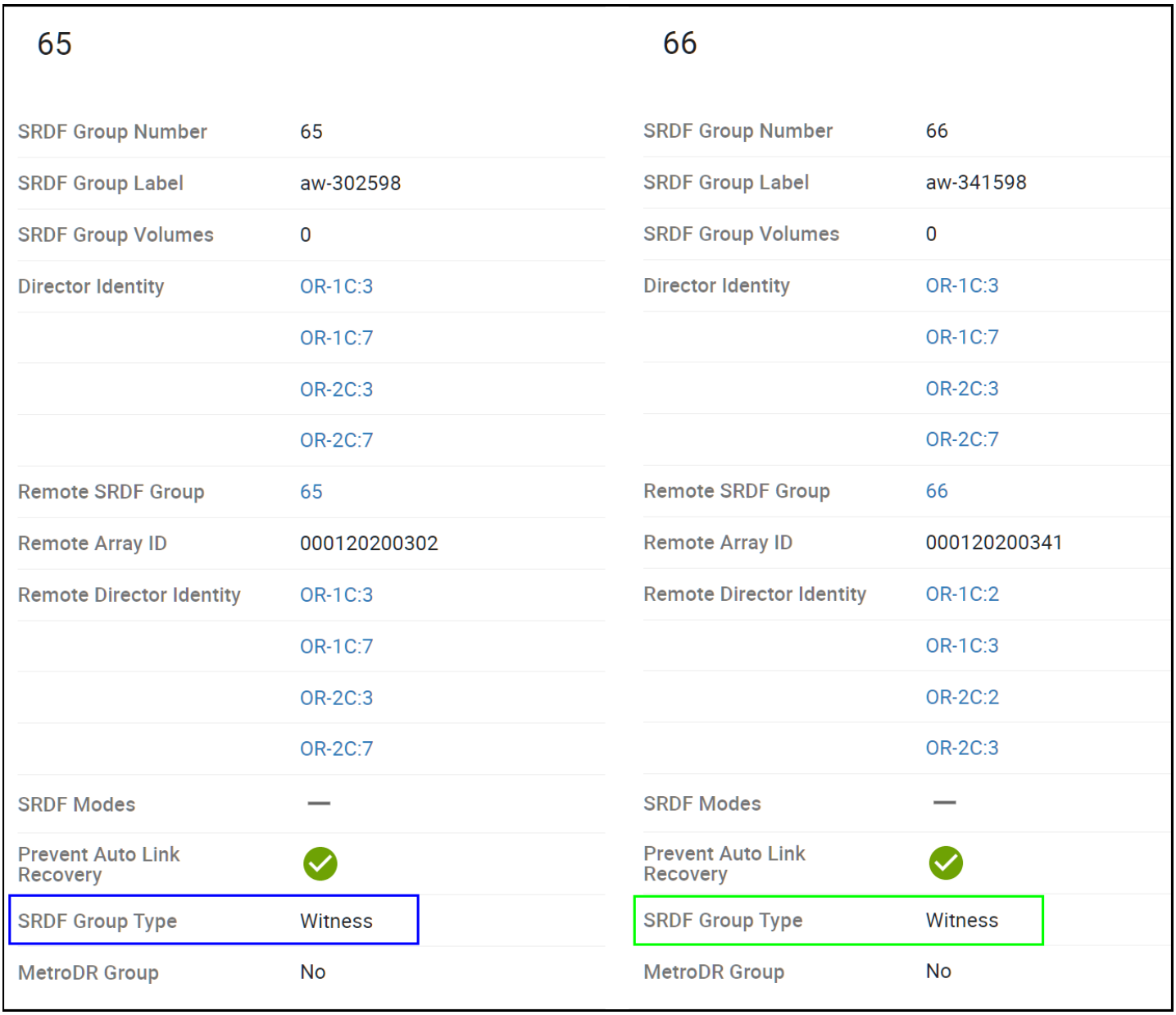

Figure 55. Witness group summary

To add an SRDF group for the witness using Solutions Enabler, specify the **`-witness`** flag to a standard SRDF group creation statement shown in Figure 56. Note different arrays are used here in the CLI than the GUI.

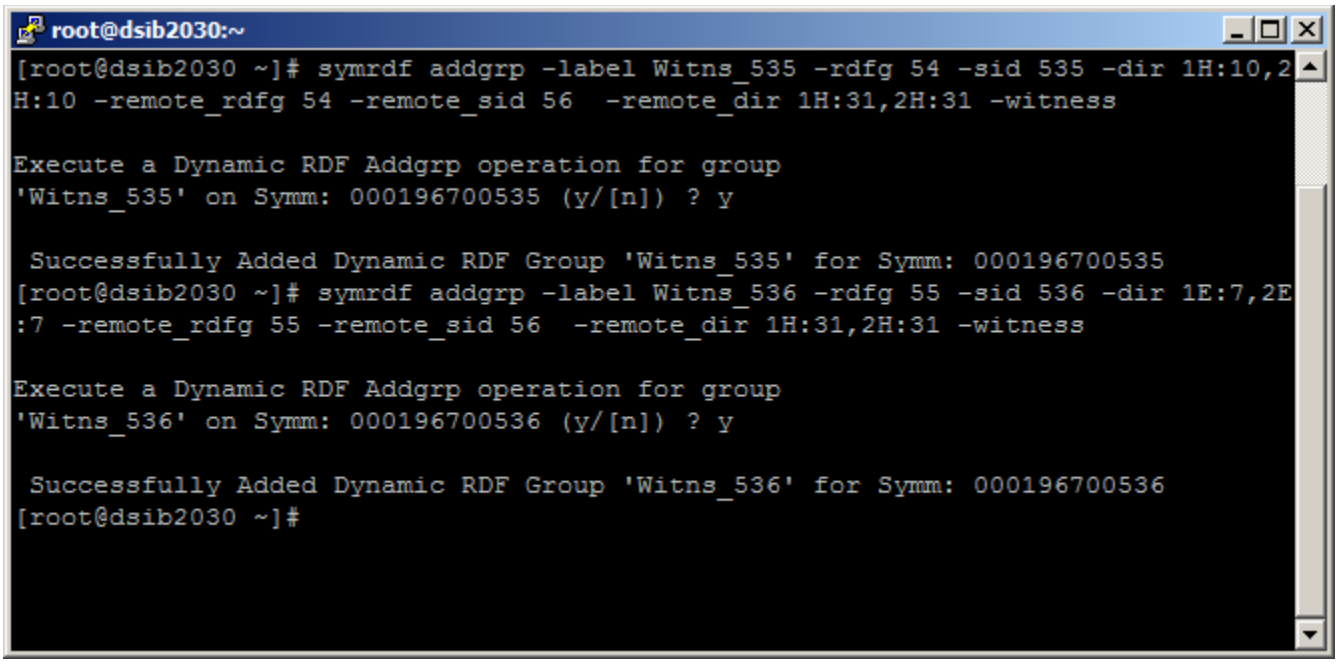

Figure 56. Creating SRDF groups for the Witness through CLI

If a Witness group already exists, whether or not it is in use, Solutions Enabler will return the following error: *A Witness RDF group already exists between the two Symmetrix arrays*.

1.1.2 vWitness creation

Beginning with PowerMaxOS 10, the virtual witness must be deployed on a supported OS as the virtual appliance is deprecated. The software is packaged like Solutions Enabler and is deployed similarly. After deployment, the Lock Service Daemon should be running as in Figure 57.

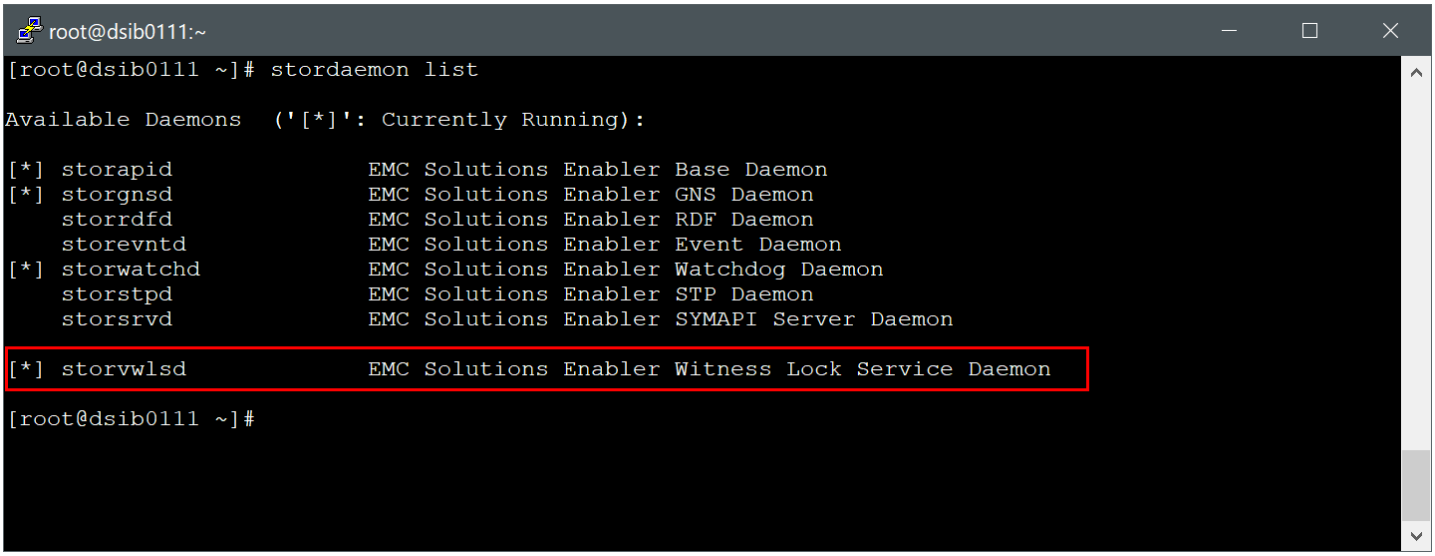

Figure 57. vWitness deployment

1.1.3 Witness use

Though the status of an SRDF/Metro pair will designate whether the witness is in use, it is not possible to tell what type, array or virtual, within Unisphere. Only in Solutions Enabler CLI is this information found. The `symcfg` command can be used to reveal the name and type of witness. Figure 58 shows the command to issue to see witness detail. Note the column **Witness Identifier**. If this field has a value for a row, it indicates there is a witness associated with the group. There are two types of values in this row. If the array witness is being used, the value will be the witness array SID as is present for **RA-Grp** 11. If a virtual witness is in use, the value will be the name given to the virtual witness as in **RA-Grp** 40 and 41.

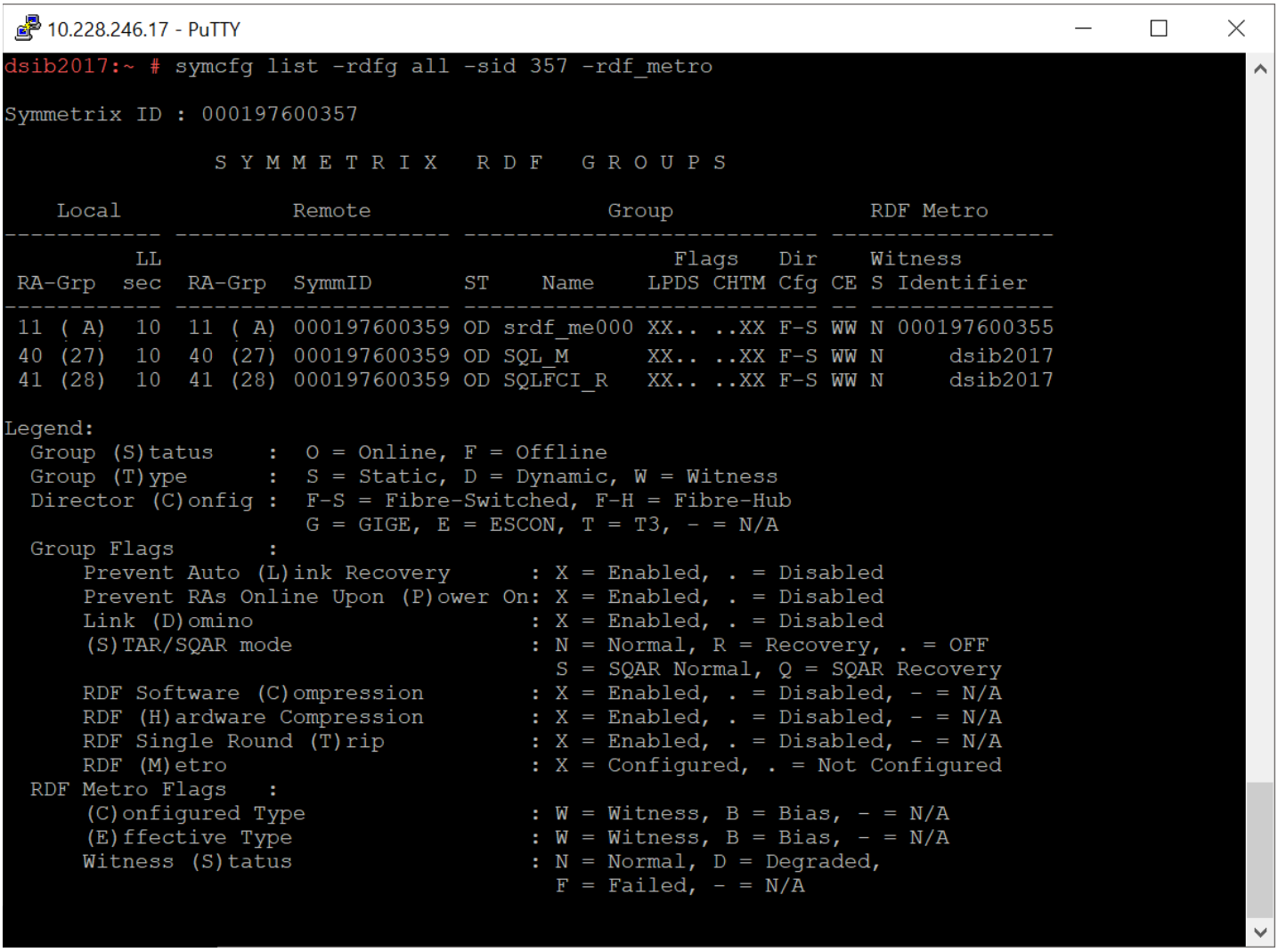

Figure 58. Witness detail

1.1.4 SRDF/Metro pair creation

The Storage Groups screen in Unisphere provides an easy-to-use wizard to enable SRDF/Metro on a storage group. Before starting the wizard, it is important to remember that though this is an active-active device configuration, the R2 should NOT be presented to the hosts until all devices are fully synchronized.

Start by navigating to the **Storage -> Storage Groups** seen in Figure 59. Highlight the desired storage group for protection and select the button **Protect** at the top menu.

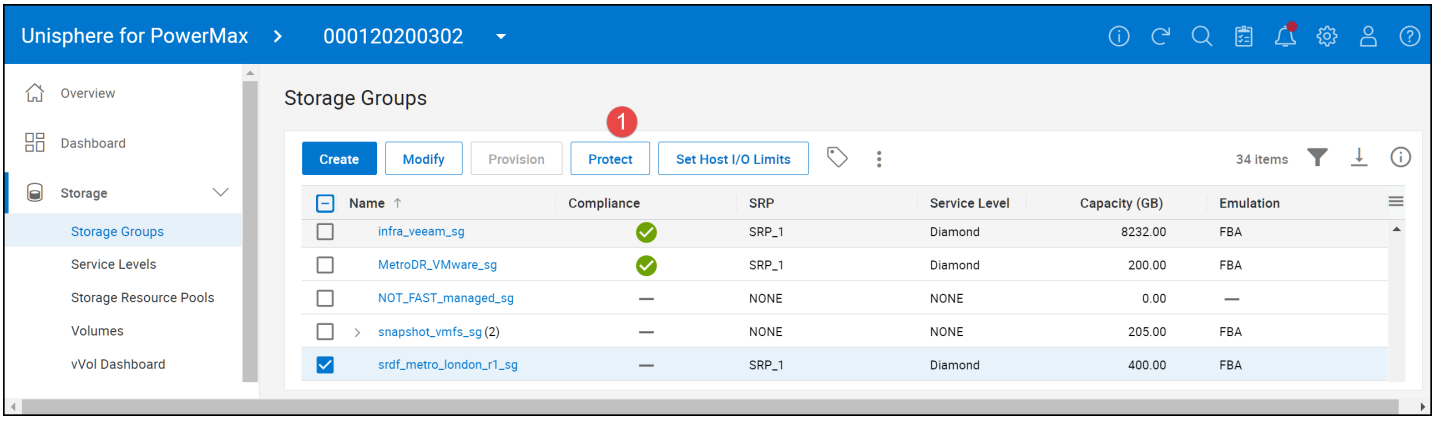

Figure 59. Storage group selection for SRDF/Metro setup in Unisphere

With the storage group selected, Unisphere provides the available protection types on the array. Select **Setup High Availability using SRDF/Metro** radio button and then **Next**. This step is shown in Figure 60.

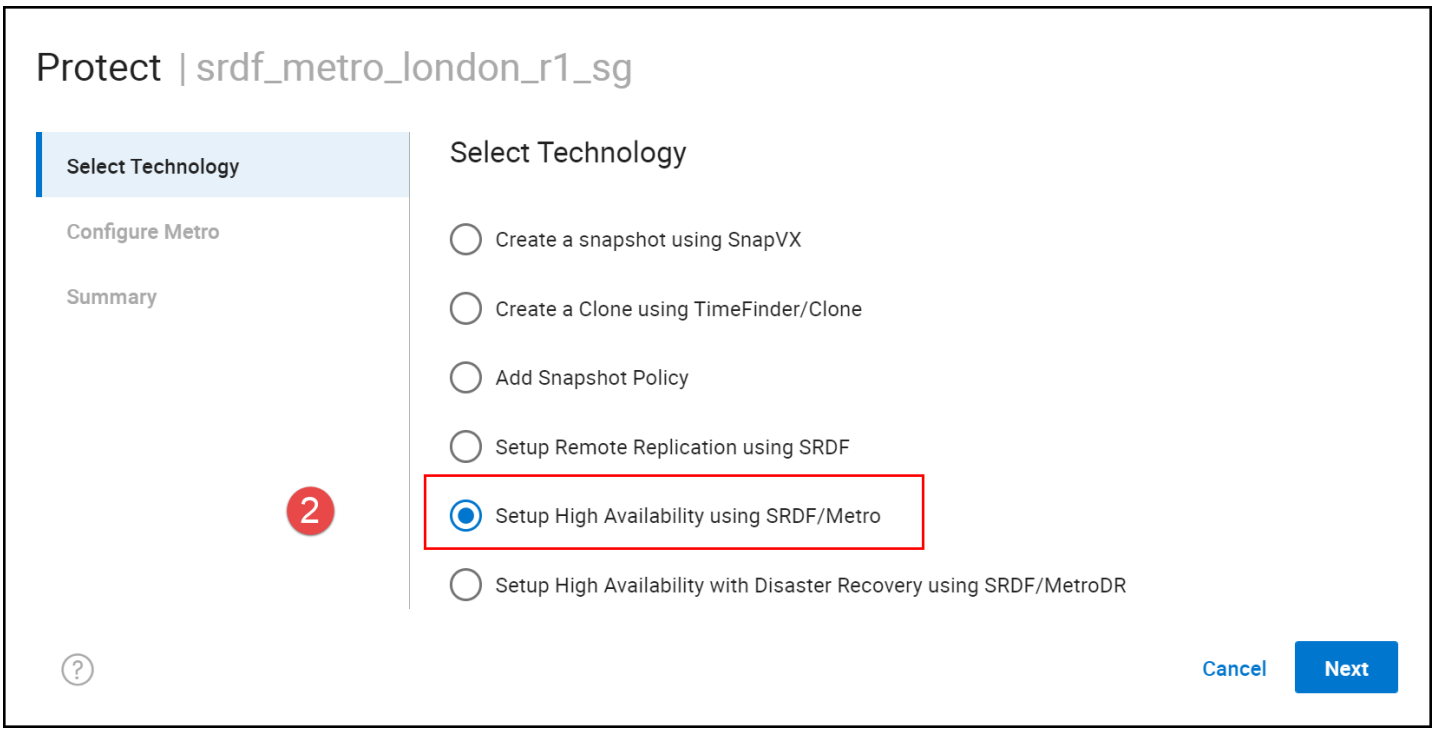

Figure 60. Protection Type in protection wizard

In step 3, Figure 61, Unisphere will default the following:

- Automatic for the **SRDF Group**
- Box checked for Establish SRDF Pairs
- Bias for **Protected By** – this should be changed to Witness as best practice. Note that if a witness is unavailable, the option will be grayed out.

- The remote storage group name will be set to the same name as the group being protected, though it can be changed.
- The same service level as the group being protected
- Box checked for **Enabled Data Reduction** if enabled on the R1

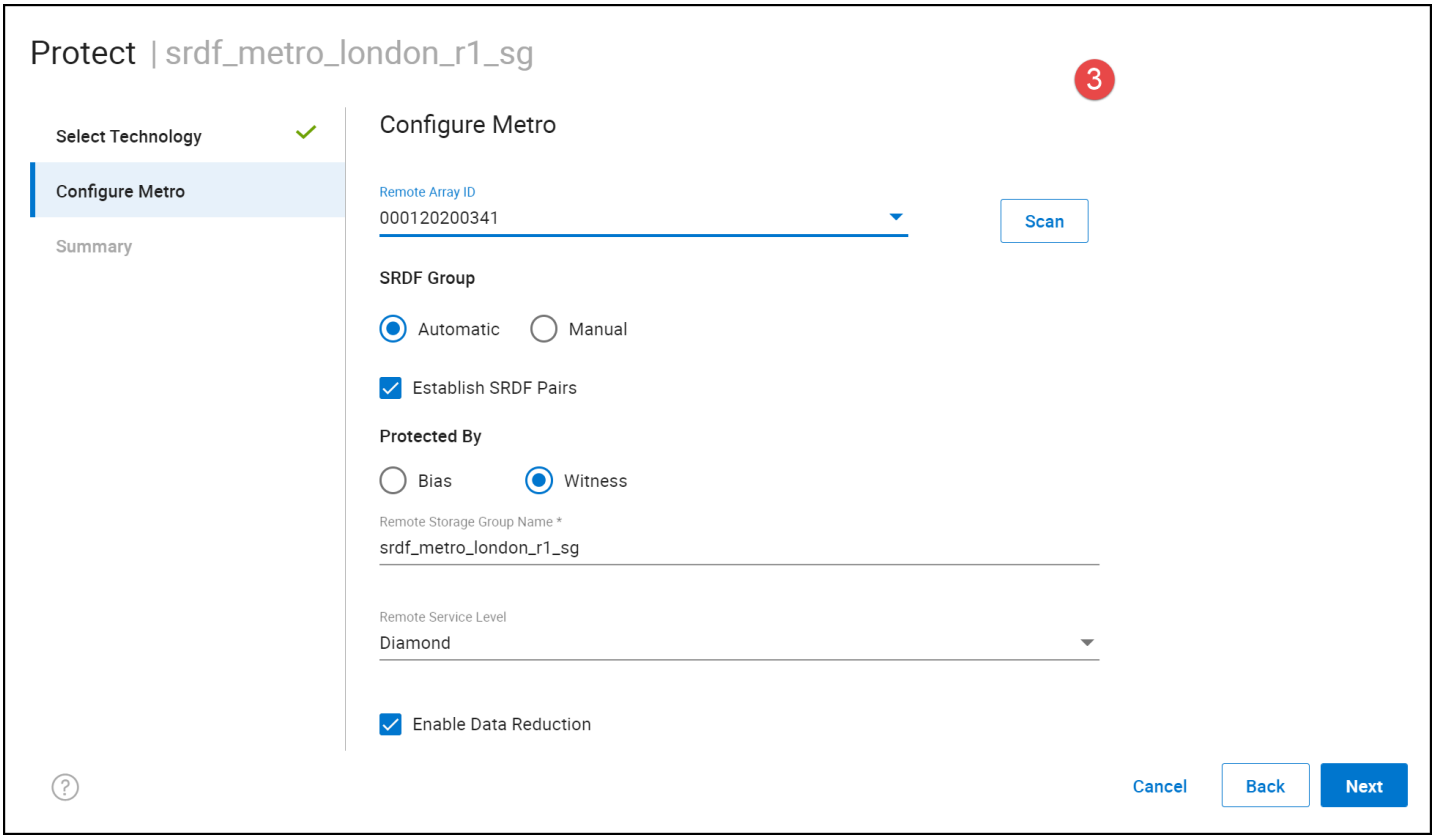

Figure 61. SRDF/Metro Connectivity in in protection wizard

Finally review the proposed changes in step 4 in Figure 62 and run the task.

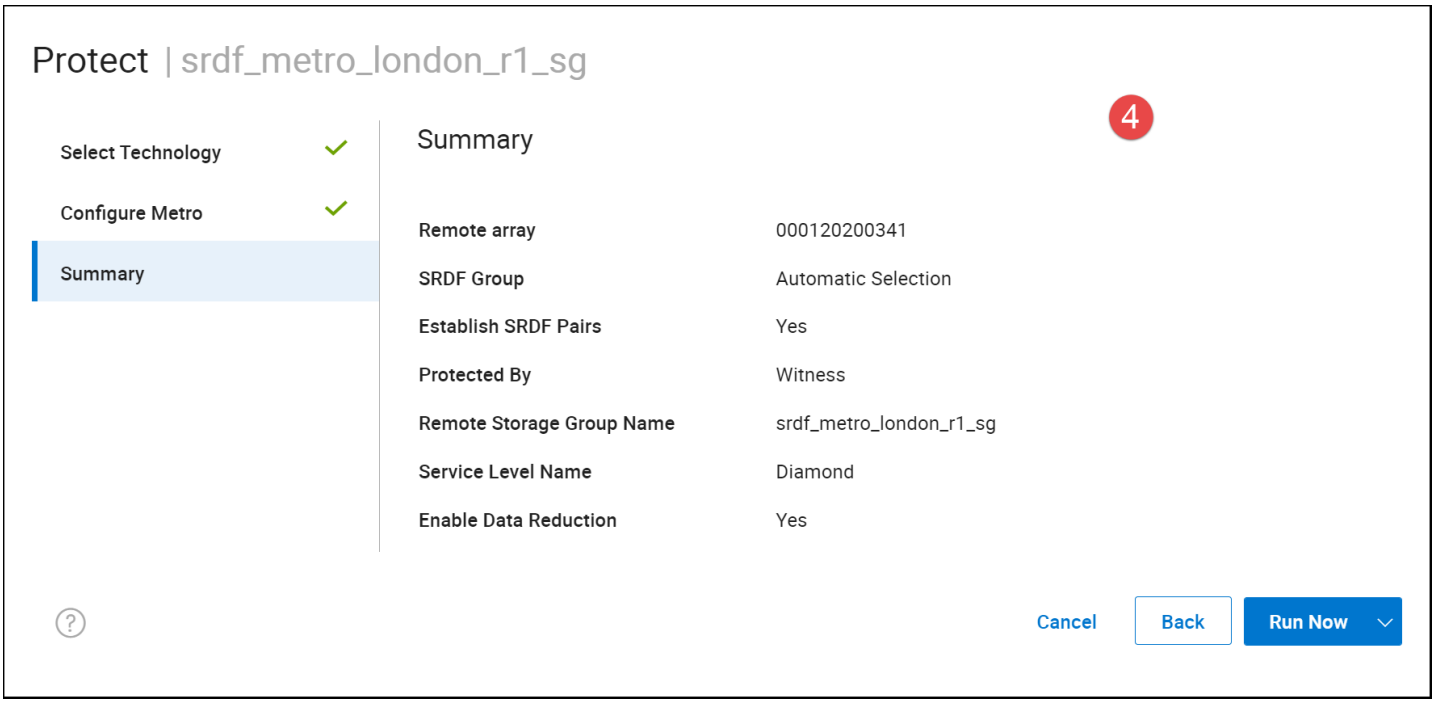

Figure 62. Complete protection wizard

Once completed, the SRDF/Metro dashboard in Figure 63 will show the group syncing.

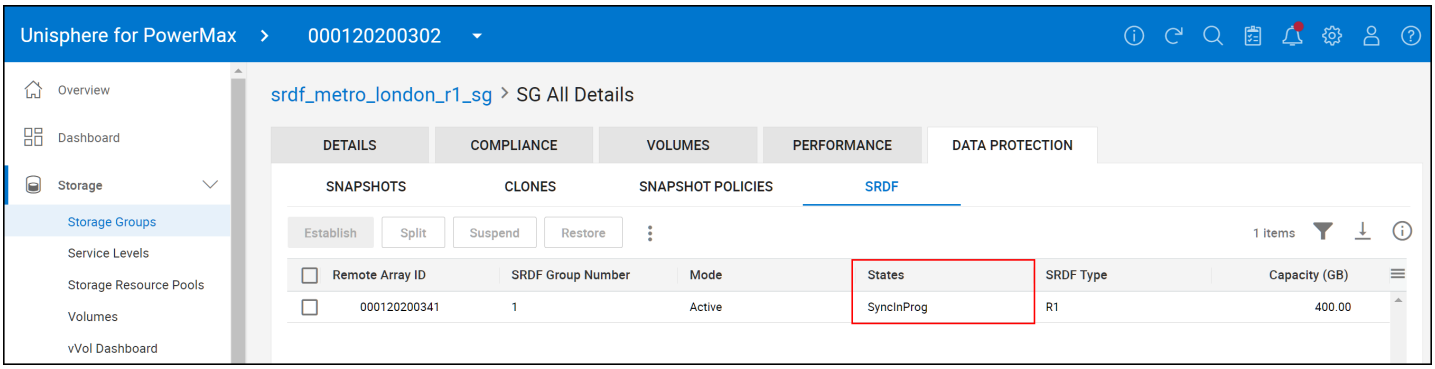

Figure 63. SRDF/Metro dashboard

The user should wait until the State in Figure 63 reads **ActiveActive** before creating a masking view for the R2 devices. That will mean the synchronization is complete. Alternatively, if using bias, the synchronized state is **ActiveBias**. Solutions Enabler may also be used to verify the state of the RDF group as in Figure 64.

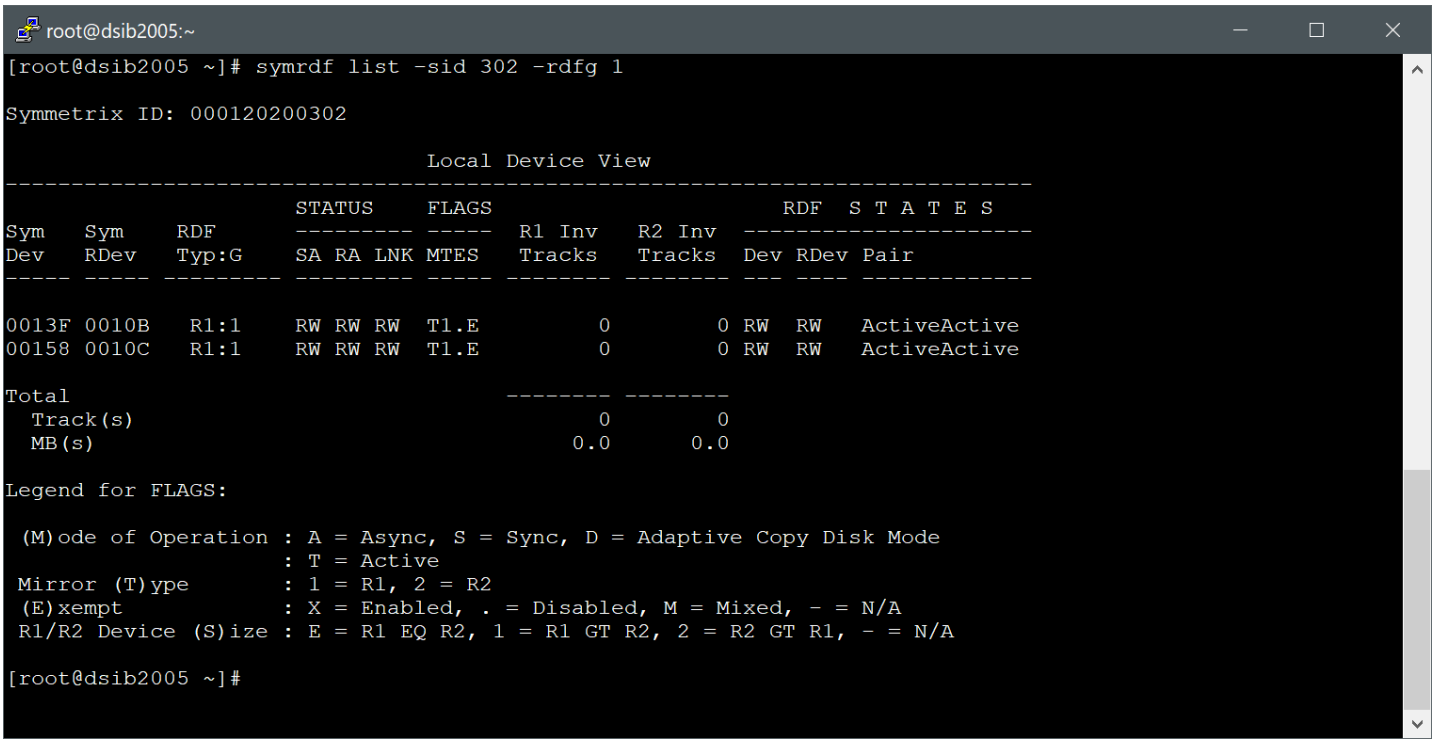

Figure 64. Verify SRDF/Metro state in CLI

Once the state is in the proper active state, the R2 devices can be presented to the R2 host(s) using the masking view wizard in Unisphere, displayed in Figure 65. In this example the masking view name changes the r1 to r2. This is useful initially, but recall if there is a failure, or a re-establish after a change in architecture (e.g., a cascaded leg added), the R2 may become an R1.

Figure 65. Create R2 masking view

As shown above, when setting up SRDF/Metro pairs, Unisphere for PowerMax wizards work at the storage group level. If the devices are not already in a storage group, Unisphere can still be utilized in more of a manual fashion. For instance, pairs can be created in the **SRDF Groups** screen in the Data Protection area of Unisphere. Here, using the checkboxes, check one of the groups and then select **Create Pairs** as in Figure 66.

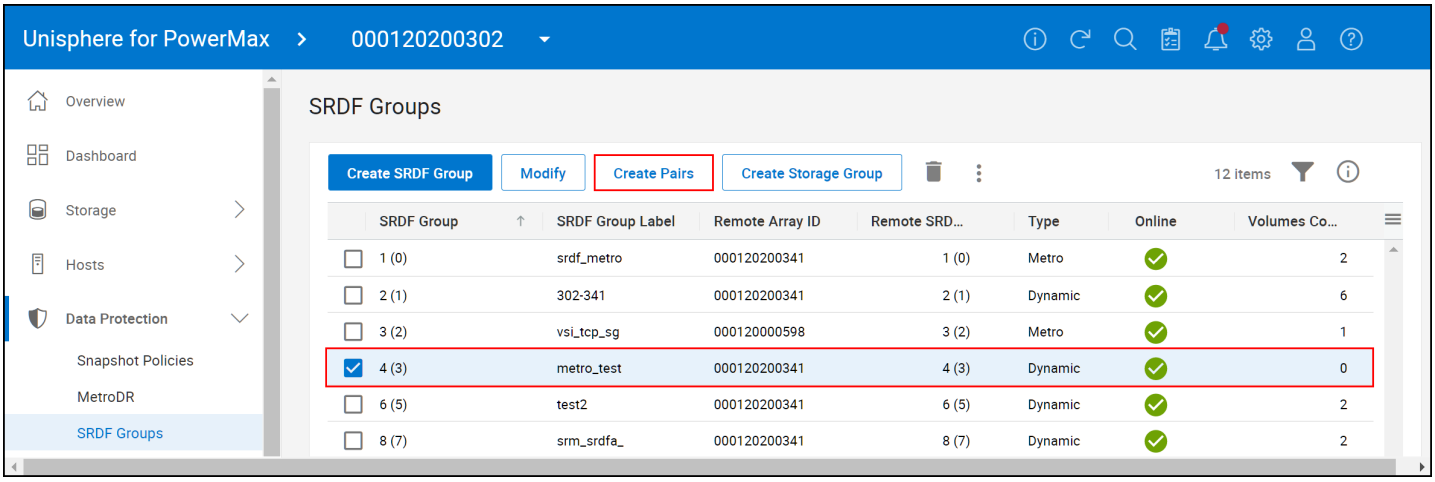

Figure 66. SRDF/Metro create pairs in an empty RDF group

A dialog box appears where the user can select the SRDF Mode **Active** from the drop-down list. At that point there is the ability to set **Establish** or **Restore** and **Bias** or **Witness** (if available). Choosing between Establish and Restore sets which device is the source for the data. Establish means SRDF will copy the data from the R1 to the R2, Restore means it will copy the data from the R2 to the R1. Note that choosing Restore does not mean the R2 becomes an R1. Unisphere then provides the ability to manually select existing local and remote devices, or Unisphere can look for available devices based on a designated size. If Unisphere cannot find available devices, it will create them. The devices may also be placed in an existing storage group on one or both sides. All four steps are shown in Figure 67.

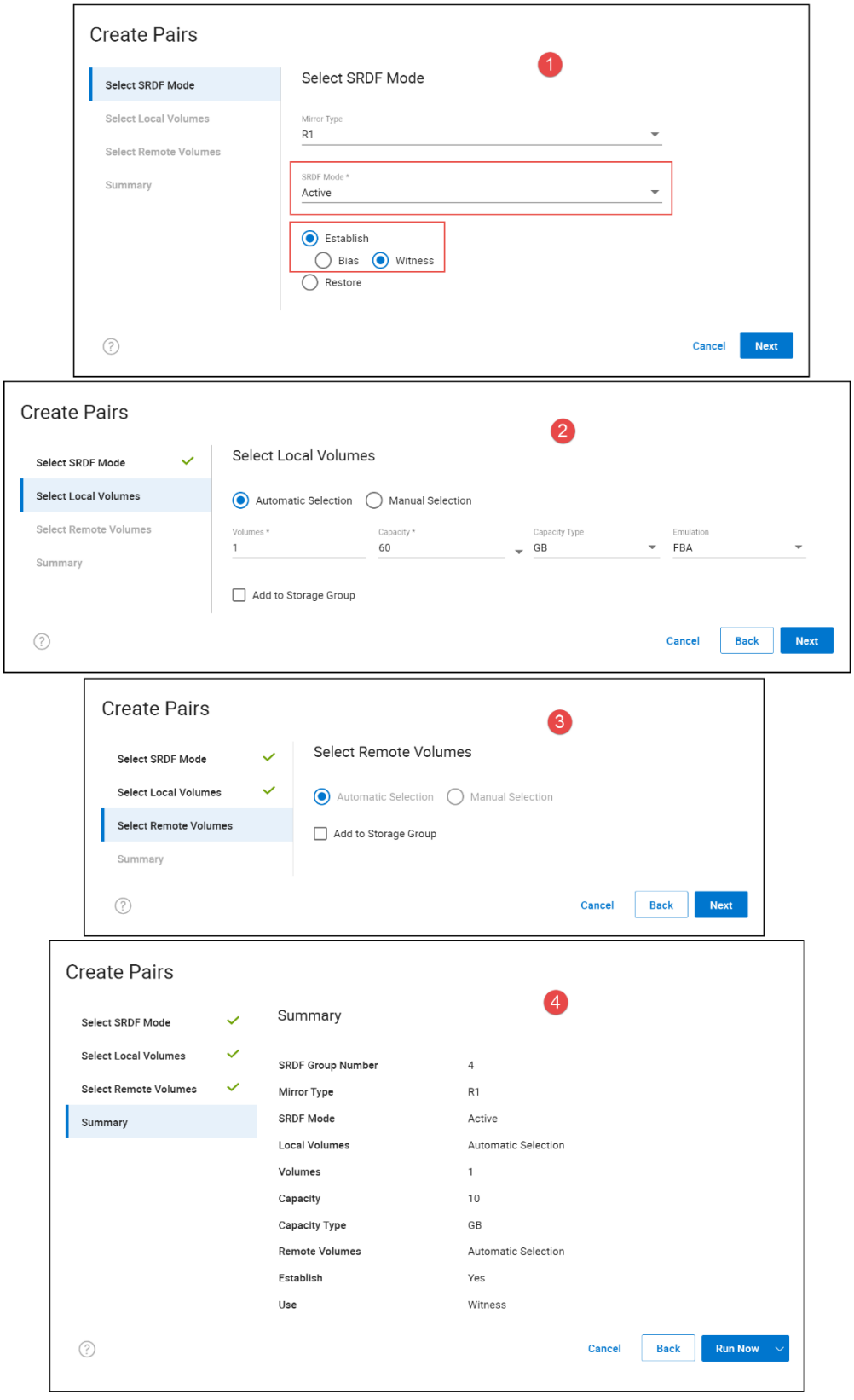

Figure 67. Creating SRDF/Metro pairs in Unisphere without a storage group

Note: If the user selects an existing SRDF/Metro RDF group, the SRDF mode will automatically be set to **Active** and cannot be adjusted as an RDF group with SRDF/Metro pairs cannot have a second SRDF mode present. Though Dell recommends using Unisphere, it is possible to setup SRDF/Metro using Solutions Enabler. In general, setting up SRDF/Metro is akin to any SRDF configuration, save for during the ***createpair*** command when the switch ***–metro*** should be specified for an SRDF/Metro pair.

1.1.5 Adding new pairs to an existing RDF group online

Beginning with PowerMaxOS 5978.144.144, existing devices can be dynamically added to a SRDF/Metro RDF group without any loss of data. Prior to this release, only new or formatted devices could be added to an existing group with the `-format` option, potentially causing data loss.