EMC® VPLEX™ Metro Cross-Cluster Host Connectivity

Implementation Planning and Best Practices

- Cross-Cluster Host Connectivity
- Host Multi-pathing with Cross-Cluster Connectivity
- OS Specific Cross-Connect Considerations

Abstract

This document reviews host connectivity SAN topologies for VPLEX Metro and describes the cross-cluster host connectivity best practices for VPLEX Metro.

December 2014

Published December 2014

EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

The information in this publication is provided as is. EMC Corporation makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose. Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.

For the most up-to-date listing of EMC product names, see EMC Corporation Trademarks on EMC.com.

EMC VPLEX Metro Cross Connect Best Practices

Part Number H13813

Contents

Chapter 1 Introduction

This chapter presents the following topics:

Audience

These technical notes are for EMC field personnel, partners, and customers who configure, install, and support VPLEX. An understanding of these technical notes requires familiarity with the following:

- SAN technology and network design
- Fibre Channel block storage concepts
- VPLEX concepts and components

Summary

This document reviews VPLEX Metro host connectivity models and addresses the best practices for the optional cross-cluster host connectivity. The best practices for standard host connectivity to VPLEX are contained in the *EMC VPLEX SAN Connectivity Best Practices Guide* available online at https://support.emc.com.

Document Organization

This technical note is one of a set of documents that supersede the monolithic Implementation Planning and Best Practices for EMC VPLEX Technical Notes document that had previously been available.

All documents are available on https://support.emc.com.

The following list represents the technical note best practice documents that will be made available:

- EMC VPLEX Overview and General Best Practices
- EMC VPLEX SAN Connectivity
- EMC VPLEX Metro Cross-Cluster Host Connectivity
- EMC VPLEX Host Multipathing
- EMC VPLEX IP Networking
- EMC VPLEX and RecoverPoint Interoperability
- EMC VPLEX Data Mobility and Migrations
- EMC VPLEX REST API
- EMC VPLEX Upgrades
- EMC VPLEX Virtual Edition

Host Connectivity with VPLEX Metro Review

Standard host to VPLEX connectivity, also referred to as front-end connectivity, consists of each host having multiple, redundant fibre channel SAN connections to a VPLEX cluster. The host to VPLEX connectivity design shown in Figure 1 remains the same regardless of whether the VPLEX consists of one site (VPLEX Local) or two sites (VPLEX Metro and Geo). Standard VPLEX Metro host connectivity is referred to as non-uniform connectivity as hosts are not required to have the exact same connectivity to each VPLEX cluster in order to access storage. VPLEX caching and hardware architecture is designed for extremely efficient non-uniform host connectivity.

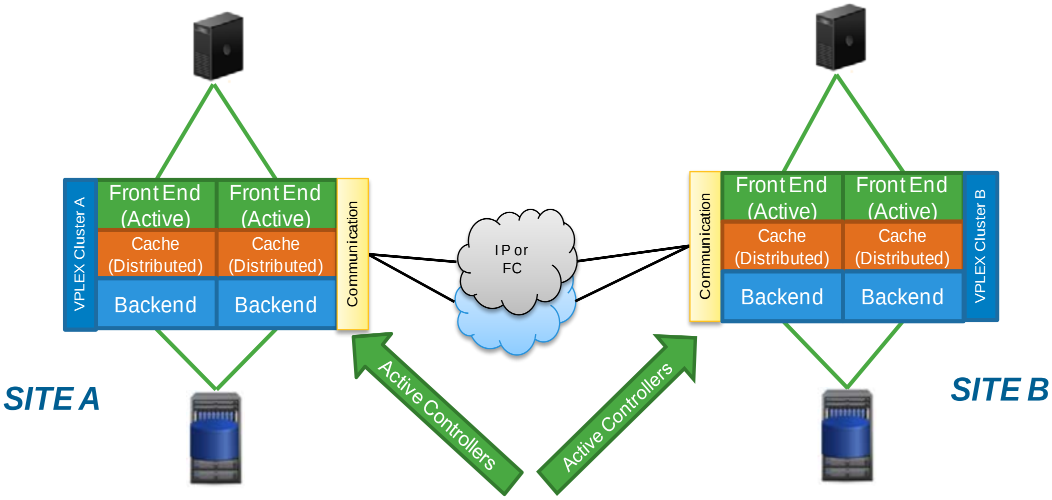


Figure 1 VPLEX Metro Standard Non-Uniform Host Connectivity

Non-uniform host connectivity is a significant advantage for VPLEX Metro compared to other solutions that require uniform host connectivity. VPLEX Metro eliminates the need for cross site host zoning, additional fibre SAN switches, additional fibre channel adapters, additional infrastructure, and added complexity. In fact, for VPLEX Metro IP there is no cross-site fibre channel requirement whatsoever. These advantages translate into more efficient WAN bandwidth utilization, fewer round trips on reads and writes, lower implementation costs, and lower total cost of ownership.

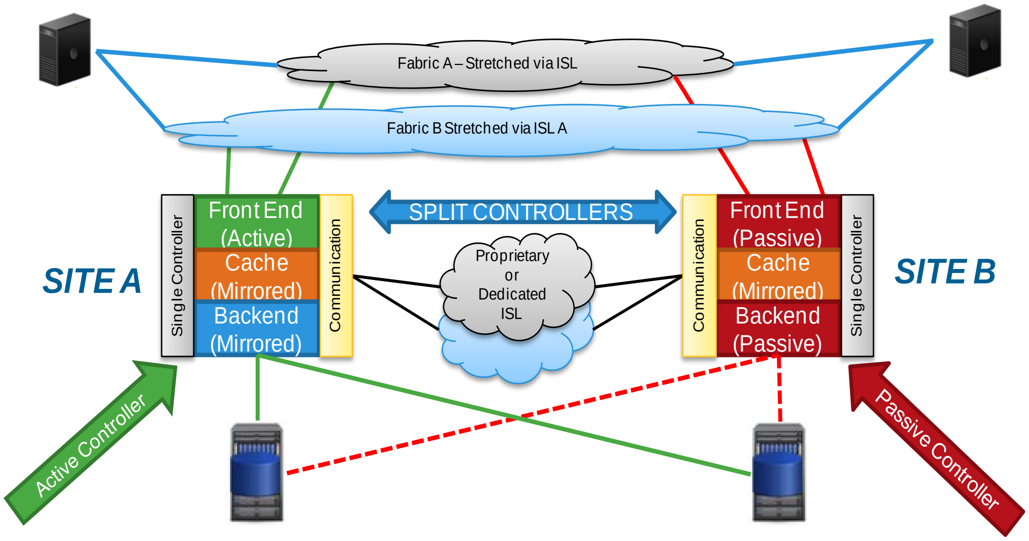


Figure 2 Non-VPLEX Solutions Require Uniform Host Connectivity

As shown in Figure 2, non-VPLEX split controller solutions require uniform host connectivity. Hosts at each site must connect to both controllers by way of the stretched SAN fabric; however the active controller (for any given LUN) is only at one of the sites (in this case site A).

While not as efficient as VPLEX, under normal operating conditions (i.e. where the active host is at the same location as the active controller) this type of configuration functions satisfactorily, however the host I/O access pattern starts to become sub-optimal if the active host is performing its I/O at the same location where the **passive controller** resides. This suboptimal I/O situation would occur for any active/active host application or any time the active controller fails.

Note: Some split controller solutions claim to be active / active across sites by enabling host IO to pass through their *front-end* ports (i.e. an ALUA model) and then across the SAN ISL to be terminated and replicated at the remote storage controller. Architecturally, however, hosts are still interacting with one active controller (lun owner) that is responsible for all front-end I/O and back-end array I/O for a given storage volume.

Because VPLEX Metro provides synchronous, active/active copies of data at two sites it is possible to configure hosts using cross-cluster (uniform) connectivity. Cross-cluster host connectivity provides each host with connections to both the local and the remote VPLEX cluster. It is important to understand that the availability benefits of this type of host connectivity to Metro are limited. As illustrated later in this chapter, hosts using cross-cluster connectivity to VPLEX Metro only achieve improved application availability in very specific multi-failure circumstances and **only when the proper host, SAN, and WAN infrastructure is in place.**

Why Cross-Cluster Host Connectivity?

What are the primary reasons behind using cross-cluster host connectivity with VPLEX Metro?

- Protection against scenarios where one of the VPLEX Metro sites fails or suspends I/O and there are no other host, SAN, WAN, or power issues.

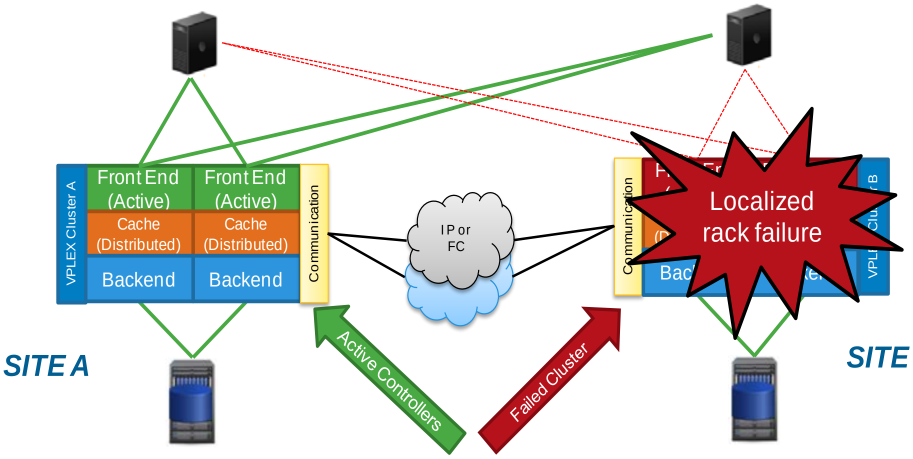


Figure 3 Localized VPLEX Rack Failure

- Protection against scenarios where there is a mismatch between the site where an application is running and the site bias rule setting for distributed devices used by the application <u>during</u> a dual WAN partition event.

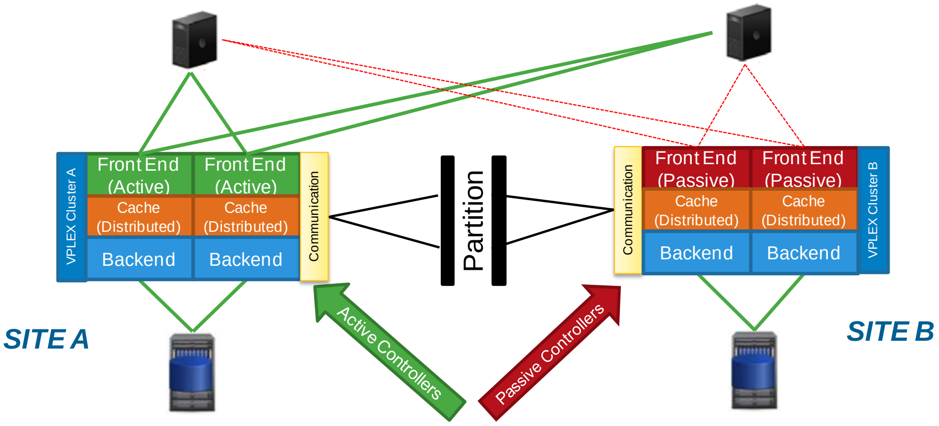


Figure 4 Host at Site B but VPLEX Preference Rule is Site A Wins during WAN Partition

- Protection against a limited set of APD (all paths down/dead) conditions for VMware Metro Storage Clusters.

- Some non-disruptive site migration scenarios benefit from the increased flexibility cross-cluster host connectivity provides. For example, the VPLEX cluster at one of the two sites can be powered down, moved, and brought back online while the host applications remain online. At some point the hosts would need to move, but the dependency on one VPLEX within a site eliminated.

Note: VPLEX Metro suspends host I/O and/or fails a site due to **multiple** SAN, WAN, power, and/or infrastructure related failures. It is likely (though not guaranteed) these multiple failures would also impact hosts, storage, fibre channel switches, and IP networks.

Cross-Cluster SAN Architectures and Failure Scenarios

The failure scenarios that a cross-cluster host connectivity topology can protect against will vary depending on the number of host fibre channel adapter ports, number of WAN and cross-cluster host connectivity channels, and number of SAN fabrics. There are several different types of SAN fabric topologies that can be employed with cross-cluster host connectivity to VPLEX. These topologies can be grouped according to the following characteristics:

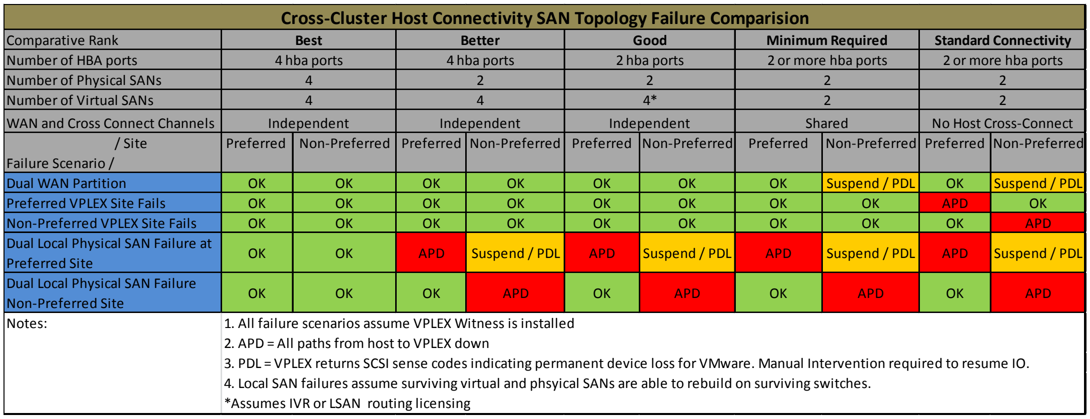
Dual or Quad SAN fabrics

- Dual SAN fabrics are merged across both data centers (sites) and each HBA port is zoned into the local and the remote VPLEX front end ports.
- Quad SAN fabrics use independent HBA ports to access local fabrics, and another set of independent HBA ports for access merged fabrics across data centers (sites).

Shared or Independent WAN channels

- A cross-cluster host configuration is deemed shared when it is routed along the same physical WAN as the VPLEX WAN traffic
- A cross-connect configuration is deemed dedicated when the VPLEX WAN uses a physically separate channel to the cross connect network.

Table 1 below shows high level failure scenarios that the various cross-cluster host connectivity SAN topologies provide protection against. The table indicates the effect to host I/O at the preferred and non-preferred locations based on each topology and type of dual failure that occurs.



Cross-Cluster Host Connectivity SAN Topology Failure Comparision										
Comparative Rank	**Best**		**Better**		**Good**		**Minimum Required**		**Standard Connectivity**	
Number of HBA ports	4 hba ports		4 hba ports		2 hba ports		2 or more hba ports		2 or more hba ports	
Number of Physical SANs	4		2		2		2		2	
Number of Virtual SANs	4		4		4*		2		2	
WAN and Cross Connect Channels	Independent		Independent		Independent		Shared		No Host Cross-Connect	
/ Site Failure Scenario /	Preferred	Non-Preferred	Preferred	Non-Preferred	Preferred	Non-Preferred	Preferred	Non-Preferred	Preferred	Non-Preferred
Dual WAN Partition	OK	OK	OK	OK	OK	OK	OK	Suspend / PDL	OK	Suspend / PDL
Preferred VPLEX Site Fails	OK	OK	OK	OK	OK	OK	OK	OK	APD	OK
Non-Preferred VPLEX Site Fails	OK	OK	OK	OK	OK	OK	OK	OK	OK	APD
Dual Local Physical SAN Failure at Preferred Site	OK	OK	APD	Suspend / PDL	APD	Suspend / PDL	APD	Suspend / PDL	APD	Suspend / PDL
Dual Local Physical SAN Failure Non-Preferred Site	OK	OK	OK	APD	OK	APD	OK	APD	OK	APD
Notes:	1. All failure scenarios assume VPLEX Witness is installed 2. APD = All paths from host to VPLEX down 3. PDL = VPLEX returns SCSI sense codes indicating permanent device loss for VMware. Manual Intervention required to resume IO. 4. Local SAN failures assume surviving virtual and phsyical SANs are able to rebuild on surviving switches. *Assumes IVR or LSAN routing licensing									

Table 1 Cross-Cluster Host Connectivity Failure Comparison

Note: VPLEX has no single point of failure, so each of the scenarios in Table 1 and many non-SAN and non-WAN related failures require two or more components to fail before there is a chance of data becoming unavailable.

As shown in Table 1, the best practice to deploy the cross-cluster connected hosts with additional HBAs (therefore not merging fabrics between sites) and also to use separate dedicated channels so that host cross connect traffic is not shared with the VPLEX WAN.

Chapter 2 Cross-Cluster Host Connectivity

This chapter presents the following topics:

Cross-Cluster Host Connectivity Best Practices

Each of the bulleted best practices items below must be followed to ensure the highest levels of availability with cross-cluster hosts connectivity:

- Use quad (4) fabric and four (4) fibre channel adapter port (single or dual port fibre channel adapters) SAN topologies for cross-cluster host connectivity.

Note: Two (2) fabric and two (2) fibre channel adapter port (2 x single port fibre channel adapter) designs are supported and can provide the equivalent logical connectivity. As shown in Chapter 1, however, these topologies lack the redundancy and fault isolation as quad (4) fabric and 4 adapter port designs.

Note: With dual (2) fabric design the SAN ISL links are also transporting VPLEX fibre channel WAN traffic (for Metro FC). When a site partition event (WAN loss) occurs the SAN switches will segment and host connectivity across clusters will be lost. In this case, if the site preference rules do not match where the application is running, the added cross-cluster connectivity does not improve the availability.

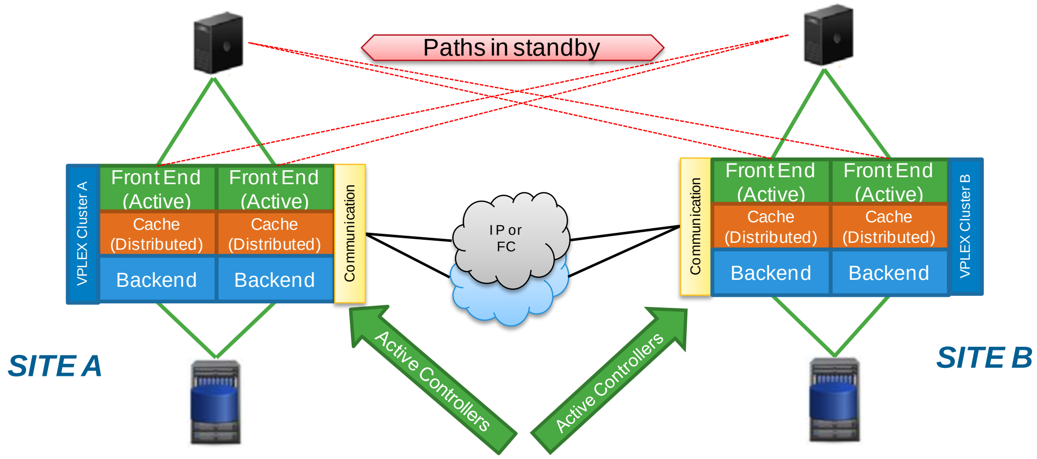


Figure 5 Cross-Cluster Host Connectivity SAN Topology with Stand-by Paths

As shown in Figure 3 hosts using cross-cluster connectivity will have 2 logical paths **per** fabric, for a total of eight (8) logical connections across 4 fabrics (two local fabrics and two stretched fabrics). The red standby paths in Figure 3 are connected to a pair of redundant stretched fabrics that span the Site A and Site B data centers.

- The front-end I/O modules on each director require a minimum of two physical connections one to each fabric
- Each host should have at least one path to an A director and one path to a B director on each fabric at each site for a total of eight (8) logical paths

- Each host should have fabric zoning that provides redundant access to each LUN from a minimum of an A and B director from each fabric at each site
- Dual and Quad engine VPLEX Clusters require host connectivity to span engines on each fabric at each site

Note: For cluster upgrades from a single engine to a dual engine or from a dual to a quad engine you must rebalance the host connectivity across the newly added engines. Adding additional engines and then not connecting host paths to them is of no benefit. The NDU pre-check will flag host connectivity that does not span engines in a multi-engine VPLEX cluster as a configuration issue. When scaling up a single engine cluster to a dual, the ndu pre-check may have passed initially but will fail after the addition of the new engine which is why the host paths must be rebalanced across both engines. Dual to Quad upgrade will not flag an issue provided there were no issues prior to the upgrade. You may choose to rebalance the workload across the new engines or add additional hosts to the pair of new engines.

- PowerPath 5.7+ provides an auto standby feature created specifically for this environment
- Cross-cluster host connectivity applies to specific host OS and multipathing configurations as listed in the VPLEX ESSM only
- Host initiators are zoned to both VPLEX clusters in a Metro
- Host multipathing software can be configured for active path/passive path with active path going to the local VPLEX cluster. Configure the multipathing driver to prefer all local cluster paths over remote cluster paths
- Separate HBA ports should be used for the remote cluster connection to avoid merging of the local and remote fabrics
- Host connectivity at both VPLEX sites follows same rules as single host connectivity
- Supported stretch clusters can be configured using cross-cluster host connectivity (Please refer to VPLEX ESSM on https://support.emc.com)
- Cross-cluster host connectivity is typically limited to a VPLEX cluster separation of no more than 1ms RTT latency subject to the VPLEX ESSM

Note: Latencies up to 5ms can be supported, but may require RPQ if not shown on the latest VPLEX ESSM.

- Cross-cluster host connectivity requires the use of VPLEX Witness

- Cross-cluster host connectivity must be configured using VPLEX distributed devices and consistency groups only. For example, exporting local VPLEX volumes across clusters for the purpose of cross-cluster host connectivity is **not** supported. Local private volumes are not resilient to site failures and will not benefit from cross-cluster host connectivity.
- Cross-cluster host connectivity is supported in a VPLEX Metro environment only
- At least one backend storage array is required at each site with redundant connections to the VPLEX cluster at that site. Back-end arrays are not cross connected to each VPLEX cluster
- All VPLEX consistency groups used in a cross-cluster host connectivity configuration are required to have the auto-resume attribute set to **true**

The unique solution provided by a cross-cluster host connectivity topology requires hosts have fibre channel access to both datacenters. The latency requirements for cross-cluster connectivity can be achieved using extended fabrics or fabrics that span both datacenters. The WAN RTT latency must be within the limits stated for cross-cluster host connectivity in the VPLEX ESSM.

If using PowerPath, you must enable the autostandby feature:

```
#powermt set autostandby=on trigger=prox host=xxx
```

PowerPath will take care of setting to autostandby to those paths associated with the remote/non-preferred VPLEX cluster. PowerPath groups the paths by VPLEX cluster and the one with the lowest minimum path latency is designated as the local/preferred cluster.

VBLOCK and VPLEX Front End Connectivity Rules

Note: All rules in **BOLD** cannot be broken, however Rules in *Italics* can be adjusted depending on customer requirement, but if these are general requirements simply use the suggested rule.

1. **Physical FE connectivity**

   a. **Each VPLEX Director has 4 front end ports. 0, 1, 2 and 3. In all cases even ports connect to fabric A and odd ports to fabric B.**

      i. **For single VBLOCKS connecting to single VPLEX**

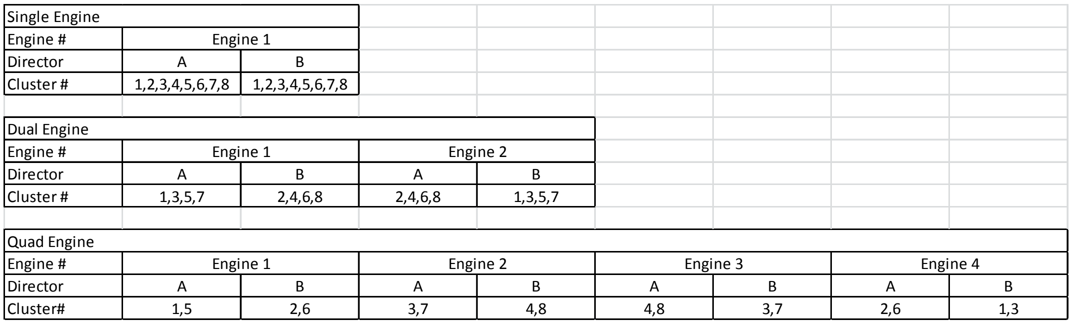
      - **Only ports 0 and 1 will be used on each director. 2 and 3 are reserved.**
      - **Connect even VPLEX front end ports to fabric A and odd to fabric B.**

      ii. **For two VBLOCKS connecting to a single VPLEX**

      - **Ports 0 and 1 will be used for VBLOCK A**
      - **Ports 2 and 3 used for VBLOCK B**
      - **Connect even VPLEX front end ports to fabric A and odd to fabric B.**

2. ESX Cluster Balancing across VPLEX Frontend

All ESX clusters are evenly distributed across the VPLEX front end in the following patterns:



Single Engine		
Engine #	Engine 1	
Director	A	B
Cluster #	1,2,3,4,5,6,7,8	1,2,3,4,5,6,7,8

Dual Engine				
Engine #	Engine 1		Engine 2	
Director	A	B	A	B
Cluster #	1,3,5,7	2,4,6,8	2,4,6,8	1,3,5,7

Quad Engine								
Engine #	Engine 1		Engine 2		Engine 3		Engine 4	
Director	A	B	A	B	A	B	A	B
Cluster#	1,5	2,6	3,7	4,8	4,8	3,7	2,6	1,3

3. Host / ESX Cluster rules

   a. Each ESX cluster must connect to a VPLEX A and a B director.

   b. For dual and quad configs, A and B directors must be picked from different engines (see table above for recommendations)

   c. Minimum directors that an ESX cluster connects to is 2 VPLEX directors.

   d. *Maximum directors that an ESX cluster connects to is 2 VPLEX directors.*

e. Any given ESX cluster connecting to a given VPLEX cluster must use the same VPLEX frontend ports for all UCS blades regardless of host / UCS blade count.

f. Each ESX host should see four paths to the same datastore

   i. 2 across fabric A

   - A VPLEX A Director port 0 (or 2 if second VBLOCK)
   - A VPLEX B Director port 0 (or 2 if second VBLOCK)

   ii. 2 across fabric B

   - The same VPLEX A Director port 1 (or 3 if second VBLOCK)
   - The same VPLEX B Director port 1 (or 3 if second VBLOCK)

4. Pathing policy

   a. Non cross connected configurations may use adaptive pathing policy or round robin policies. See EMC VPLEX Host Multipathing Best Practices Guide at https://support.emc.com for further details.

   b. For cross-cluster host to VPLEX connectivity, fixed pathing should be used for VMware NMP and preferred paths set per Datastore to the local VPLEX path only taking care to alternate and balance over the whole VPLEX front end (i.e. so that all datastores are not all sending IO to a single VPLEX director).

   If using PowerPath/VE, you must enable the autostandby feature:

   #powermt set autostandby=on trigger=prox host=xxx

   PowerPath will take care of setting to autostandby to those paths associated with the remote/non-preferred VPLEX cluster. PowerPath groups the paths by VPLEX cluster and the one with the lowest minimum path latency is designated as the local/preferred cluster.

Chapter 3 ESXi Path Loss Handling

This chapter presents the following topics:

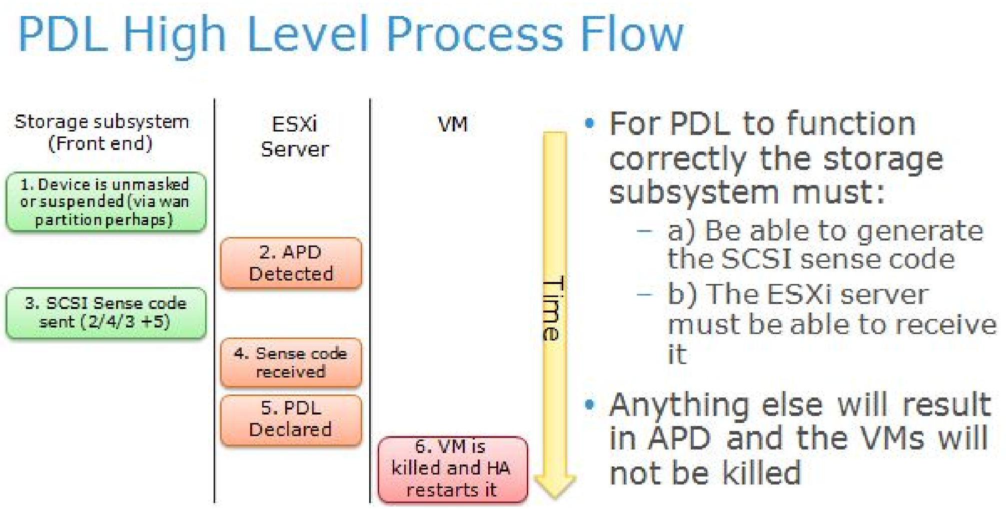
Path loss handling semantics (PDL and APD)

In cross-cluster connected host environments it is critical to set the proper VMware path failure settings. Cross-cluster host connectivity eliminates some, but not all APD conditions. VMware vSphere can recognize two different types of total path failures to an ESXi 5.0 u1 and newer ESXi server. These are known as "All Paths Down" (APD) and "Persistent Device Loss" (PDL). Either of these conditions can be declared by the ESXi server depending on the failure condition.

Persistent device loss (PDL)

A storage device is considered to be in the permanent device loss (PDL) state when it becomes permanently unavailable to your ESXi host. Typically, the PDL condition occurs when a device is unintentionally removed, its unique ID changes, when the device experiences an unrecoverable hardware error, or in the case of a vSphere Metro Storage Cluster WAN partition. When the storage determines that the device is permanently unavailable, it sends SCSI sense codes to the ESXi host. The sense codes allow your host to recognize that the device has failed and register the state of the device as PDL. The sense codes must be received on all paths to the device for the device to be considered permanently lost. If virtual machine files do not all reside on the same datastore and a PDL condition exists on one of the datastores, the virtual machine will not be killed. VMware recommends placing all files for a given virtual machine on a single datastore, ensuring that PDL conditions can be mitigated by vSphere HA.

When a datastore enters a Permanent Device Loss (PDL) state, High Availability (HA) can power off virtual machines and restart them later. A virtual machine is powered off only when issuing I/O to the datastore. Otherwise, it remains active. A virtual machine that is running memory-intensive workloads without issuing I/O to the datastore might remain active in such situations. VMware offers advanced options to regulate the power off and restart operations for virtual machines. The following settings apply only to a PDL condition and not to an APD condition.



PDL High Level Process Flow

Storage subsystem (Front end)

ESXi Server

VM

1. Device is unmasked or suspended (via wan partition perhaps)

2. APD Detected

3. SCSI Sense code sent (2/4/3 +5)

4. Sense code received

5. PDL Declared

6. VM is killed and HA restarts it

Time

- For PDL to function correctly the storage subsystem must:
  - a) Be able to generate the SCSI sense code
  - b) The ESXi server must be able to receive it
- Anything else will result in APD and the VMs will not be killed

Figure 6 Persistent device loss process flow

Advanced settings have been introduced in VMware vSphere 5.0 Update 1 and 5.5 to enable vSphere HA to respond to a PDL condition. The following settings are for the hosts and VMs in the stretched cluster consuming the virtual storage.

Note: PDL response works in conjunction with DRS rules. If the rule is set to “must”, VMware HA will not violate the rule. If the rule is set to “should”, VMware HA will violate it. The DRS rule should be set to “should” to provide availability.

vSphere Storage for vSphere 5.0 U1

disk.terminateVMonPDLDefault set to true

For each host in the cluster, create and edit /etc/vmware/settings with Disk.terminateVMOnPDLDefault=TRUE, then reboot each host.

das.maskCleanShutdownEnabled set to true:

HA Advanced Option. If the option is unset in 5.0U1, a value of false is assumed, whereas in ESXi 5.1 and later, a value of true is assumed. When a virtual machine powers off and its home datastore is not accessible, HA cannot determine whether the virtual machine should be restarted. So, it must make a decision. If this option is set to false, the responding FDM master will assume the virtual machine should not be restarted, while if this option is set to true, the responding FDM will assume the virtual machine should be restarted.

vSphere Storage for vSphere 5.5

disk.terminateVMOnPDLDefault set to default:

Advanced Virtual Machine Option. Default value is FALSE. When TRUE, this parameter powers off the virtual machine if any device that backs up the virtual machine's datastore enters the PDL state. HA will not restart this virtual machine. When set to DEFAULT, VMkernel.Boot.terminateVMOnPDL is used.

VMkernel.Boot.terminateVMOnPDL set to true:

Advanced Vmkernel Option. Default value is FALSE. When set to TRUE, this parameter powers off all virtual machines on the system when storage that they are using enters the PDL state. Setting can be overridden for each virtual machine by disk.terminateVMOnPDLDefault parameter. This parameter can be set only to TRUE or to FALSE. With vSphere web client:

1. Browse to the host in the vSphere Web Client navigator.
2. Click the Manage tab and click Settings.
3. Under System, click Advanced System Settings.
4. In Advanced Settings, select the appropriate item.
5. Click the Edit button to edit the value.
6. Click OK
7. **Reboot the host**

das.maskCleanShutdownEnabled set to default:

HA Advanced Option. This option is set to TRUE by default. It allows VMware HA to restart virtual machines that were powered off while the PDL condition was in progress. When this option is set to true, HA restarts all virtual machines, including those that were intentionally powered off by a user.

disk.AutoremoveOnPDL set to 0:

Advanced Vmkernel option. Default is 1. In the case of a vMSC environment the PDL's are likely temporary because one site has become orphaned from the other, in which case a failover has occurred. If the devices in a PDL state are removed permanently when the failure or configuration error of the vMSC environment is fixed they will not automatically be visible to the hosts again. This will require a manual rescan in order to bring the devices back into service. The whole reason for having a vMSC environment is that it handles these types of things automatically. So you don't want to have to do manual rescans all the time. For this reason the PDL AutoRemove functionality should be disabled on all hosts that are part of a vMSC configuration. Please note that this is recommended for Uniform or Non-Uniform vMSC configurations. Any vMSC configuration that could cause a PDL should have the setting changed.
To disable this feature:

1. Connect to the ESXi host using the console or SSH. For more information, see Using Tech Support Mode in ESXi 4.1 and ESXi 5.x (KB article 1017910).
2. Run this command to disable AutoRemove: esxcli system settings advanced set -o "/Disk/AutoremoveOnPDL" -i 0

Or with vSphere web client:

1. Browse to the host in the vSphere Web Client navigator.
2. Click the Manage tab and click Settings.
3. Under System, click Advanced System Settings.
4. In Advanced Settings, select the appropriate item.
5. Click the Edit button to edit the value.
6. Click OK

Permanent Device Loss

- Remove device from VPLEX, remove or offline a LUN from backend.
- WAN partition, disable wan ports from switch or log in vplex and disable wan ports using vplexcli

All Paths Down

- Remove volume from storage view.

- Remove FC ports from ESXi host, can cause other errors.
- Disable FC ports on switch.

APD Handling

A storage device is considered to be in the all paths down (APD) state when it becomes unavailable to your ESXi host for an unspecified period of time. The reasons for an APD state can be, for example, a failed switch.

In contrast with the permanent device loss (PDL) state, the host treats the APD state as transient and expects the device to be available again.

The host indefinitely continues to retry issued commands in an attempt to reestablish connectivity with the device. If the host's commands fail the retries for a prolonged period of time, the host and its virtual machines might be at risk of having performance problems and potentially becoming unresponsive.

With vSphere 5.1, a default APD handling feature was introduced. When a device enters the APD state, the system immediately turns on a timer and allows your host to continue retrying non-virtual machine commands for a limited time period.

vSphere Storage for vSphere 5.1

Disable Storage APD Handling

The storage all paths down (APD) handling on your ESXi host is enabled by default. When it is enabled, the host continues to retry I/O commands to a storage device in the APD state for a limited time period. When the time period expires, the host stops its retry attempts and terminates any I/O. You can disable the APD handling feature on your host.

If you disable the APD handling, the host will indefinitely continue to retry issued commands in an attempt to reconnect to the APD device. Continuing to retry is the same behavior as in ESXi version 5.0. This behavior might cause virtual machines on the host to exceed their internal I/O timeout and become unresponsive or fail. The host might become disconnected from vCenter Server.

Procedure

1. Browse to the host in the vSphere Web Client navigator.
2. Click the **Manage** tab, and click **Settings**.
3. Under System, click **Advanced System Settings**.
4. Under Advanced System Settings, select the **Misc.APDHandlingEnable** parameter and click the Edit icon.
5. Change the value to 0.

If you disabled the APD handling, you can re-enable it when a device enters the APD state. The internal APD handling feature turns on immediately and the timer starts with the current timeout value for each device in APD.

Note: The host cannot detect PDL conditions and continues to treat the device connectivity problems as APD when a storage device permanently fails in a way that does not return appropriate SCSI sense codes.

Change Timeout Limits for Storage APD

The timeout parameter controls how many seconds the ESXi host will retry non-virtual machine I/O commands to a storage device in an all paths down (APD) state. If needed, you can change the default timeout value.

The timer starts immediately after the device enters the APD state. When the timeout expires, the host marks the APD device as unreachable and fails any pending or new non-virtual machine I/O. Virtual machine I/O will continue to be retried.

The default timeout parameter on your host is 140 seconds. You can increase the value of the timeout if, for example, storage devices connected to your ESXi host take longer than 140 seconds to recover from a connection loss.

Procedure

1. Browse to the host in the vSphere Web Client navigator.
2. Click the **Manage** tab, and click **Settings**.
3. Under System, click **Advanced System Settings**.
4. Under Advanced System Settings, select the **Misc.APDTimeout** parameter and click the Edit icon.
5. Change the default value.

PDL/APD References

http://www.emc.com/collateral/software/white-papers/h11065-vplex-with-vmware-ft-ha.pdf for ESXi 5.0 U1 test scenarios

http://www.vmware.com/files/pdf/techpaper/vSPHR-CS-MTRO-STOR-CLSTR-USLET-102-HI-RES.pdf for ESXi 5.0 U1 vmware vsphere metro storage cluster case study

http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2007545

http://www.emc.com/collateral/software/white-papers/h11065-vplex-with-vmware-ft-ha.pdf

http://www.vmware.com/files/pdf/techpaper/vSPHR-CS-MTRO-STOR-CLSTR-USLET-102-HI-RES.pdf

http://pubs.vmware.com/vsphere-50/topic/com.vmware.ICbase/PDF/vsphere-esxi-vcenter-server-50-storage-guide.pdf

http://pubs.vmware.com/vsphere-55/topic/com.vmware.ICbase/PDF/vsphere-esxi-vcenter-server-55-storage-guide.pdf

https://support.emc.com/search/?text=implementation%20and%20planning%20best%20practices%20for%20emc%20vplex

http://www.boche.net/blog/index.php/2014/07/14/yet-another-blog-post-about-vsphere-ha-and-pdl

http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2033250

http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2059622