EMC® VPLEX™ WITH IBM AIX VIRTUALIZATION AND CLUSTERING

ABSTRACT

This white paper provides best practices, planning, and use cases for implementing EMC VPLEX with IBM AIX virtualization and host cluster technologies.

September 2014

EMC WHITE PAPER

To learn more about how EMC products, services, and solutions can help solve your business and IT challenges, contact your local representative or authorized reseller, visit www.emc.com, or explore and compare products in the EMC Store

Part Number H8138.4

Table of Contents

Executive summary

The EMC® VPLEX™ family removes physical barriers within, across, and between data centers. VPLEX Local provides simplified management and nondisruptive data mobility across heterogeneous arrays. VPLEX Metro provides data access and mobility between two VPLEX clusters within synchronous distances. With a unique scale-up and scale-out architecture, VPLEX's advanced data caching and distributed cache coherency provide workload resiliency, automatic sharing, balancing and failover of storage domains, and enable both local and remote data access with predictable service levels.

Introduction

This white paper explores implementing VPLEX storage in advanced IBM POWER platform AIX configurations and will cover three major sections. First introduced is VPLEX technology. The next major section covers virtualization with PowerVM, particularly Virtual I/O Server (VIOS). The final section discusses implementing local and distributed VPLEX volumes with PowerHA clusters.

The focus of this white paper is on the special considerations of implementing VPLEX storage with advanced features of PowerVM (VIOS and LPM) and PowerHA host clusters with an emphasis on a VPLEX Metro-Plex environment.

This white paper discusses two distinct technologies: PowerHA and PowerVM. While PowerHA and PowerVM are frequently implemented together, there are no dependencies to do so, either by IBM or EMC. PowerVM can be implemented in a nonclustered environment, and PowerHA can be implemented on standalone systems. In addition, if PowerVM is implemented without VIOS, that is the dedicated LPARs have direct access to physical HBAs, VPLEX implementation is the same as with a standalone system. This is described in the *EMC Host Connectivity Guide for IBM AIX*, available on the E-Lab Interoperability Navigator at https://elabnavigator.emc.com.

Audience

This white paper is intended for technology architects, storage administrators, and system administrators who are responsible for architecting, creating, managing IT environments that utilize EMC VPLEX technologies. The white paper assumes the reader is familiar with EMC VPLEX, IBM POWER platform, the AIX operating system, and the PowerVM and PowerHA product suites. It is also assumed that the reader is familiar with the *EMC Host Connectivity Guide for IBM AIX*, available at https://elabnavigator.emc.com.

Terminology

This section provides terminology, acronyms, and abbreviation information.

Table 1. Operational definitions

Term	Definition
Storage volume	LUN or unit of storage presented by the back-end arrays
Metadata volume	System volume that contains metadata about the devices, virtual volumes, and cluster configuration
Extent	All or part of a storage volume
Device	Protection scheme applied to an extent or group of extents
Virtual volume	Unit of storage presented by the VPLEX front-end ports to hosts
Front-end port	Director port connected to host initiators (acts as a target)
Back-end port	Director port connected to storage arrays (acts as an initiator)
Director	A director is the central processing and intelligence of the VPLEX solution. There are redundant (A and B) directors in each VPLEX Engine
Engine	An engine consists of two directors and is the unit of scale for the VPLEX solution
VPLEX cluster	A collection of VPLEX Engines in one rack, using redundant, private Fibre Channel connections as the cluster interconnect
Metro-Plex	A cooperative set of two VPLEX clusters, each serving their own storage domain
PowerHA	IBM host clustering product for AIX hosts on the POWER platform. Sometimes referred to by its older name, HACMP
HACMP	Older name for IBM's AIX host clustering product
PowerVM	Suite of IBM software products allowing host virtualization on the POWER platform
LPAR	Logical Partition acting as an independent server on a IBM POWER platform
Logical Partition Mobility (LPM)	A feature of PowerVM that permits migrating a live system between physical hosts
VIOC	Virtual I/O Client LPAR that uses VIO servers for its I/O, accessing storage virtualized by VIOS
VIOS	Virtual I/O Server, a special LPAR that virtualizes I/O to other VIOC on the same physical machine. Runs a dedicated OS based on AIX version 6
POWER	IBM processor and host family, originally based on the PowerPC CPU, capable of running AIX, pLinux, and i/OS. Successor to both the pSeries (RS/6000) and iSeries (AS/400) lines
rootvg	Root Volume Group. The AIX logical volume manager object that contains the operating system and boot device. It can consist of one or more disks

Table 2. Acronyms and abbreviations

Acronym/Abbreviation	Definition
HBA	Host bus adapter
FE port	Front-end (target ports visible to hosts)
BE port	Back-end (initiator ports visible to storage arrays)
HMC	Hardware Management Console; standalone system used to manage LPARs, VIOS, and VIOC on an IBM POWER server
PVID	Physical Volume Identifier. Unique identifier written onto a disk by AIX
ODM fileset	ODM fileset distributed by EMC for storage attachment support
PowerPath	EMC multipathing software

VPLEX technology

EMC VPLEX virtual storage

EMC VPLEX encapsulates traditional physical storage array devices and applies three layers of logical abstraction to them. The logical relationships of each layer are shown in Figure 1.

Extents are the mechanism used by VPLEX to divide storage volumes. Extents may be all or part of the underlying storage volume. EMC VPLEX aggregates extents and applies RAID protection in the device layer. Devices are constructed using one or more extents, and can be combined into more complex RAID schemes and device structures as desired. At the top layer of the VPLEX storage structures are virtual volumes. Virtual volumes are created from devices and inherit the size of underlying device. Virtual volumes are the elements VPLEX exposes to hosts via its FE ports. Access to virtual volumes is controlled using storage views. Storage views are analogous to Auto-provisioning Groups on EMC Symmetrix® or to storage groups on EMC CLARiiON® and VNX®. They act as logical containers determining host initiator access to VPLEX FE ports and virtual volumes.

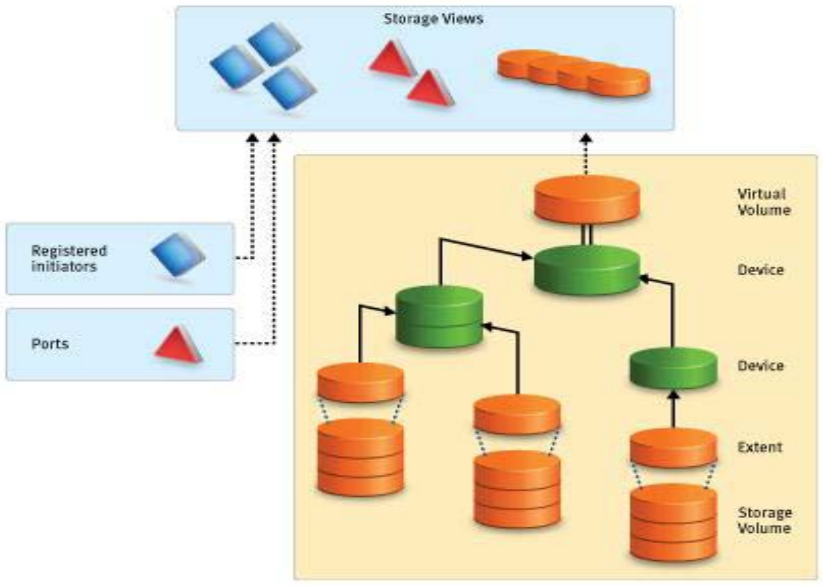


Figure 1. EMC VPLEX logical storage structures

EMC VPLEX architecture

EMC VPLEX represents the next-generation architecture for data mobility and information access. The new architecture is based on EMC's more than 20 years of expertise in designing, implementing, and perfecting enterprise-class intelligent cache and distributed data protection solutions.

As shown in Figure 2, VPLEX is a solution for federating both EMC and non-EMC storage.

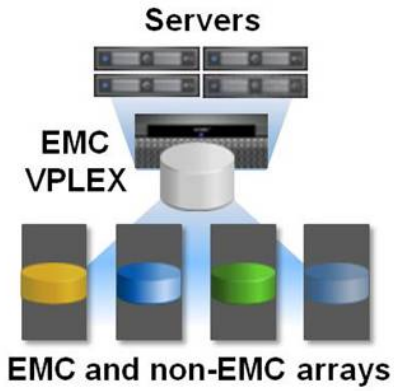


Figure 2. Capability of the EMC VPLEX system

VPLEX resides between the servers and heterogeneous storage assets and introduces a new architecture with unique characteristics:

- Scale-out clustering hardware, which lets customers to start small and grow big with predictable service levels

- Advanced data caching utilizing large-scale SDRAM cache to improve performance and reduce I/O latency and array contention
- Distributed cache coherence for automatic sharing, balancing, and failover of I/O across the cluster
- Consistent view of one or more LUNs across VPLEX clusters separated either by a few feet with a data center or across synchronous distances, enabling new models of high availability and workload relocation

EMC VPLEX family

The following two EMC VPLEX offerings are discussed in this paper:

- **VPLEX Local:** This solution is appropriate for customers that would like federation of homogeneous or heterogeneous storage systems within a data center and for managing data mobility between the physical data storage entities.
- **VPLEX Metro:** This solution is for customers that require concurrent access and data mobility across two locations separated by synchronous distances. The VPLEX Metro offering also includes the unique capability where a remote VPLEX Metro site can present LUNs without the need for physical storage for those LUNs at the remote site.

The EMC VPLEX family architectural capabilities are shown in Figure 3.

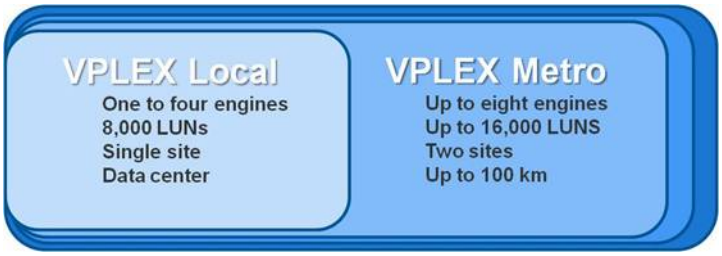


Figure 3. Architectural capabilities

EMC VPLEX clustering architecture

VPLEX uses a unique clustering architecture to help customers break the boundaries of the data center and allow servers at multiple data centers to have concurrent read and write access to shared block storage devices. A VPLEX cluster, shown in Figure 4, can scale up through the addition of more engines, and scale out by connecting multiple clusters to form a VPLEX Metro configuration. In the initial release, a VPLEX Metro system supports up to two clusters, which can be in the same data center or at two different sites within synchronous distances (approximately up to 60 miles or 100 kilometers apart). VPLEX Metro configurations help users to transparently move and share workloads, consolidate data centers, and optimize resource utilization across data centers. In addition, VPLEX clusters provide nondisruptive data mobility, heterogeneous storage management, and improved application availability.

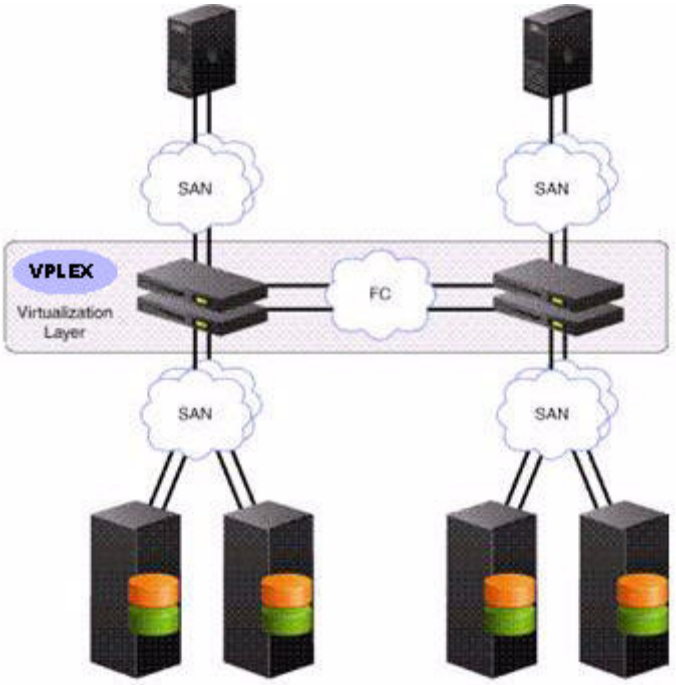


Figure 4. Schematic representation of supported EMC VPLEX cluster configurations

A VPLEX cluster is composed of one, two, or four engines. The engine is responsible for virtualizing the I/O stream, and connects to hosts and storage using Fibre Channel connections as the data transport. A single-engine VPLEX cluster consists of the following major components:

- Two directors, which run the GeoSynchrony™ software
- Dedicated 8 Gb FC front-end, back-end, and local COM and WAN ports
- One Standby Power Supply, which provides backup power to sustain the engine through transient power loss

Each cluster also consists of:

- A management server that provides a GUI and CLI interface to manage a VPLEX cluster
- An EMC standard 40U cabinet to hold all of the equipment of the cluster

Additionally, clusters containing more than one engine also have:

- A pair of Fibre Channel switches used for inter-director communication between various engines
- A pair of Universal Power Supplies that provide backup power for the Fibre Channel switches and allow the system to ride through transient power loss

AIX host virtualization with VPLEX

The PowerVM product allows configuration of multiple virtual machines on a single POWER-architecture server. These virtual machines, called LPARs, can run AIX, pLinux, the IBM i operating system (successor to OS/400), and VIOS. This paper focuses on the AIX operating system; pLinux and i/OS will not be discussed. The POWER hardware architecture, non-VIOS and non-VIOC LPARs, requires the physical I/O adapters (serial, Ethernet, SCSI, Fibre Channel) to be dedicated to a single LPAR. Physical I/O adapters cannot be "split/shared" among LPARs. As an alternative to configuring multiple physical I/O adapters in a dedicated LPAR, the PowerVM VIOS controls physical I/O and virtualizes HBAs, ports, and devices to other LPARs. Storage devices are then virtualized as generic SCSI LUNs, known as a Virtual SCSI Disk Drive over a Virtual SCSI Client Adapter.

In a PowerVM virtualized I/O environment, storage representation is in two stages:

- The VPLEX virtual volumes are presented to the VIOS over the physical Fibre Channel connections. To increase availability, each VIOS should have at least two Fibre Channel connections to the VPLEX FE ports. VPLEX virtual volumes are presented to all VIOS FC HBA ports by including all of the HBA WWPNs in the VPLEX storage view. EMC PowerPath® is installed on the VIOS to provide multipath support for the VPLEX volumes.
- The Virtual I/O Servers then virtualize the VPLEX volumes and present them to the VIOC LPAR as Virtual SCSI Disk Drive devices over a Virtual SCSI Client Adapter. This Virtual SCSI Client Adapter is created when the VIOC LPAR is defined at the HMC. A Virtual SCSI Client Adapter on the VIOC LPAR is mapped to a Virtual SCSI target adapter on the VIOS. The VIOC LPAR sees the VPLEX volumes that are presented from a VIOS.

For increased availability and multiple paths to the client LPARs, two Virtual I/O Servers need to be configured in each physical system. Each VIOS partition would control one or more HBA. Since the same VPLEX volumes are presented through both Virtual I/O Servers, the VIOC LPAR sees the same volume through two different Virtual SCSI Client Adapter paths. The VIOS presents the VPLEX volumes as Virtual SCSI Disk Drive (rather than VPLEX disks). This allows the AIX native MPIO kernel driver on the VIOC LPAR to use the volumes as a multipath device. This is illustrated in Figure 5.

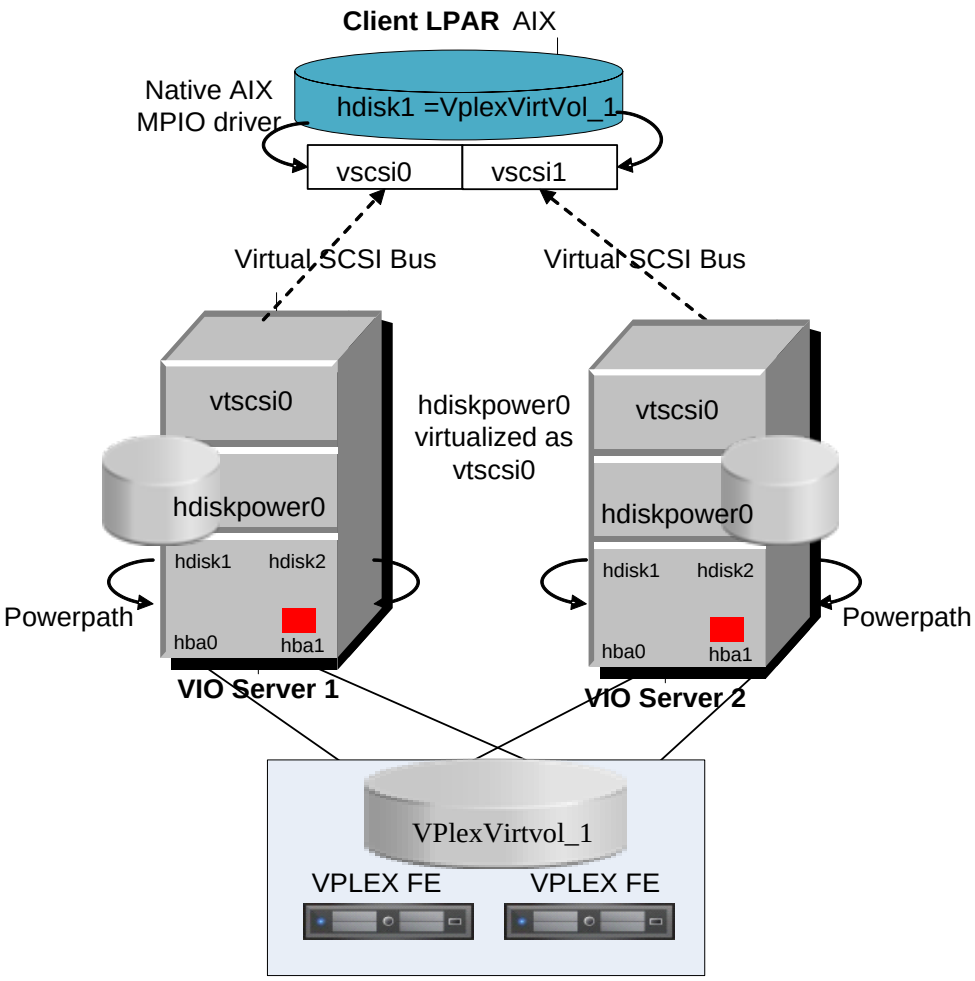


Figure 5. Two VIO servers in a physical system

Installation and configuration

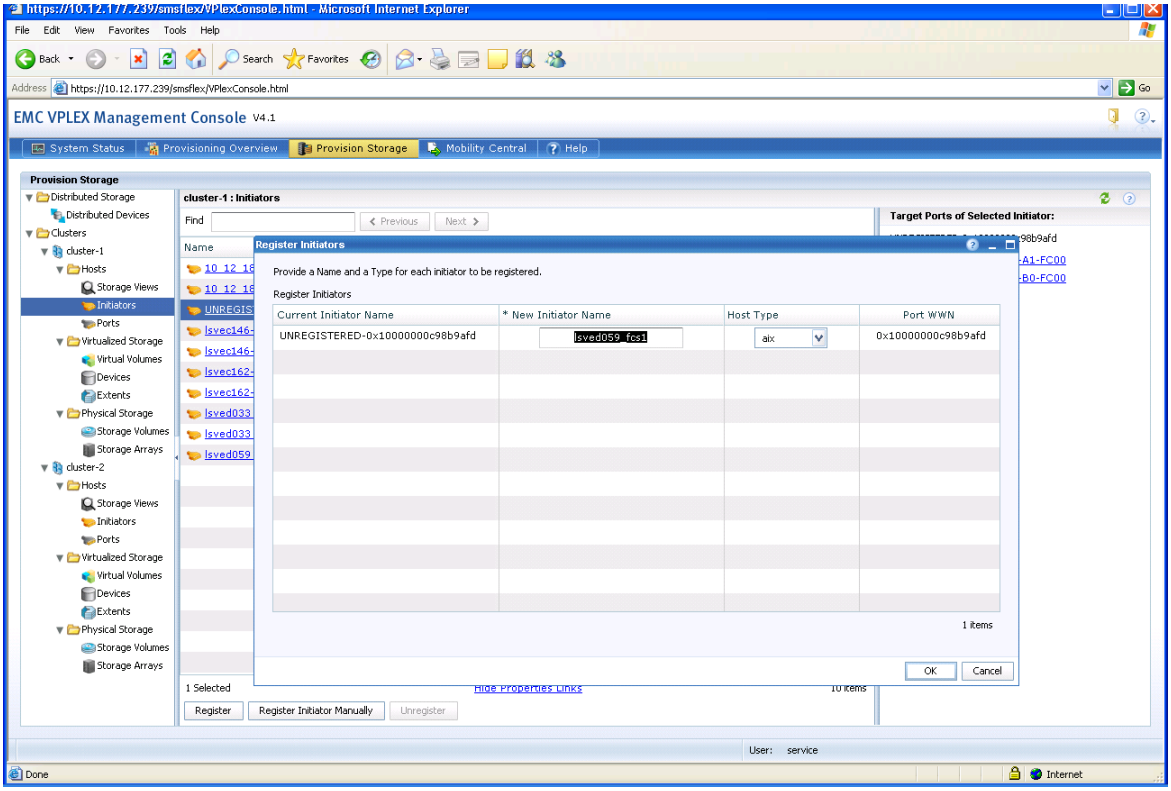
In a VIOS environment, there is no need to install any additional software on the VIOC LPAR to support VPLEX. The VIOS presents the LUNs to the VIOC as generic "Virtual SCSI Disk Drive" devices which are supported by the AIX kernel and AIX native multipath driver MPIO.

The steps required are as follows with detailed descriptions immediately following:

1. Define the required storage.
2. Zone the VIOS HBAs to VPLEX.
3. Define initiators of all physical HBAs as Machine Type AIX.
4. Create a storage view in VPLEX with all physical HBAs used by the VIOS.
5. Install EMC-specific files on VIOS.
6. Configure EMC PowerPath devices on VIOS.
7. Virtualize devices to VIOC LPARs.
8. Configure devices on VIOC LPARs.

Steps 1 through 8 details:

1. Define the required storage. For assistance in back-end array discovery please refer to the *EMC VPLEX 4.0 Installation and Setup Guide* and *Implementation and Planning Best Practices for EMC VPLEX Technical Notes* located on EMC Support Zone at http://support.emc.com (access limited to EMC customers and employees).
2. Zone the VIOS HBAs to the VPLEX. Follow the documentation provided by your Fibre Channel switch manufacturer.
3. Define the initiators.
    a. Using the GUI. Note that host type needs to be set to **aix**.



    b. Using the VPLEX cli (the following is all on one line):

    VPlexcli:/clusters/cluster-1/exports/initiator-ports› initiator-port register -p ‹WWN› --type aix -i ‹initiator name›

    Example:

    VPlexcli:/clusters/cluster-1/exports/initiator-ports› initiator-port register -p 0x10000000c98b9afd --type **aix** -i lsved059_fcs1

4. Create a storage view and include initiators, VPLEX front-end ports, and virtual volumes.
5. Install EMC-specific files on the VIOS. EMC-specific software (PowerPath, ODM filesets) is installed on the VIOS. The setup steps required on a VIOS connected to

VPLEX are nearly identical to the steps required to configure a VIOS with Symmetrix or CLARiiON storage. These steps are documented in more detail in the *EMC Host Connectivity Guide for IBM AIX*, "Virtual I/O Server" chapter.

As is the case for non-virtual I/O environments, the EMC ODM Support Software package is required. The ODM package is implemented in the partition to which devices are physically attached; in this case, the VIOS. Obtain the minimum ODM package version 5.3.0.8 or 6.0.0.3 from the EMC FTP site:

ftp://ftp.emc.com/pub/elab/aix/ODM_DEFINITIONS

Install the EMC ODM Support Package, EMC Invista® AIX, and FCP Support Software filesets on the VIOS. Invista support is required because the VPLEX emulates the Invista virtualization system.

After the EMC ODM filesets are installed, install PowerPath 5.5 or later on the VIOS.

Reboot the VIOS after all installations are complete.

6. Configure devices on the VIOS.

   The VPLEX-specific configuration steps required for the VIOS are the same as those for a standalone AIX version 6 host. This is documented in more detail in the *EMC Host Connectivity Guide for IBM AIX*, in Part 4, Chapter 13, in the "Configuring IBM AIX to recognize VPLEX volumes" section.

   a. Verify that the VIOS recognizes the VPLEX volumes and has properly initialized PowerPath. From the VIOS CLI execute the following command:

      $ lsdev | grep disk

      Output should be similar to the following:

      ```
      hdisk0          Available   SAS Disk Drive
      hdisk1          Available   SAS Disk Drive
      hdisk2          Available   EMC INVISTA FCP Disk
      hdisk3          Available   EMC INVISTA FCP Disk
      hdisk4          Available   EMC INVISTA FCP Disk
      hdisk5          Available   EMC INVISTA FCP Disk
      hdisk6          Available   EMC INVISTA FCP Disk
      hdisk7          Available   EMC INVISTA FCP Disk
      hdisk8          Available   EMC INVISTA FCP Disk
      hdisk9          Available   EMC INVISTA FCP Disk
      hdiskpower0     Available   PowerPath Device
      hdiskpower1     Available   PowerPath Device
      ```

If the hdisk devices do not appear as INVISTA FCP Disk, verify the installation of the EMC ODM filesets, including Invista support.

If the hdisk devices do appear correctly, but no PowerPath devices are listed, do the following at the VIOS CLI:

$ oem_setup_env

powermt config

Rerun the lsdev command to see that the PowerPath devices have been added.

b. Place a PVID on the hdiskpower devices. Type the following command (where ‹x› is the device instance) on the VIOS CLI:

chdev -l hdiskpower‹x› -a pv=yes

Execute the lspv command to verify the hdiskpower devices have a PVID. Output should be similar to the following:

lspv

NAME	PVID	VG	STATUS
hdisk0	00cdcd243a431021	rootvg	active
hdisk1	00cdcd24f86abb9c	None	
hdisk2	none	None	
hdisk3	none	None	
hdisk4	none	None	
hdisk5	none	None	
hdisk6	none	None	
hdisk7	none	None	
hdisk8	none	None	
hdisk9	none	None	
hdiskpower0	00cdcd241cb4b35a	None	
hdiskpower1	00cdcd2412f6b1ea	None	

c. Set the reserve_policy attribute on each hdiskpower device on the VIOS server to no_reserve before mapping it as a virtual device to the VIOC. Use a command similar to the following:

chdev -l hdiskpower‹x› -a reserve_policy=no_reserve

7. Virtualize devices to VIOC LPARs. The basic command to map the Virtual SCSI Disk Drive with the physical hdiskpower in the restricted shell is:

$ mkvdev -vdev TargetDevice -vadapter VirtualSCSIServerAdapter

8. Configure VIOC LPARs to see new volumes. On the VIOC LPARs, execute the cfgmgr command, or reboot. Use the lsdev and lspv commands to verify that the VIOC LPAR can see the virtualized volumes. The PVIDs will be the same for each virtualized volume as they are on the VIOS.

Logical Partition Mobility: Additional considerations

Live Partition Mobility (LPM) is the PowerVM feature that allows a client LPAR to be moved from one physical system to another. This is not a disaster-recovery/high-availability solution. LPM is supported with both VPLEX Local and VPLEX Metro. It is possible to migrate an LPAR from one side of the Metro-Plex to the other. If the Metro-Plex is configured across two data centers, for example, it is possible to migrate an LPAR from a physical server in one data center to a different physical server in the other remote data center.

In order to use LPM, the LPAR must access all its storage (including rootvg) through VIO, and all volumes used by the LPAR must be visible to all physical hosts that can contain the LPAR.

For general considerations on boot volume configuration, see the section "Creating a Fibre Channel boot device on the EMC array" in Chapter 2, "Virtual I/O Server," in the *EMC Host Connectivity Guide for IBM AIX*.

Note the following items when configuring VPLEX for presenting storage when LPM will be used:

- The VPLEX storage view must contain *all* the HBAs from *all* the VIO servers on *all* the physical systems connected to the local VPLEX cluster that could host the client LPAR. This includes both the source and destination LPN physical systems.
- The VPLEX storage view must contain all the virtual volumes used by the LPAR to be migrated. This includes both the volumes in rootvg and volumes used by applications.
- To use LPM across a VPLEX Metro-Plex, all virtual volumes used by the LPAR must be created on distributed devices, and added to the VIO servers' storage views on both VPLEX clusters.

Use case

Customers implement logical partitioning to allow greater flexibility in allocating system resources to business workloads. The addition of a VIO server extends that flexibility to I/O resources such as IP networking and storage networking connectivity (host adapters, switch ports, cabling). VPLEX completes this flexible, virtualized environment by extending this configurability to storage provisioning.

LPM is not designed as a high-availability/disaster-recovery solution. It is more often used for scheduled maintenance and operational tasks. Some cases for moving LPAR virtual host 1 from physical host A to physical host B would be:

- Planned hardware maintenance on physical host A scheduled

- Load balancing between physical hosts A and B
- Physical host A being decommissioned, going off lease, or so on

AIX host clustering with VPLEX

PowerHA is IBM's AIX high-availability host clustering product. The current marketing name is PowerHA, but the older term HACMP is still commonly used in the industry.

As a reminder, PowerHA does not require LPARs, and works equally well on standalone systems. Nor does the use of LPARs require PowerHA. However, the two are often implemented together. PowerHA has been tested successfully with VPLEX in both standalone systems and LPARs using VIO.

General considerations for PowerHA and EMC storage can be found in Chapter 11, "High Availability with Symmetrix/CLARiiON and AIX," of the *EMC Host Connectivity Guide for IBM-AIX*. The section provides a useful overview but much of the information is specific to Symmetrix and CLARiiON and is not applicable to VPLEX.

Note especially the following:

- The emcpowerreset utility is not used with VPLEX.
- General Parallel File System (GPFS) has been tested with VPLEX.
- AIX native MPIO has been tested with VPLEX storage in a VIOS environment. PowerPath is the preferred supported multipath solution for VIO servers with VPLEX.

  Note: Always refer to the *EMC Simple Support Matrix* , *EMC VPLEX and GeoSynchrony* for the most current support statements. ESSMs can be found on the E-Lab Navigator at https://elabnavigator.emc.com.

- The use of disk heartbeating networks (diskhb) on VPLEX volumes virtualized through VIO *has been tested successfully and is supported*.

With most distance-replication solutions, one side is the active side while the other is passive (source/target, master/slave, etc.). In order to make the passive side the active side, a command or commands need to be sent to the replication product. Most distance-replication solutions have an API or a scripting interface for automating this. PowerHA allows customization of failover events to send the necessary commands and has pre-defined "hooks" into the major distance replication products.

Marketers of competitive products emphasize that VPLEX has no API to customize PowerHA. However, because of the VPLEX dual-active architecture, this is not needed. The VPLEX appears like one very large array to all nodes in the PowerHA cluster, local or extended-distance. VPLEX replication is completely transparent to PowerHA. The data is always available at both sides. There is no need to "enable" access at the passive side as part of a failover event.

With VPLEX 4.x, manual intervention on the VPLEX side is required to recover from a failure in a VPLEX WAN link. There are also some configuration considerations, which are documented in this White Paper.

Installation and configuration (General)

In a PowerHA cluster configuration, the VPLEX setup considerations are typical for any Fibre Channel-attached storage in a host-cluster environment.

On the hosts:

1. Standalone hosts or LPARs with non-virtualized I/O:
   - Install the EMC ODM Support Package including Invista AIX support
   - Install PowerPath
2. On VIOS and VIOC LPARs, use the procedure described above:
   - Install the EMC ODM Support Package including Invista AIX support on the VIO server
   - Install PowerPath on the VIO server
3. On all systems running PowerPath (VIO servers or hosts with non-virtualized I/O):
   - Change the hdiskpower devices to turn off the reserve_policy
   - Use the command chdev -l hdiskpower(x) -a reserve_policy=no_reserve

Disabling the reserve_policy attribute on the hdiskpower device is required on all systems running PowerPath, not just VIO servers.

On the VPLEX:

- All HBAs in all nodes of the cluster must be included in the storage view of the shared storage.
- The HBAs must be defined with the IBM AIX host type in the VPLEX GUI or CLI UI. This applies to regular hosts as well as VIO servers, if used.
- The host cluster must be configured to minimize the risk of split-brain, including the use of redundant heartbeat networks. Both IP and non-IP networks should be used. VPLEX virtual volumes can be used for disk heartbeat networks.

All HBAs from all nodes in the PowerHA cluster must be included in the local VPLEX storage view. If VIO servers are used, all HBAs from all VIO servers virtualizing VPLEX volumes for any client LPAR nodes in the PowerHA cluster must be included in the local storage view.

Local and geographically extended clusters

In the following, the terms "local" and "extended" are in the context of PowerHA, not VPLEX. A local cluster is a cluster where all the nodes are in the same physical location and have access to all the same resources, most notably storage. This is the default configuration. PowerHA has optional additional support for geographically

distributed clusters, where the nodes are in different locations, and do not necessarily have direct physical access to the same storage resources. The marketing terms used for this include HAGEO and XD. The geographic cluster has an additional management object, the site, which is a group of one or more cluster nodes. Geographical separation of cluster nodes implies the need of a remote replication solution. PowerHA has software hooks for a number of remote replication products. These include IBM's software remote mirroring over IP, IBM proprietary array-to-array replication, and EMC SRDF®.

In a PowerHA local cluster configuration, the VPLEX setup considerations are typical for any Fibre Channel-attached storage in a host-cluster environment:

- All HBAs in all nodes of the cluster must be included in the storage view of the shared storage.
- The HBAs must be defined with the IBM AIX host type in the VPLEX UI. This applies to regular hosts as well as VIO servers, if used.
- The host cluster must be configured to minimize the risk of split-brain, including the use of redundant heartbeat networks. Both IP and non-IP networks should be used. VPLEX virtual volumes can be used for disk heartbeat networks.

Installation and configuration (Extended clusters)

In addition to all the setup steps and considerations listed above, the following are required to implement Metro-Plex volumes in a geographically distributed PowerHA cluster

- Virtual volumes must be created on distributed devices.
- The host HBAs at each site must be registered as initiators on the VPLEX cluster at that site.
- Storage views must be created on the VPLEX cluster at each site, containing the host initiators located at that site.
- The distributed volumes must be included in the corresponding storage views on each VPLEX cluster.

First storage view

Adding distributed volumes to the first storage view is done in the usual way, with the GUI, as shown in Figure 6.

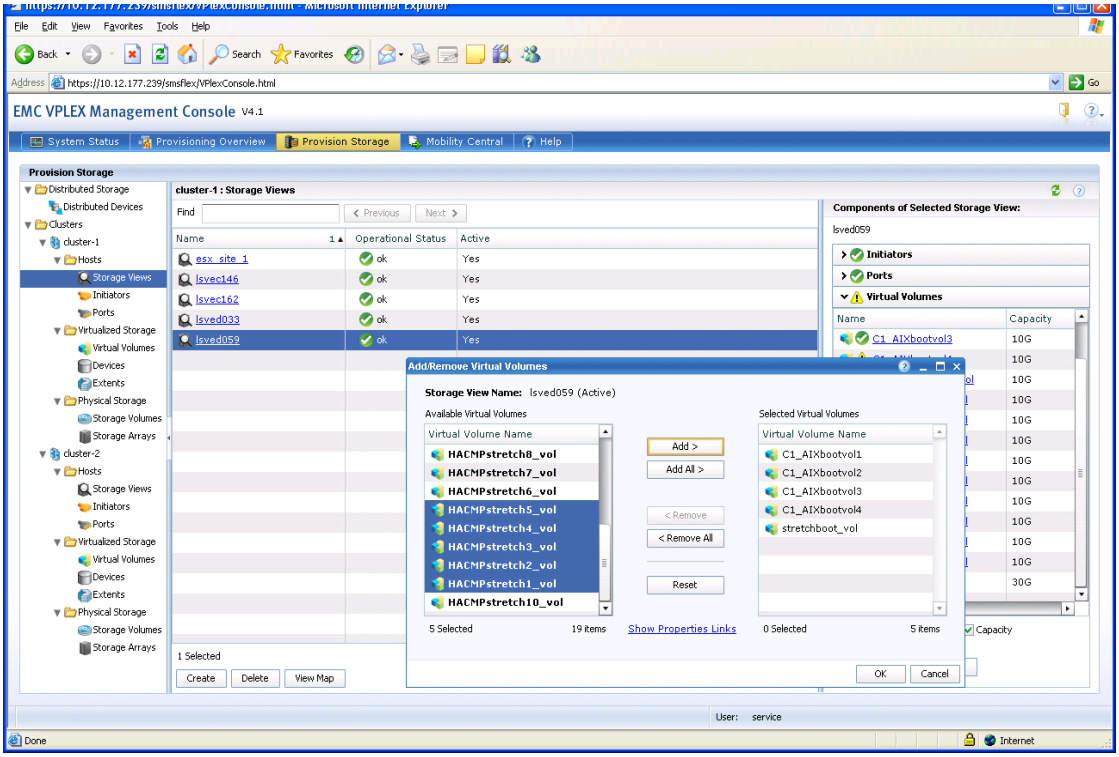


Figure 6. Adding distributed volumes to a storage view

Or it can be done with the CLI:

VPlexcli:/clusters/cluster-1/exports› storage-view addvirtualvolume --view ‹view name› --virtual-volumes=‹volume name›

Example:

VPlexcli:/clusters/cluster-1/exports› storage-view addvirtualvolume --view lsved059 --virtual-volumes=/clusters/cluster-1/virtual-volumes/HACMPstretch1_vol

Second storage view

In VPLEX version 4.1, the GUI *cannot* be used when adding the distributed volumes to the storage view on the VPLEX cluster at the second site. The CLI must be used. In addition, the force option must be used. By default, VPLEX does not allow a virtual volume to be in more than one storage view at a time.

VPlexcli:/clusters/cluster-2/exports› storage-view addvirtualvolume --view ‹view name› --virtual-volumes=‹volume name› --force

Example:

VPlexcli:/clusters/cluster-2/exports› storage-view addvirtualvolume --view lsved060 --virtual-volumes=/clusters/cluster-1/virtual-volumes/HACMPstretch1_vol **--force**

Executing this command will generate a warning message similar to the following. This is expected.

CAUTION: Exporting a volume through two or more views is a valid configuration in only very specific circumstances. Ensure that initiator ports lsved060_fcs0, lsved059_fcs1, lsved060_fcs1, and lsved059_fcs0 are participating in host cluster and should all have access to volume 'HACMPstretch1_vol'. Proceed.

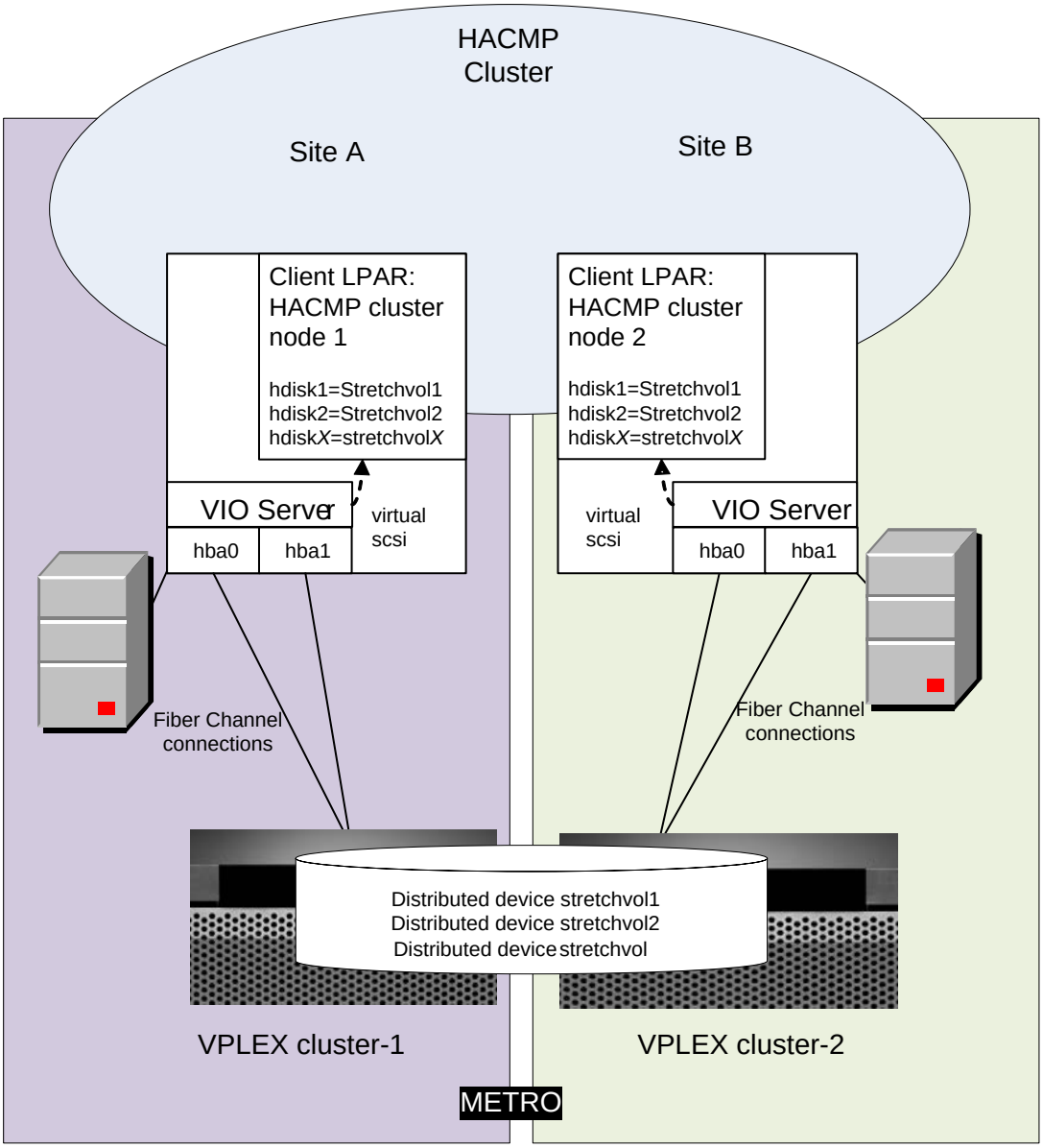


Figure 7. PowerHA Implementation on Metro-Plex

Metro-Plex and PowerHA geographic clusters

In a stretched volume, AIX volume identifier information (Physical Volume Identifier, Volume Group Descriptor Area) is replicated between sites along with the user data. From the AIX point of view, the stretched volume is the identical volume that happens to be available through different FC paths. As a result, a Metro-Plex appears to all AIX

hosts as one unified array that is physically connected to all sites. This simplifies implementation since no replication resources or disk failover control commands need to be defined within AIX. It is simply a disk resource like any other. If the hosts fail at one site, the failover to the other site will complete without user intervention.

Note that this is true only for a host failure. This transparency does not extend to failures within a VPLEX or in the communication link between sites in a Metro-Plex. By design, a failure will suspend I/O to one side of the Metro-Plex until the administrator intervenes to unsuspend I/O based on the nature of the failure. The detach rules, that is, the determination of which side gets suspended, are determined individually per distributed volume.

PowerHA can be customized to send commands to an array to perform the necessary reconfiguration after a failure. The current release of VPLEX does not have a scriptable CLI that is accessible from outside the management server, comparable to SYMCLI for the Symmetrix family. Therefore, PowerHA cannot be configured to send array management commands required to recover from site failures or inter-site link failures. Manual intervention will be required to recover from VPLEX site failures and inter-site communication failures. Users will need to configure site detach rules to meet the requirements of their installations.

CAUTION: Use caution when overriding VPLEX's default behavior in a partitioned Metro-Plex situation. If you are re-enabling I/O to the losing Metro-Plex side (such as with resume link-down or declare-winner), it is critical to ensure that no hosts perform I/O on the winning Metro-Plex side. If this occurs, users **can** and **will** experience data loss or corruption once the Metro-Plex inter-site link is restored. EMC best practices recommend configuring Metro-Plex detach rules to minimize the need for manual VPLEX intervention.

In addition, it is a best practice to configure the VPLEX in parallel with the PowerHA behavior, so that the same VPLEX and PowerHA sides will win in the event of a partitioned cluster.

Assuming a configuration where site-1 is the production site and site-2 is a warm standby/disaster recovery site, the site detach rules will have the following effects:

- **Site-1-detaches:**

  Disadvantage:

  Will require manual intervention during a failover from site-1 to site-2 to make the stretched volumes available at site-2.

  Advantage:

  Inter-site outages and site-2 VPLEX outages will not affect site-1 production operations.

- **Site-2-detaches:**

  Advantage:

  Failovers from site-1 to site-2 will complete without intervention since the stretched volumes will be available at site-2.

  Disadvantage:

  Communication and site-2 VPLEX outages (planned or not) can interfere with production operations.

The specific results are shown in the following tables.

Table 3. Stretched volumes configured with a cluster-1-detaches ruleset (VPLEX site-1 is the winning site)

Site-1 host	VPLEX cluster-1	Site-2 host	VPLEX cluster-2	Result
Failure	Up	Up	Up	Normal failover
Up	Failure	Up	Up	**Manual intervention required** [a]
Failure	Failure	Up	Up	**Manual intervention required** [b] [f]
Up	Up	Failure	Up	Site-1 continues normal operation
Up	Up	Up	Failure	Site-1 continues normal operation
Up	Up	Failure	Failure	Site-1 continues normal operation

Inter-Site Link activity	PowerHA Winner	VPLEX Winner	Result
Failure	site-1	site-1	No intervention necessary; distributed devices are in winner-running state.
Restoration	site-1	site-1	No intervention necessary; distributed devices return to healthy state.

Table 4. Stretched volumes configured with a cluster-2-detaches ruleset (VPLEX site-2 is the winning site)

Site-1 host	VPLEX cluster-1	Site-2 host	VPLEX cluster-2	Result
Failure	Up	Up	Up	Normal failover
Up	Failure	Up	Up	Failover [c]
Failure	Failure	Up	Up	Normal failover
Up	Up	Failure	Up	Site-1 continues normal operation
Up	Up	Up	Failure	**Manual intervention required** [a]
Up	Up	Failure	Failure	**Manual intervention required** [a]

Table 5. Stretched volumes configured with a cluster-2-detaches ruleset (VPLEX site-2 is the winning site) cont.

Inter-Site Link activity	PowerHA Winner	VPLEX Winner	Result
Failure	site-1	site-2	**Manual intervention required** [a] [e]
Restoration	site-1	site-2	**Manual intervention required** [d] [e]

Notes:

a. The VPLEX device resume link-down command needs to be executed on all stretched devices before site-1 I/O can continue.
b. A custom error notification was required to initiate the failover based on the disk errors caused by this scenario.
c. Loss of path was not detected for 10 minutes. This is a PowerHA issue, not a VPLEX issue.
d. The declare-winner command needs to be executed on all stretched devices to allow site-1 I/O to resume
e. This difference in behavior between failures depending on the ruleset is because failures will not necessarily trigger the same results in PowerHA and VPLEX. If the VPLEX Fibre Channel ISL goes down but the PowerHA cluster does not fail over, VPLEX site-2 wins but the site-1 hosts will still be running and trying to perform I/O. One scenario would be if there is a fibre failure between sites but the IP connectivity stays up.
f. A custom error notification was required to initiate the failover based on the disk errors caused by this scenario.

Front-End Cross-Connect (Cross-Cluster Connect)

An additional layer of resiliency in Metro HA environments can be added by using VPLEX Metro HA Cross-Cluster Connect. It can be deployed when two sites are within campus distance of each other (up to 1ms round trip latency). A VPLEX Metro distributed volume can then be deployed across the two sites using a cross connect front end configuration. In a cross connect front end environment, the front-end fabrics are interconnected and zoned so that the hosts (or VIO Servers) can access both VPLEX clusters simultaneously. The host HBAs are then defined in storage views in both sides of the Metro-Plex.

Further information and implementation details can be found in the EMC TechBook *EMC VPLEX Metro Witness Technology and High Availability*, part number H7113, available at http://support.emc.com. This feature is described in the "Combining VPLEX High Availability and VPLEX Witness" chapter. The configuration examples provided in that document are for VMWare HA, but the principles are equally applicable, and supported, with PowerHA host clusters.

Additional notes: Non-VPLEX specific

These issues are not specific to VPLEX but should be considered when planning any implementation of PowerHA on FC-attached storage. Customers with AIX and PowerHA experience will probably already be aware of some or all of these. This information is on non-EMC products and such is provided for reference only.

SAN boot and PowerHA

PowerHA is tightly coupled with the AIX operating system and depends on AIX facilities such as the errlogger daemon for event detection. If a system is booted from a SAN volume and has its rootvg on a SAN volume, the loss of access to that volume will result in abnormal behavior. It will still respond to IP network commands but will be unable to execute any command requiring any disk I/O – including reading and writing its own error log. Loss of the rootvg will not result in the node shutting itself or initiating any PowerHA events. The other node will not mark it as down and force a takeover since the failed node still responds on the heartbeat network(s). The net result is application and data unavailability. This is a known and documented limitation of SAN boot.

There is an additional issue when SAN booting from virtual SCSI (vSCSI) disks, that is, from disks presented from a VIO server to a VIO client. The following is from the AIX53SANBoot Wiki on the IBM developerWorks website:

"It seems imprudent to boot from SAN (or allocate paging space on SAN for) any cluster node unless the cluster's shared volume groups are protected by disk reservation locks. (In this context, booting from vSCSI disks mapped to LUNs is equivalent to booting from SAN.)

Disk and path failure detection on a VIO client host

Single path failures will be logged at the VIO server and not propagated up to the VIO client hosts. That is, if one path to the storage is lost, the error will be logged on the VIO server but not on the VIO client. VIO servers cannot be part of a PowerHA cluster, and must be monitored separately.

PowerHA has a facility for initiating a failover when it loses access to a disk, but by default it is only triggered by a loss of quorum, logged as LVM_SA_QUORCLOSE. If the volume groups are not mirrored, this specific error may not be logged even though numerous other I/O errors are logged as a result of the lost access. Possible workarounds include mirroring a small logical volume in each volume group (for example, the jfslog if file systems are used) or creating custom PowerHA error notify methods to initiate failover rather than notification on disk I/O errors.

Conclusion

The IBM PowerVM product suite, with its Logical Partitions, Virtual I/O, and Live Partition Mobility, provides the ability to flexibly distribute AIX computing resources across physical hardware depending on changing business needs. Implementing VPLEX in this environment extends that flexibility into the storage domain.

PowerHA enables the grouping of resources for high availability — even over a geographically distributed area. VPLEX's GeoSynchrony makes data transparently available at both locations, simplifying resource failover between both sites. Subject to the configuration caveats outlined in this paper, implementing VPLEX with PowerHA leverages the synergies of both products, adding value to the customer's information resources.

References

The following in available on EMC Online support at http://support.emc.com and https://elabnavigator.emc.com:

- *EMC Host Connectivity Guide for IBM AIX*

The following IBM Redbooks provide more information:

- *PowerVM Virtualization on IBM System p: Introduction and Configuration* (SG24-7940-03)
- *Virtualization and Clustering Best Practices Using IBM System p Servers* (SG24-7349-00)