EMC® VPLEX™

with GeoSynchrony™ 5.0 and Point Releases

Product Guide

P/N 300-012-307
REV A03

EMC Corporation
Corporate Headquarters:
Hopkinton, MA 01748-9103
1-508-435-1000
www.EMC.com

Contents

Figures

Tables

Preface

As part of an effort to improve and enhance the performance and capabilities of its product line, EMC® from time to time releases revisions of its hardware and software. Therefore, some functions described in this document may not be supported by all revisions of the software or hardware currently in use. Your product release notes provide the most up-to-date information on product features.

If a product does not function properly or does not function as described in this document, please contact your EMC representative.

About this guide

This document provides a high level description of the VPLEX™ product and GeoSynchrony™ 5.0 features.

Audience

This document is part of the VPLEX system documentation set and introduces the VPLEX Product and its features. The document provides information for customers and prospective customers to understand VPLEX and how it supports their data storage strategies.

Table 1 Document Change History

Revision	Changes since previous revision
Revision A03	Chapter 1 — Clarified description of failure handling with VPLEX Witness. Chapter 2 — Included additional information on power subsystem, redundancy, and power failure handling. Chapter 3 — Clarified descriptions of consistency groups. Provided more information about cache vaulting. Added section on recovery after power failure. Chapter 4 — Clarified description of failure handling with VPLEX Witness. Clarified description of power failure data protection. Defined DLFM.
Revision A02	Chapter 3 — Corrected information about data mobility. Chapter 4 — Clarified information about data mobility.

Related documentation

Related documentation (available on EMC Powerlink®) includes:

- *EMC VPLEX with GeoSynchrony 5.0 and Point Releases Release Notes*
- *Implementation and Planning Best Practices for EMC VPLEX Technical Notes*
- *EMC VPLEX Security Configuration Guide*
- *EMC VPLEX Site Preparation Guide*
- *EMC Best Practices Guide for AC Power Connections in Two-PDP Bays*
- *EMC VPLEX Hardware Installation Guide*
- *EMC VPLEX with GeoSynchrony 5.0 Configuration Worksheet*

- *EMC VPLEX with GeoSynchrony 5.0 Configuration Guide*
- *EMC VPLEX with GeoSynchrony 5.0 and Point Releases CLI Guide*
- VPLEX Procedure Generator

The VPLEX GUI also provides online help.

For additional information on all VPLEX publications, contact the EMC Sales Representative or refer to the EMC Powerlink website at:

http://powerlink.EMC.com

Conventions used in this guide

EMC uses the following conventions for special notices.

Note: A note presents information that is important, but not hazard related.

Typographical conventions

EMC uses the following type style conventions in this document:

Normal	In running text: • Interface elements (for example button names, dialog box names) outside of procedures • Items that user selects outside of procedures • Names of resources, attributes, pools, Boolean expressions, buttons, DQL statements, keywords, clauses, environment variables, filenames, functions, menu names, utilities • URLs, pathnames, filenames, directory names, computer names, links, groups, service keys, file systems, environment variables, notifications
Bold	In procedures: • Names of dialog boxes, buttons, icons, menus, fields • Selections from the user interface, including menu items and field entries • Key names • Window names In running text: • Command names, daemons, options, programs, processes, notifications, system calls, man pages, services, applications, utilities, kernels
Italic	Used for: • Full publications titles referenced in text • Unique word usage in text
`Courier`	Used for: • System output • Filenames, • Complete paths • Command-line entries • URLs
`Courier bold`	Used for: • User entry • Options in command-line syntax
`Courier italic`	Used for: • Arguments used in examples of command-line syntax • Variables in examples of screen or file output • Variables in pathnames
< >	Angle brackets enclose parameter or variable values supplied by the user
[ ]	Square brackets enclose optional values
|	Vertical bar indicates alternate selections - the bar means “or”

{ }	Braces indicate content that you must specify (that is, x or y or z)
...	Ellipses indicate nonessential information omitted from the example

Where to get help

EMC support, product, and licensing information can be obtained as follows.

Product information — For documentation, release notes, software updates, or for information about EMC products, licensing, and service, go to the EMC Powerlink website (registration required) at:

http://Powerlink.EMC.com

Technical support — For technical support, go to EMC Powerlink. To open a case, you must be a customer. Information about your site configuration and the circumstances under which the problem occurred is required.

Your comments

Your suggestions will help us continue to improve the accuracy, organization, and overall quality of the user publications. Please send your opinion of this document to:

techpubcomments@EMC.com

If you have issues, comments or questions about specific information or procedures, please include the title and, if available, the part number, the revision (for example, A01), the page numbers, and any other details that will help us locate the subject you are addressing.

1 Introducing VPLEX

This chapter provides an overview of the EMC VPLEX product family and covers several key features of the VPLEX system. Topics include:

VPLEX overview

EMC VPLEX is unique virtual storage technology that federates data located on multiple storage systems – EMC and non-EMC – allowing the storage resources in multiple data centers to be pooled together and accessed anywhere. When combined with virtual servers, it is a critical enabler of private and hybrid cloud computing and the delivery of IT as a flexible, efficient, and reliable resilient service.

The VPLEX family addresses three primary IT needs:

- Mobility: The ability to move applications and data across different storage installations, whether within the same data center, across a campus, within a geographical region—and now, with VPLEX Geo, across even greater distances.
- Availability: The ability to create high-availability storage infrastructure across these same varied geographies with unmatched resiliency.
- Collaboration: The ability to provide efficient real-time data collaboration over distance for such big data applications as video, geographic/ oceanographic research, and others.

All of this can be done within or across data centers, located synchronous or asynchronous distances apart, in a heterogeneous environment.

The VPLEX family brings many unique innovations and advantages:

- VPLEX technology enables new models of application and data mobility, leveraging distributed/federated virtual storage. For example, VPLEX is specifically optimized for virtual server platforms (e.g., VMware ESX, Hyper-V, Oracle Virtual Machine, AIX VIOS) and can streamline and even accelerate transparent workload relocation over distances, which includes moving virtual machines over distances.
- With its unique, highly available, scale-out clustered architecture, VPLEX can be configured with one, two, or four engines—and engines can be added to a VPLEX cluster non-disruptively. All virtual volumes presented by VPLEX are always accessible from every engine in a VPLEX cluster. Similarly, all physical storage connected to VPLEX is accessible from every engine in the VPLEX cluster. Combined, this scale-out architecture uniquely ensures maximum availability, fault tolerance, and scalable performance.
- Advanced data collaboration, through AccessAnywhere, provides cache-consistent active-active access to data across two VPLEX clusters over synchronous distances with VPLEX Metro and asynchronous distances with VPLEX Geo.

Mobility

Application and data mobility provides the movement of virtual machines (VM) without downtime.

Storage administrators have the ability to automatically balance loads through VPLEX, using storage and compute resources from either cluster's location. When combined with server virtualization, VPLEX can transparently move and relocate virtual machines and their corresponding applications and data over distance. This provides a unique capability to relocate, share, and balance infrastructure resources between sites, which can be within a campus or between data centers, up to 5 ms round trip time (RTT) latency apart with VPLEX Metro, or further apart (50ms RTT) across asynchronous distances with VPLEX Geo.

Availability

By providing redundancy, flexibility, and awareness (through VPLEX Witness), GeoSynchrony 5.0 supports small recovery time objective (RTO) and recovery point objective (RPO). Chapter 4, "System Integrity and Resiliency" provides details on the redundancies built into the VPLEX Metro and VPLEX Geo configurations, and describes how these configurations handle failures to reduce recovery point objective. All of these features allow the highest resiliency possible in the case of an outage like the one shown in Figure 1.

Figure 1, shows a VPLEX Metro configuration where storage has become unavailable at one of the cluster sites. Because data is being mirrored using the GeoSynchrony AccessAnywhere feature, both sites access the identical copies of the same data. At the point of failure, the applications can continue to function using the back-end storage at the unaffected site. This is just one example of the resiliency provided in the VPLEX architecture. VPLEX also supports uninterrupted access even in the event of port, engine, director, cluster, or inter-cluster link failures as described in Chapter 2, "VPLEX Hardware Overview" and Chapter 4, "System Integrity and Resiliency."

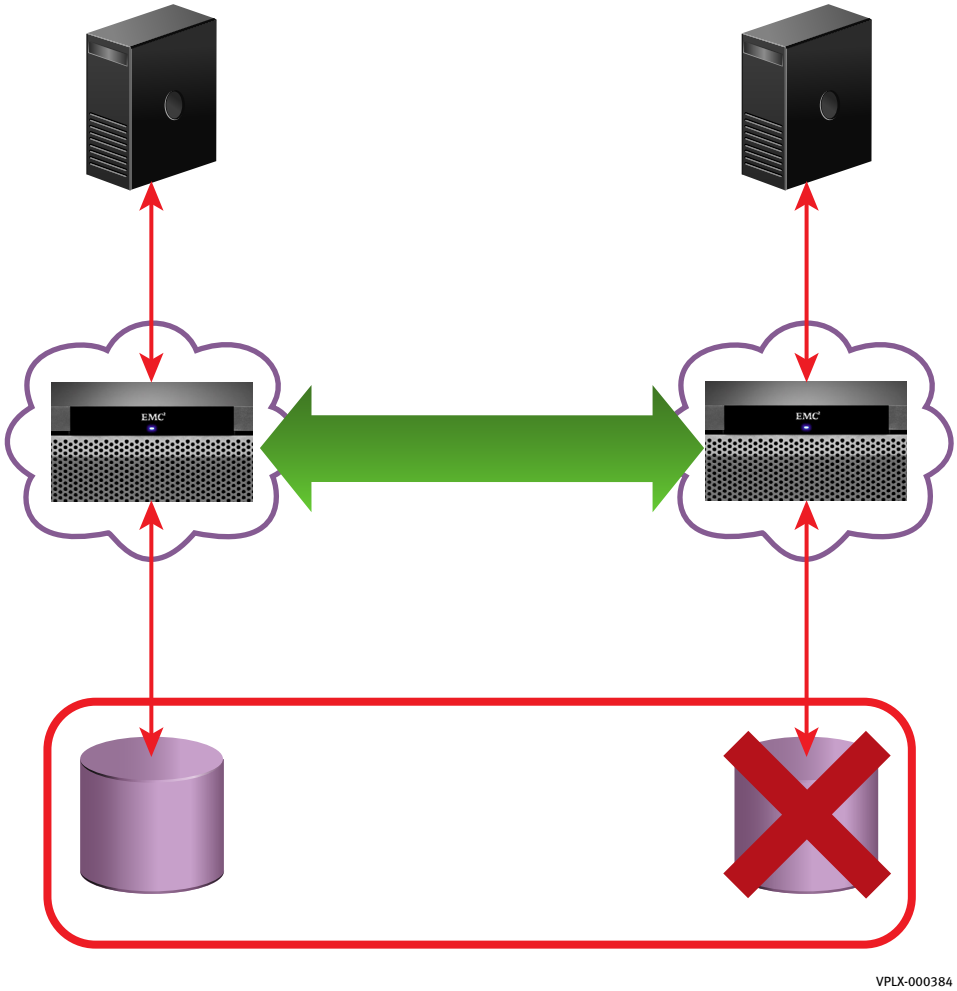

Figure 1 High availability infrastructure example

Figure 1, shows a VPLEX Metro configuration where storage has become unavailable at one of the cluster sites. Because data is being mirrored using the GeoSynchrony AccessAnywhere feature, both sites access the identical copies of the same data. At the point of failure, the applications can continue to function using the back-end storage at the unaffected site. This is just one example of the resiliency provided in the VPLEX architecture. VPLEX Metro also supports uninterrupted access even in the event of port, engine, director, cluster, or inter-cluster link failures as described in Chapter 2, "VPLEX Hardware Overview" and Chapter 4, "System Integrity and Resiliency."

Note: Behavior in a VPLEX Geo configuration performing active/active writes differs in its handling of access during these link failures. Chapter 4, "System Integrity and Resiliency" for a description of how VPLEX Geo handles cluster and inter-cluster failures.

Collaboration

Collaboration increases utilization of passive *data recovery* assets and provides simultaneous access to data. Figure 2 shows an example of how you can use distributed data collaboration.

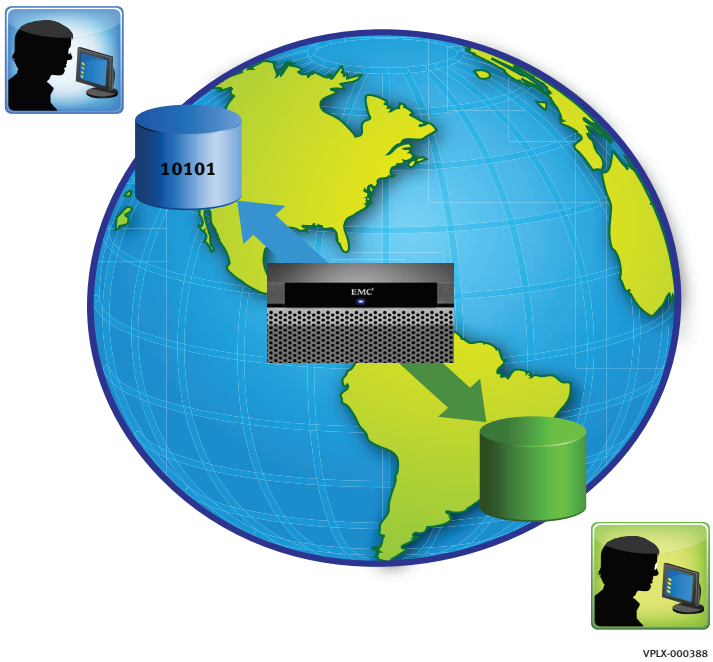

Figure 2 Distributed data collaboration example

When a workforce has multiple users at different sites who need to work on the same data and maintain consistency in the dataset, the distributed data collaboration scenario supported by VPLEX provides a solution. A common example would be a geographically separated company where co-development of software requires collaborative workflows among engineering, graphic arts, video, educational programs, design, research, and so forth.

With traditional solutions, when you try to build collaboration across distance, you normally have to save the entire file at one location and then send it to another site using FTP. This is slow, can incur heavy bandwidth costs for large files (or even small files that move regularly) and it negatively impacts productivity because the other sites can sit idle while they wait to receive the latest data from another site. If teams decide to do their own work independent of each other, then the dataset quickly becomes inconsistent, as multiple people are working on it at the same time and are unaware of each other's most recent changes. Bringing all of the changes together in the end is time-consuming, costly, and grows more complicated as your data-set gets larger.

VPLEX provides a scalable solution for collaboration.

VPLEX product offerings

VPLEX first meets high-availability and data mobility requirements and then scales up to the I/O throughput you require for the front-end applications and back-end storage.

The three available VPLEX product offerings are:

- VPLEX Local
- VPLEX Metro
- VPLEX Geo[1]

Figure 3 shows an example of each product offering.

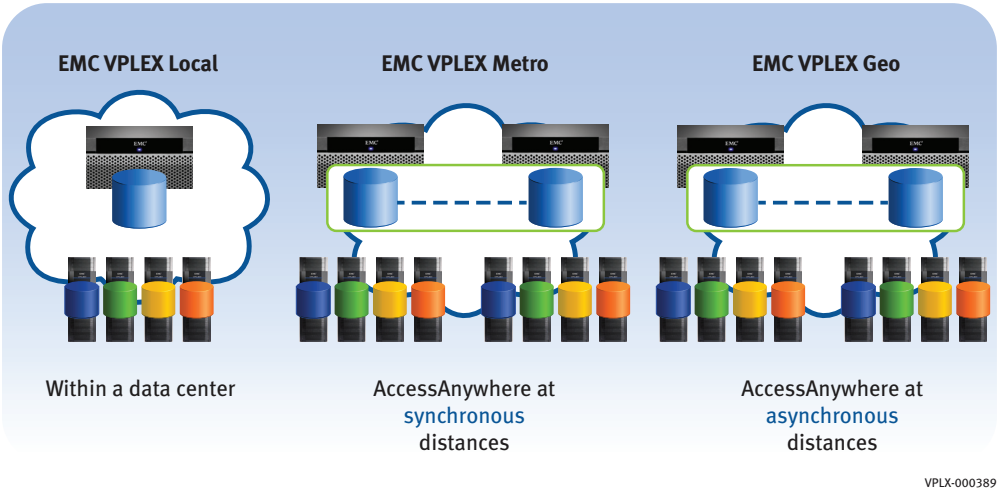

Figure 3 VPLEX offerings

GeoSynchrony 5.0 runs on both the VS1 hardware and the VS2 hardware offerings. VS2 hardware is new with this release of VPLEX.

A VPLEX cluster (both VS1 and VS2) consists of a single-engine, dual-engines, or quad-engines and a management server. Each engine contains two directors. A dual-engine or quad-engine cluster also contains a pair of Fibre Channel switches for communication between directors and a pair of UPS (Uninterruptible Power Sources) for battery power backup of the Fibre Channel switches and the management server.

The management server has a public Ethernet port, which provides cluster management services when connected to the customer network.

VPLEX Local

VPLEX Local provides seamless, non-disruptive data mobility and the ability to manage and mirror data between multiple heterogeneous arrays from a single interface within a data center. VPLEX Local consists of a single VPLEX cluster.

VPLEX Local is a next-generation architecture that allows increased availability, simplified management, and improved utilization and availability across multiple arrays.

1. VPLEX Geo requires a minimum release of GeoSynchrony 5.0.1 or later.

VPLEX Metro

VPLEX Metro enables active/active, block level access to data between two sites within synchronous distances. The distance is limited not only by physical distance but also by host and application requirements. Depending on the application, VPLEX clusters should be installed with inter-cluster links that can support not more than 5ms[1] round trip delay (RTT)

The combination of virtual storage with VPLEX Metro and virtual servers enables the transparent movement of virtual machines and storage across synchronous distances. This technology provides improved utilization and availability across heterogeneous arrays and multiple sites.

VPLEX Geo

VPLEX Geo enables active/active, block level access to data between two sites within asynchronous distances. VPLEX Geo enables more cost-effective use of resources and power.

VPLEX Geo extends the distance for distributed devices up to and within 50ms RTT. As with any asynchronous transport media, you must also consider bandwidth to ensure optimal performance. Due to the asynchronous nature of distributed writes, VPLEX Geo has different availability and performance characteristics than Metro.

Note: VPLEX Geo requires a minimum release of GeoSynchrony 5.0.1 or greater.

Architecture highlights

VPLEX with GeoSynchrony 5.0 is open and heterogeneous, supporting both EMC storage and arrays from other storage vendors, such as HDS, HP, and IBM. VPLEX conforms to established world wide naming (WWN) guidelines that can be used for zoning.

Refer to the EMC Simplified Support Matrix for VPLEX located on Powerlink.

VPLEX provides storage federation for operating systems and applications that support clustered file systems, including both physical and virtual server environments with VMware ESX and Microsoft Hyper-V. VPLEX supports network fabrics from Brocade and Cisco.

Refer to the *EMC Simple Support Matrix, EMC VPLEX and GeoSynchrony*, available at http://elabnavigator.EMC.com under the Simple Support Matrix tab.

An example of the architecture is shown in Figure 4 on page 21.

1. Refer to VPLEX and vendor-specific White Papers for confirmation of latency limitations.

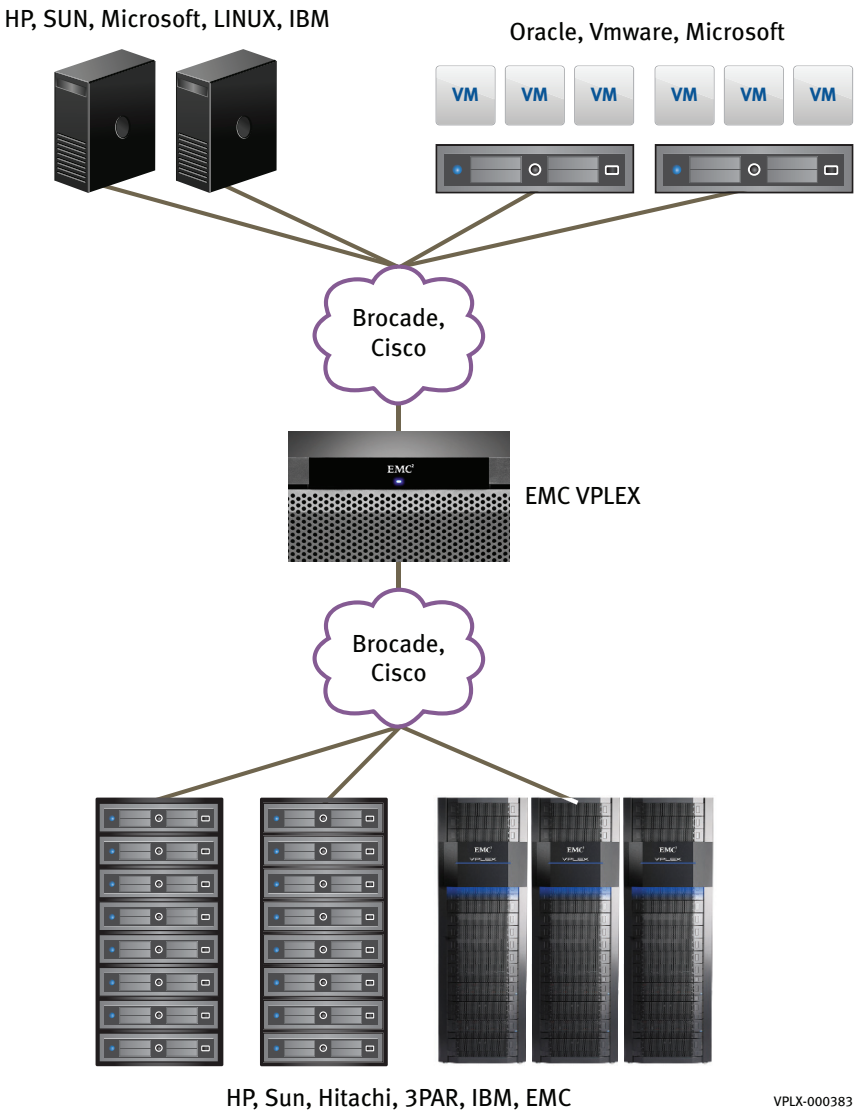

Figure 4 Architecture highlights

Table 1 on page 22 lists an overview of VPLEX features along with the benefits.

Table 1 Overview of VPLEX features and benefits

Features	Benefits
Mobility	**Migration:** Move data and applications without impact on users. **Virtual Storage federation:** Achieve transparent mobility and access in a data center and between data centers. **Scale-out cluster architecture:** Start small and grow larger with predictable service levels.
Availability	**Resiliency:** Mirror across arrays within a single data center or between data centers without host impact. This increases availability for critical applications. **Distributed Cache Coherency:** Automate sharing, balancing, and failover of I/O across the cluster and between clusters whenever possible. **Advanced data caching:** Improve I/O performance and reduce storage array contention.
Collaboration	**Distributed Cache Coherency:** Automate sharing, balancing, and failover of I/O across the cluster and between clusters whenever possible.

For all VPLEX products, GeoSynchrony:

- Presents storage volumes from back-end arrays to VPLEX engines
- Federates the storage volumes into hierarchies of VPLEX virtual volumes with user-defined configuration and protection levels
- Presents virtual volumes to production hosts in the SAN via the VPLEX front-end
- For VPLEX Metro and VPLEX Geo products, presents a global, block-level directory for distributed cache and I/O between VPLEX clusters

Location and distance determine high-availability and data mobility requirements.

When back-end storage arrays or application hosts span two data centers, the AccessAnywhere feature in VPLEX Metro or a VPLEX Geo federates storage in an active/active configuration between VPLEX clusters. Choosing between VPLEX Metro or VPLEX Geo depends on distance, availability, and data synchronicity requirements.

Application and back-end storage I/O throughput, along with availability requirements determine the number of engines in each VPLEX cluster. High-availability features within the VPLEX cluster allow for non-disruptive software upgrades and hardware expansion as I/O throughput increases.

Robust high availability with VPLEX Witness

VPLEX uses rule sets to define how a failure should be handled in a VPLEX Metro or VPLEX Geo configuration. If two clusters lose contact or if one cluster fails, the rule set defines which cluster continues operation and which suspends I/O. This works in many cases of link failure or cluster failure. However, there are still cases in which all I/O must be suspended resulting in a data unavailability. VPLEX with GeoSynchrony 5.0 is introduces the new functionality of VPLEX Witness. VPLEX Metro combined with VPLEX Witness provides the following features:

- High availability for applications in a VPLEX Metro configuration leveraging synchronous consistency groups (no single points of storage failure)
- Fully automatic failure handling of synchronous consistency groups in a VPLEX Metro configuration (provided these consistency groups are configured with a specific preference)
- Better resource utilization

When VPLEX Witness is deployed with a VPLEX Geo system, it can be used for diagnostic purposes but it *does not* automate any fail-over decisions for asynchronous consistency groups.

Typically data centers implement highly available designs *within* a data center, and deploy disaster recovery functionality *between* data centers. Traditionally, within the data center, components operate in active/active mode (or active/passive with automatic failover). However, between data centers, legacy replication technologies use active/passive techniques and require manual failover to use the passive component.

When using VPLEX Metro active/active replication technology in conjunction with VPLEX Witness, the lines between local high availability and long-distance disaster recovery are somewhat blurred because high availability is stretched beyond the data center walls. Because the idea of replication is a by-product of federated and distributed storage disaster avoidance, it is achievable within these geographically dispersed high-availability environments.

VPLEX Witness augments the failure handling for distributed virtual volumes placed into synchronous consistency groups by providing perspective as to the nature of a failure and providing the proper guidance in the event of a cluster failure or inter-cluster link failure.

Note: VPLEX Witness has no effect on failure handling for distributed volumes outside of consistency groups or volumes in asynchronous consistency groups. Witness also has no effect on distributed volumes in synchronous consistency groups when the preference rule is set to no-automatic-winner.

See Chapter 4, "System Integrity and Resiliency" for more information on VPLEX Witness including the differences in how VPLEX Witness handles failures and recovery.

Figure 5 on page 24 shows a high level architecture of VPLEX Witness. The VPLEX Witness server ***must*** reside in a failure domain separate from Cluster 1 and Cluster 2.

Note: The VPLEX Witness server supports round trip time latency of 1 second over the management IP network.

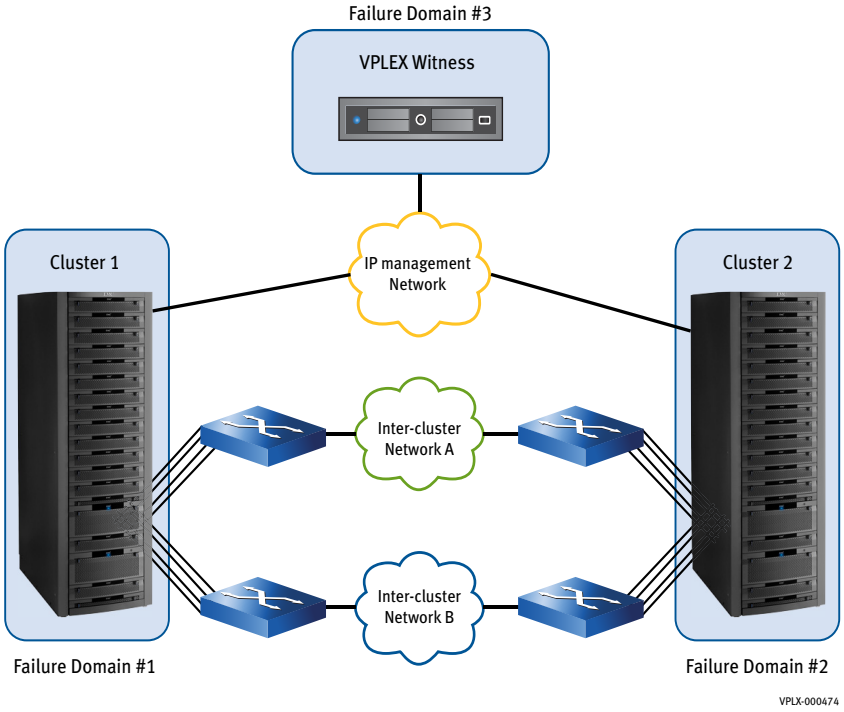

Figure 5 **High level VPLEX Witness architecture**

Because the VPLEX Witness server resides in a separate failure domain to both of the VPLEX clusters, it can gain more perspective as to the nature of a failure and provide correct guidance. It is this perspective that is vital to distinguishing between a site outage and a link outage because either one of these scenarios requires VPLEX to take a different action.

Additional new features in GeoSynchrony 5.0

GeoSynchrony 5.0 provides support for some features already provided by existing array and software packages that might be in use in your storage configuration. Specifically, GeoSynchrony 5.0 now supports the following features:

- Cross Connect
- ALUA

Cross connect

You can deploy a VPLEX Metro high availability Cross Connect when two sites are within campus distance of each other (up to 1ms round trip time latency) and the sites are running VMware HA and VMware Distributed Resource Scheduler (DRS). You can then deploy A VPLEX Metro distributed volume across the two sites using a cross connect front-end configuration and install a VPLEX Witness server in a different failure domain.

Note: Cross Connect is supported in VPLEX Metro deployments only.

Figure 6 shows a high level schematic of a VPLEX Metro high availability Cross Connect solution for VMware. This type of configuration brings the ability to relocate virtual machines over distance which is extremely useful in disaster avoidance, load balancing, and cloud infrastructure use cases all relying on out-of-the-box features and functions. Additional value can be derived from deploying the VPLEX Metro HA Cross Connect solution to ensure total availability.

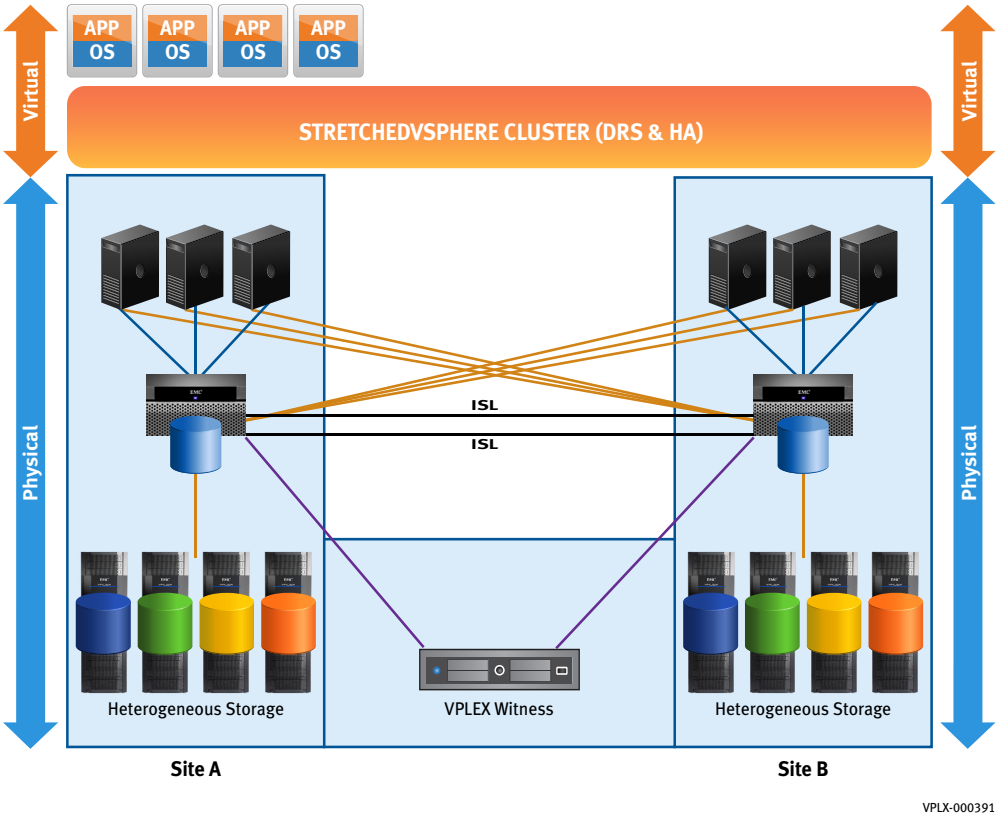

Figure 6 Metro HA Cross Connect solution for VMware

If a physical host failure occurs at either Site A or Site B the VMware high availability cluster restarts the affected virtual machines on the surviving ESX servers.

For more information on Metro HA Cross Connect, see "VPLEX Metro HA in a campus" on page 101.

Leveraging ALUA

GeoSynchrony 5.0 supports Asymmetric Logical Unit Access (ALUA), a feature provided by many new active/passive arrays. VPLEX with GeoSynchrony 5.0 can now take advantage of arrays that support ALUA. In active/passive arrays, logical units or LUNs are normally exposed through several array ports on different paths and the characteristics of the paths might be different. ALUA calls these paths characteristics *access states*. ALUA provides a framework for managing these access states.

The most important access states are active/optimized and active/non-optimized.

- Active optimized paths usually provide higher bandwidth than active non-optimized paths. Active/optimized paths are paths that go to the service processor of the array that owns the LUN.
- I/O that goes to the active non-optimized ports must be transferred to the service processor that owns the LUN internally. This transfer increases latency and has an impact on the array.

VPLEX is able to detect the active optimized paths and the active/non-optimized paths and performs round robin load balancing across all of the active optimized paths. Because VPLEX is aware of the active/optimized paths, it is able to provide better performance to the LUN.

With *implicit ALUA*, the array is in control of changing the access states of the paths. Therefore, if the controller that owns the LUN being accessed fails, the array changes the status of active/non-optimized ports into active/optimized ports and trespasses the LUN from the failed controller to the other controller. Figure 7 shows an example of implicit ALUA.

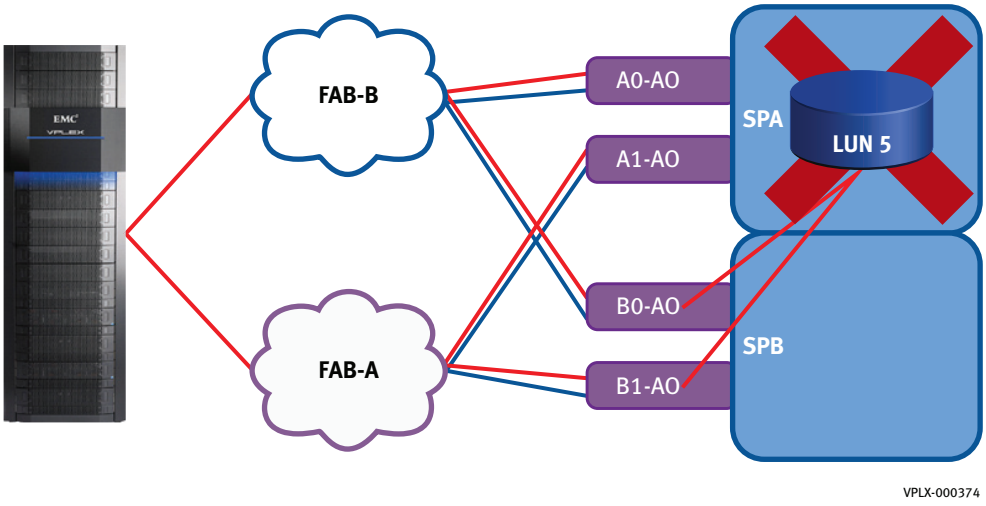

Figure 7 Implicit ALUA

With *explicit ALUA*, the host (or VPLEX) is able to change the ALUA path states. If the active/optimized path fails, VPLEX causes the active/non-optimized paths to become active/optimized paths and as a result, increase the performance. I/O can go between the controllers to access the LUN through a very fast bus. There is no need to trespass the LUN in this case. Figure 8 shows an example of explicit ALUA.

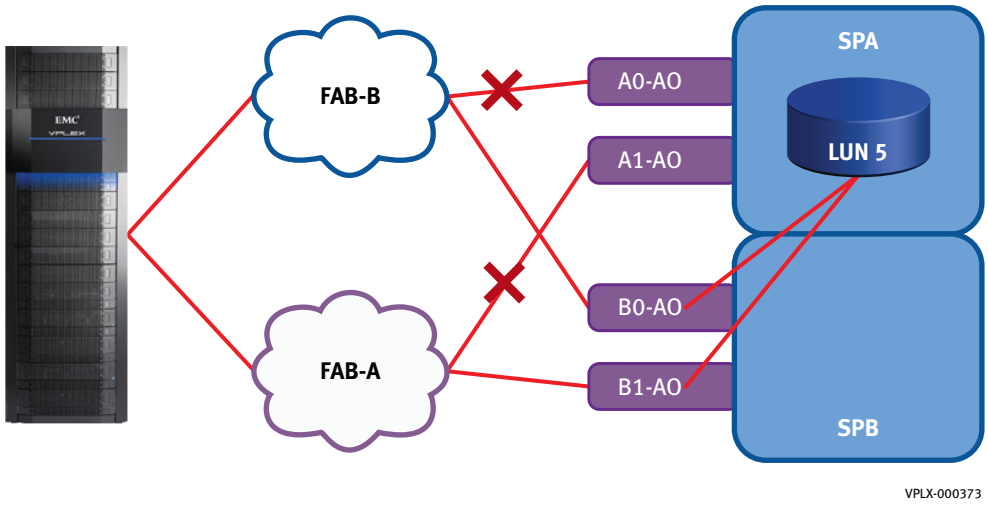

Figure 8 Explicit ALUA

Upgrade paths

VPLEX facilitates application and storage upgrades without a disruption.

This flexibility means that VPLEX is always servicing I/O and never has to be completely shut down.

Storage, application, and host upgrades

The mobility features of VPLEX enable the easy addition or removal of storage, applications and hosts. When VPLEX encapsulates back-end storage, the block-level nature of the coherent cache allows the upgrade of storage, applications, and hosts. You can configure VPLEX so that all devices within VPLEX have uniform access to all storage blocks.

Hardware upgrades

When capacity demands increase in a data center, VPLEX supports hardware upgrades for single-engine VPLEX systems to dual-engine and dual-engine to quad-engine systems. These upgrades also increase the availability of front-end and back-end ports in the data center.

Software upgrades

VPLEX features a robust non-disruptive upgrade (NDU) technology to upgrade the software on VPLEX engines. Management server software must be upgraded before running the NDU.

The redundancy of ports, paths, directors, and engines in VPLEX means that GeoSynchrony on a VPLEX Local or VPLEX Metro can be upgraded without interrupting host access to storage. No service window or application disruption is required to upgrade VPLEX GeoSynchrony on VPLEX Local or VPLEX Metro. On VPLEX Geo the upgrade script ensures that the application is active/passive before allowing the upgrade.

Simple support matrix

EMC publishes storage array interoperability information in a Simple Support Matrix available on EMC PowerLink. This information details tested, compatible combinations of storage hardware and applications that VPLEX supports. The Simple Support Matrix can be located at:

http://Powerlink.EMC.com

VPLEX management interfaces

GeoSynchrony supports multiple methods of management and monitoring for the VPLEX cluster:

- Web-based GUI: for graphical ease of management from a centralized location.
- VPLEX CLI: for command line management of clusters.
- VPLEX Element Manager API: software developers and other users use the API to create scripts to run VPLEX CLI commands.
- SNMP Support for performance statistics: Supports retrieval of performance-related statistics as published in the VPLEX-MIB.mib.

2 VPLEX Hardware Overview

This chapter describes the major VPLEX hardware components including the following topics:

System components

Note: This chapter describes only the VS2 hardware for VPLEX clusters. If you are currently running on a VS1 system, see the VPLEX with GeoSynchrony 4.2 Installation and Setup Guide available in the Procedure Generator for a description of the VS1 hardware.

Figures 9 shows the main hardware components in a quad-engine VPLEX cluster. Figure 10 on page 33 shows the main hardware components in a dual-engine VPLEX cluster. Figure 11 on page 34 shows the main hardware components in a single-engine VPLEX cluster.

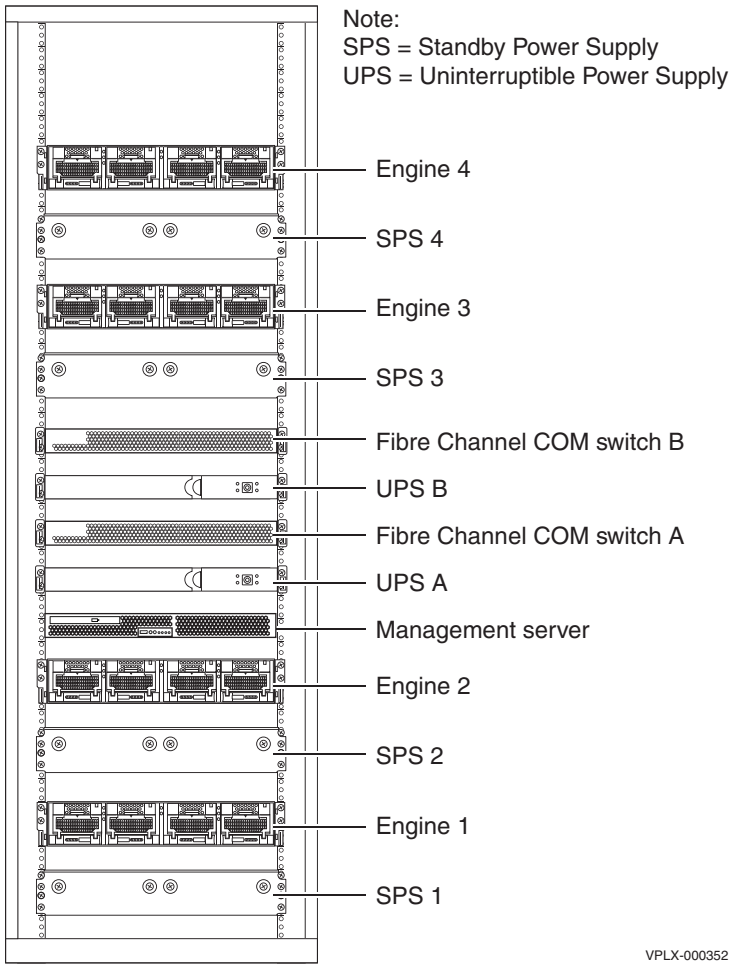

Figure 9 Quad-engine VPLEX cluster

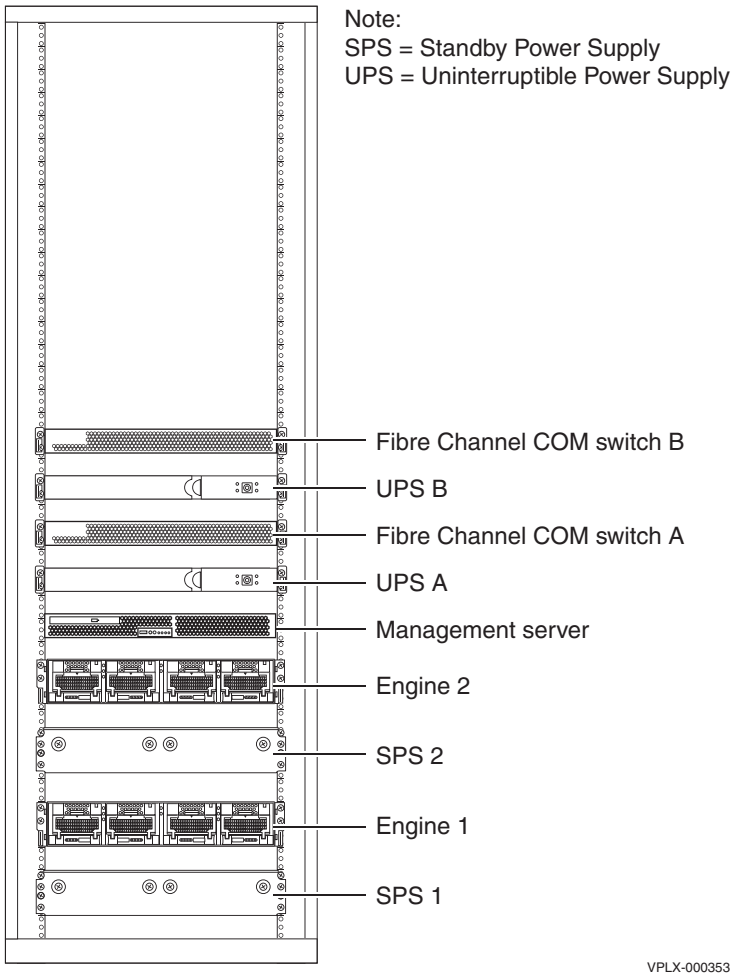

Figure 10 Dual-engine VPLEX cluster

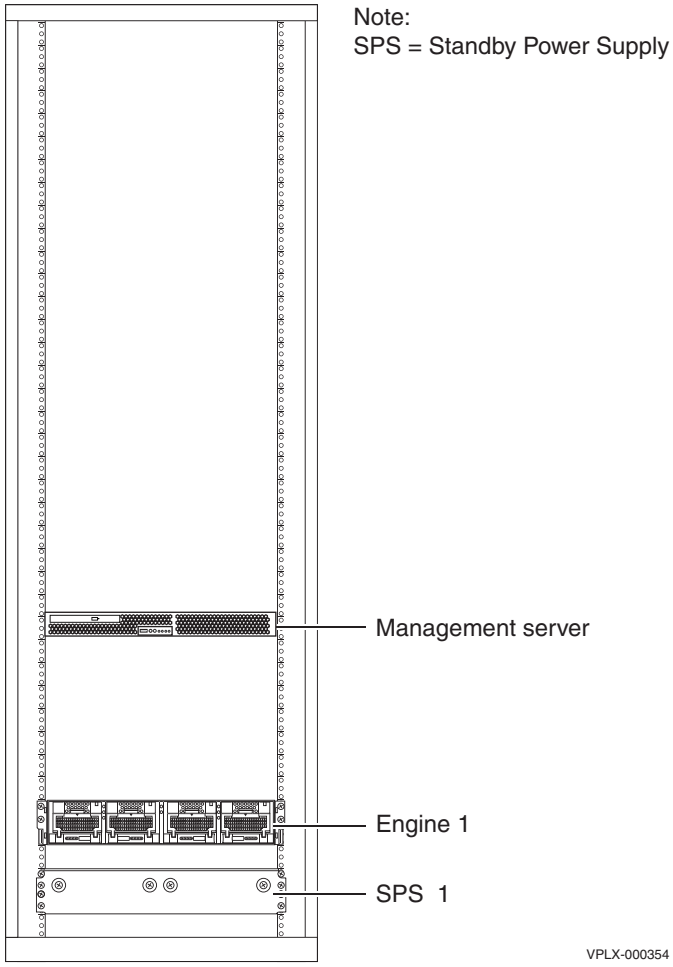

Figure 11 Single-engine VPLEX cluster

Table 2 describes the major components and their functions.

Table 2 Hardware components

Feature	Description
Engine	Contains two directors, with each providing front-end and back-end I/O connections.
Director	Contains: • Five I/O modules (IOMs), as identified in Figure 12 on page 36 • Management module, for intra-cluster communication • Two redundant 400 W power supplies with built-in fans • CPU • Solid-state disk (SSD) that contains the GeoSynchrony operating environment • RAM
Management server	Provides: • Management interface to a public IP network • Management interfaces to other VPLEX components in the cluster • Event logging service

Table 2 Hardware components

Feature	Description
Fibre Channel COM switches (Dual-engine or quad-engine cluster only)	Provides intra-cluster communication support among the directors. (This is separate from the storage I/O.)
Power subsystem	Power distribution panels (PDPs) connect to the site's AC power source, and transfer power to the VPLEX components through power distribution units (PDUs). This provides a centralized power interface and distribution control for the power input lines. The PDPs contain manual on/off power switches for their power receptacles.
Standby Power Supply (SPS)	One SPS assembly (two SPS modules) provides backup power to each engine in the event of an AC power interruption. Each SPS module maintains power for two five-minute periods of AC loss while the engine shuts down.
Uninterruptible Power Supply (UPS) (Dual-engine or quad-engine cluster only)	One UPS provides battery backup for Fibre Channel switch A and the management server, and a second UPS provides battery backup for Fibre Channel switch B.

The VPLEX engine

The VPLEX VS2 engine Contains two directors, with each providing front-end and back-end I/O connections. Each of these module types are described in more detail in "The VPLEX director."

Figure 12 identifies the modules in an engine.

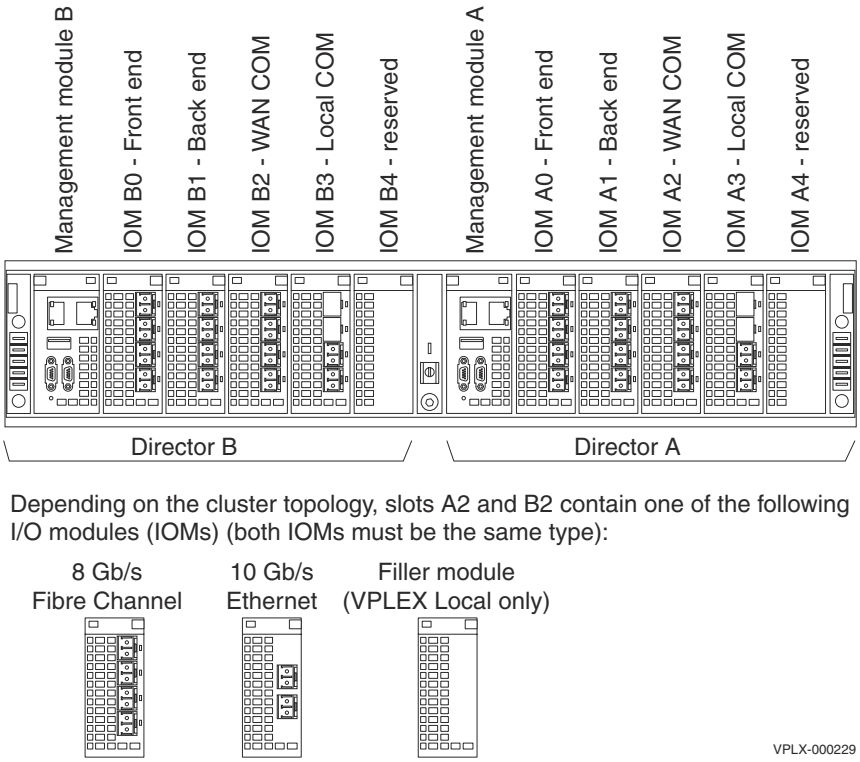

Figure 12 Engine, rear view

The VPLEX director

Each director services host I/O. The director hosts the GeoSynchrony operating environment for such VPLEX functions as I/O request processing, distributed cache management, virtual-to-physical translations, and interaction with storage arrays.

Front-end and back-end connectivity

Four 8 Gb/s Fibre Channel I/O modules are provided for front-end connectivity, and four 8 Gb/s ports are provided for back-end connectivity.

The industry-standard Fibre Channel ports connect to host initiators and storage devices.

WAN connectivity in VPLEX Metro and VPLEX Geo

WAN communication between VPLEX Metro or VPLEX Geo clusters is over Fibre Channel (8 Gbps) for VPLEX Metro, or Gigabit Ethernet (10 GbE) for VPLEX Geo.

CAUTION

The inter cluster link carries unencrypted user data. To protect the security of the data, secure connections are required between clusters.

Director redundancy

When properly zoned and configured, the front-end and back-end connections provide redundant I/O that can be serviced by any director in the VPLEX configuration.

Director redundancy is provided by connecting ports in dedicated Fibre Channel I/O modules to an internal Fibre Channel network. Directors within an engine are directly connected through an internal communication channel, directors are connected between engines through dual Fibre Channel switches. Through this network, each VPLEX director participates in intra-cluster communications.

VPLEX cluster architecture

The distributed VPLEX hardware components are connected through both Ethernet or Fibre Channel cabling and respective switching hardware.

I/O modules provide front-end and back-end connectivity between SANs and to remote VPLEX clusters in VPLEX Metro or VPLEX Geo configurations.

Management server

The management server in each VPLEX cluster provides management services that are accessible from a public IP network.

The management server coordinates event notification, data collection, VPLEX software upgrades, configuration interfaces, diagnostics, and some director-to-director communication. The management server also forwards VPLEX Witness traffic between directors in the local cluster and the remote VPLEX Witness server.

Both clusters in either VPLEX Metro or VPLEX Geo configuration can be managed from a single management server.

The management server is on a dedicated, internal management IP network that provides accessibility for all major components in the cluster. The management server provides redundant internal management network IP interfaces. In addition to the internal management IP network, each management server connects to the public network, which serves as the access point.

Fibre Channel switches

The Fibre Channel switches provide high availability and redundant connectivity between directors and engines in a dual-engine or quad-engine cluster. Each Fibre Channel switch is powered by a UPS, and has redundant I/O ports for intra-cluster communication.

The Fibre Channel switches do not connect to the front-end hosts or back-end storage.

VPLEX power supply modules

Two independent power zones in a data center feed each VPLEX cluster, providing a highly available power distribution system. To assure fault tolerant power in the cabinet system, external AC power must be supplied from independent power distribution units (PDUs) at the customer site, as shown in Figure 13.

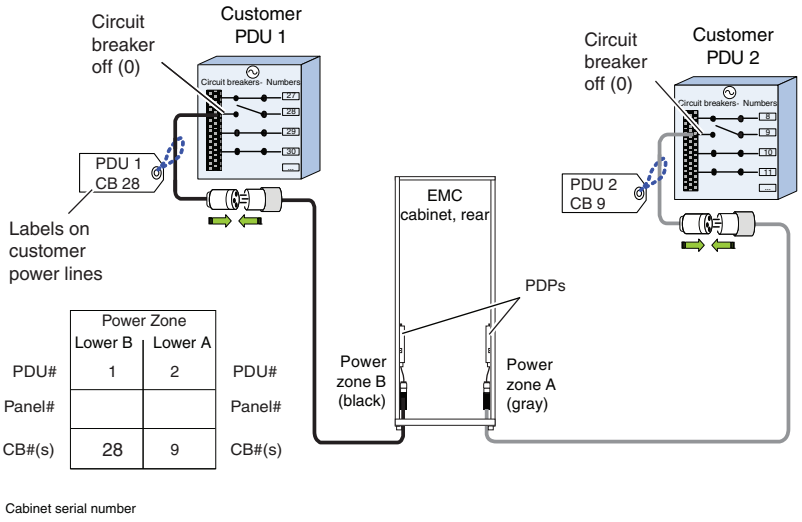

Figure 13 VPLEX cluster independent power zones

The PDPs contain manual on/off power switches for their power receptacles. For additional information on power requirements, see the *EMC Best Practices Guide for AC Power Connections in Two-PDP Bays.*

The power supply module is a FRU and can be replaced with no disruption to the services provided only one PS module is replaced at a time.

WARNING

Removal of both power supply modules on a director will initiate a vault on the cluster.

Standby Power Supplies

Each engine is connected to two standby power supplies (SPS) that provide battery backup to each director to ride through transient site power failure as well as to provide sufficient time to vault their cache in case power is not restored within 30 seconds. A single standby power supply provides enough power for the attached engine. Refer to "VPLEX distributed cache protection and redundancy" in Chapter 4, "System Integrity and Resiliency."

Uninterrupted power supplies

In the event of a power failure, in dual- and quad-engine clusters, the management server and Fibre Channel switch A draw power from UPS-A. UPS-B provides battery backup for Fibre Channel switch B. In this way, Fibre Channel switches and the management server in multi-engine configurations can continue operation for 5 minutes in the event of a power failure.

Power and Environmental monitoring

A GeoSynchrony service performs the overall health monitoring of the VPLEX cluster and provides environmental monitoring for the VPLEX cluster hardware. It monitors various power and environmental conditions at regular intervals and logs any condition changes into the VPLEX messaging system.

Any condition that indicates a hardware or power fault generates a call home event to notify the user.

VPLEX component failures

All critical processing components of a VPLEX system use at a minimum pair-wise redundancy to maximize data availability. This section describes how VPLEX component failures are handled and the best practices that should be used to allow applications to tolerate these failures.

All component failures that occur within a VPLEX system are reported through events that call back to the EMC Service Center to ensure timely response and repair of these fault conditions.

Storage array outages

To overcome both planned and unplanned storage array outages, VPLEX supports the ability to mirror the data of a virtual volume between two or more storage volumes using a RAID 1 device. Figure 14 shows a virtual volume that is mirrored between two arrays. Should one array experience an outage, either planned or unplanned, the VPLEX system can continue processing I/O on the surviving mirror leg. Upon restoration of the failed storage volume, VPLEX synchronizes the data from the surviving volume to the recovered leg.

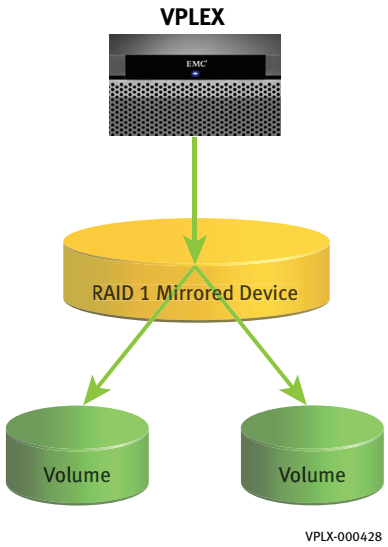

Figure 14 Local mirrored volumes

Best practices for local mirrored volumes

- For critical data, it is recommended to mirror data onto two or more storage volumes that are provided by separate arrays.
- For the best performance, these storage volumes should be configured identically and be provided by the same type of array.

Fibre Channel port failures

The small form-factor pluggable (SFP) transceivers that are used for connectivity to VPLEX are serviceable Field Replaceable Units (FRUs).

Best practices for Fibre Channel ports

Follow these best practices to ensure the highest reliability of your configuration:

Front end:

- Ensure there is a path from each host to at least one front-end port on director A and at least one front-end port on director B. When the VPLEX cluster has two or more engines, ensure that the host has at least one A-side path on one engine and at least one B-side on a separate engine. For maximum availability, each host can have a path to at least one front-end port on every director.
- Use multi-pathing software on the host servers to ensure timely response and continuous I/O in the presence of path failures.
- Ensure that each host has a path to each virtual volume through each fabric.
- Ensure that the fabric zoning provides hosts redundant access to the VPLEX front-end ports.

Back end:

- Ensure that the LUN mapping and masking for each storage volume presented from a storage array to VPLEX presents the volumes out of at least two ports from the array on at least two different fabrics from different controllers.
- Ensure that the LUN connects to at least two different back end ports of each director within a VPLEX cluster.
- Active/passive arrays must have one active and one passive port zoned to each director, and zoning must provide VPLEX with the redundant access to the array ports.
- Configure a maximum of eight paths between one director and one LUN (two directors can each have eight paths to a LUN).

Note: On VS2 hardware, only 4 physical ports are available for back end connections on each director. Refer to the *Installation and Setup Guide* for your system for details on the hardware configuration you are using.

I/O module failure

I/O modules within VPLEX serve dedicated roles. In VS2, each VPLEX director has one front-end I/O module, one back-end I/O module, and one COM I/O module used for intra- and inter-cluster connectivity. Each I/O module is a serviceable FRU. The following sections describe the behavior of the system.

Front end I/O module

Should a front end I/O module fail, all paths connected to this I/O module fail and VPLEX will call home. The "Best practices for Fibre Channel ports" on page 42 should be followed to ensure that hosts have a redundant path to their data.

During the removal and replacement of an I/O module, the affected director will be reset.

Back end I/O module

Should a back end I/O module fail, all paths connected to this I/O module fail and VPLEX will call home. The "Best practices for Fibre Channel ports" on page 42 should be followed to ensure that each director has a redundant path to each storage volume through a separate I/O module.

During the removal and replacement of an I/O module, the affected director resets.

COM I/O module

Should the local COM I/O module of a director fail, the director resets and all service provided from the director stops. The "Best practices for Fibre Channel ports" on page 42 ensure that each host has redundant access to its virtual storage through multiple directors, so the reset of a single director will not cause the host to lose access to its storage.

During the removal and replacement of a local I/O module, the affected director will be reset.

Director failure

A director failure causes the loss of all service from that director. Each VPLEX Engine has a pair of directors for redundancy. VPLEX clusters containing two or more engines benefit from the additional redundancy provided by the additional directors. Each director within a cluster is capable of presenting the same storage. The "Best practices for Fibre Channel ports" on page 42 allow a host to ride through director failures by placing redundant paths to their virtual storage through ports provided by different directors. The combination of multipathing software on the hosts and redundant paths through different directors of the VPLEX system allows the host to ride through the loss of a director.

Each director is a serviceable FRU.

Intra-cluster IP management network failure

Each VPLEX cluster has a pair of private local IP subnets that connect the directors to the management server. These subnets are used for management traffic, protection against intra-cluster partitioning, and communication between the VPLEX Witness server (if it is deployed) and the directors. Link loss on one of these subnets can result in the inability of some subnet members to communicate with other members on that subnet; this results in no loss of service or manageability due to the presence of the redundant subnet, though it might result in loss of connectivity between this director and VPLEX Witness.

Intra-cluster Fibre Channel switch failure

Each VPLEX cluster with two or more engines uses a pair of dedicated Fibre Channel switches for intra-cluster communication between the directors within the cluster. Two redundant Fibre Channel fabrics are created with each switch serving a different fabric. The loss of a single Fibre Channel switch results in no loss of processing or service.

Inter-cluster WAN links

In VPLEX Metro and VPLEX Geo configurations the clusters are connected through WAN links that you provide. Follow these best practices when setting up your VPLEX clusters.

Best practices for inter-cluster WAN links

Follow these best practices when setting up your VPLEX clusters:

- For VPLEX Metro configurations, latency must be less than 5ms round trip time (RTT).
- For VPLEX Geo configurations, latency must be less than 50ms RTT.
- Links must support a minimum of 45Mb/s of bandwidth. However, the required bandwidth is dependent on the I/O pattern and must be high enough for all writes to all distributed volumes to be exchanged between clusters.
- The switches used to connect the WAN links between both clusters should be configured with a battery backup UPS.
- Use physically independent WAN links for redundancy.

- Every WAN port on every director must be able to connect to a WAN port on every director in the other cluster.
- Logically isolate VPLEX Metro or VPLEX Geo traffic from other WAN traffic using VSANs or LSANs.
- Independent inter switch links (ISLs) are strongly recommended for redundancy.
- Use VPLEX Witness in an independent failure domain to improve availability of the VPLEX Metro solution.

Power supply failures

Each VPLEX cluster provides two zones of AC power. If one zone loses power, the modules in the cluster can continue to run using power from the other zone. When power is lost in both zones, the engines revert to power from their SPS modules. In multi-engine clusters the management server and intra cluster Fibre Channel switches revert to the power supplied by the UPS.

Standby power supply failure

Each SPS is a FRU and can be replaced with no disruption to the services provided by the system. The recharge time for an SPS is up to 5.5 hours and the batteries in the standby power supply are capable of supporting two sequential outages of no greater than 5 minutes without data loss.

Note: While the batteries can support two 5-minute power losses, the VPLEX Local, VPLEX Metro, or VPLEX Geo cluster vaults after a 30 second power loss to ensure enough battery power to complete the cache vault.

Table 3 shows the different power-loss scenarios that lead to a vault. UPS failures

Each UPS is a FRU and can be replaced with no disruption to the services provided by the system. The UPS modules provide up to 5 minutes of battery backup power to the Fibre Channel switches in a multi — engine cluster. The batteries require a 6 hour recharge time for 90% capacity.

Power failures that cause vault

GeoSynchrony monitors both the power supply and the standby power supplies (SPS). If a single power fault in each power zone occurs in a cluster, GeoSynchrony will initiate a vault. Table 3 shows the conditions under which a vault can occur.

Note: Cache vault is requires releases of GeoSynchrony 5.0.1 or greater.

The scenarios in this table are applicable to single-, dual-, and quad-engine systems provided that the system has a sufficient number of engines for the scenario to apply.

Table 3 Scenarios that cause vault

Scenario	Causes cluster vault?
One or more engines lose power in zone A (while power to all engines is still supplied by zone B).	No
One or more engines lose power in both zones A and B.	Yes
One or more engines lose power in zone A while one or more different engines in the same cluster lose power in zone B.	Yes

VPLEX Witness failure

If VPLEX Witness is deployed, failure of the VPLEX Witness has no impact on I/O as long as the two clusters stay connected with each other. However, if a cluster failure or inter-cluster network partition happens while VPLEX Witness is down, there will be data unavailability on all surviving clusters. The best practice in this situation is to disable VPLEX Witness (while the clusters are still connected) if its outage is expected to be long, and to revert to using preconfigured detach rules. Once VPLEX Witness recovers, it can be re-enabled again with the cluster-witness enable CLI command. Refer to the *EMC VPLEX with GeoSynchrony 5.0 and Point Releases CLI Guide* for information about these commands.

VPLEX management server failure

Each VPLEX cluster has a dedicated management server that provides management access to the directors and supports management connectivity for remote access to the peer cluster in a VPLEX Metro or VPLEX Geo environment. As the I/O processing of the VPLEX directors does not depend upon the management servers, in most cases the loss of a management server does not interrupt the I/O processing and virtualization services provided by VPLEX. However, VPLEX Witness traffic is sent through the Management Server. If the Management Server fails in a configuration running the VPLEX Witness, the VPLEX Witness is no longer able to communicate with the cluster. Should the remote VPLEX cluster fail, data becomes unavailable. If the inter-cluster network partitions, the remote cluster always proceeds with I/O regardless of preference because it is still connected to the Witness[1].

If the failure of the Management Server is expected to be long, it may be desirable to disable VPLEX Witness functionality while the clusters are still connected. Refer to the *EMC VPLEX with GeoSynchrony 5.0 and Point Releases CLI Guide* for information about the commands used to disable and enable the VPLEX Witness.

1. This description only applies to synchronous consistency groups with a rule setting that identifies a specific preference.

3 VPLEX Software

This chapter provides information on various components in the VPLEX software.

GeoSynchrony

GeoSynchrony is the operating system running on VPLEX directors. GeoSynchrony is an intelligent, multitasking, locality-aware operating environment that controls the data flow for virtual storage. GeoSynchrony is:

- Optimized for mobility, availability, and collaboration
- Designed for highly available, robust operation in geographically distributed environments
- Driven by real-time I/O operations
- Intelligent about locality of access
- Provides the global directory that supports AccessAnywhere

GeoSynchrony supports your mobility, availability and collaboration needs.

Table 4 AccessAnywhere capabilities

Virtualization Capability	Provides the following
Storage volume encapsulation	LUNs on a back-end array can be imported into an instance of VPLEX and used while keeping their data intact. **Considerations:** The storage volume retains the existing data on the device and leverages the media protection and device characteristics of the back-end LUN.
RAID 0	VPLEX devices can be aggregated to create a RAID 0 striped device. **Considerations:** Improves performance by striping I/Os across LUNs.
RAID-C	VPLEX devices can be concatenated to form a new larger device. **Considerations:** Provides a means of creating a larger device by combining two or more smaller devices.
RAID 1	VPLEX devices can be mirrored within a site. **Considerations:** Withstands a device failure within the mirrored pair. A device rebuild is a simple copy from the remaining device to the newly repaired device. Rebuilds are done in incremental fashion, whenever possible. The number of required devices is twice the amount required to store data (actual storage capacity of a mirrored array is 50%). The RAID 1 devices can come from different back-end array LUNs providing the ability to tolerate the failure of a back-end array.
Distributed RAID 1	VPLEX devices can be mirrored between sites. **Considerations:** Provides protection from site disasters and supports the ability to move data between geographically separate locations.
Extents	Storage volumes can be broken into extents and devices created from these extents. **Considerations:** Used when LUNs from a back-end storage array are larger than the desired LUN size for a host. This provides a convenient means of allocating what is needed while taking advantage of the dynamic thin allocation capabilities of the back-end array.
Migration	Volumes can be migrated non-disruptively to other storage systems.

Table 4 AccessAnywhere capabilities

Virtualization Capability	Provides the following
	Considerations: Use for changing the quality of service of a volume or for performing technology refresh operations.
Global Visibility	The presentation of a volume from one VPLEX cluster where the physical storage for the volume is provided by a remote VPLEX cluster. **Considerations:** Use for AccessAnywhere collaboration between locations. The cluster without local storage for the volume will use its local cache to service I/O but non-cached operations incur remote latencies to write or read the data.

Management of VPLEX

Within the VPLEX cluster, TCP/IP-based management traffic travels through a private management network to the components in one or more clusters. In VPLEX Metro and VPLEX Geo, VPLEX establishes a VPN tunnel between the management servers of both clusters.

Web-based GUI

VPLEX includes a Web-based graphical user interface (GUI) for management. The EMC VPLEX Management Console Help provides more information on using this interface.

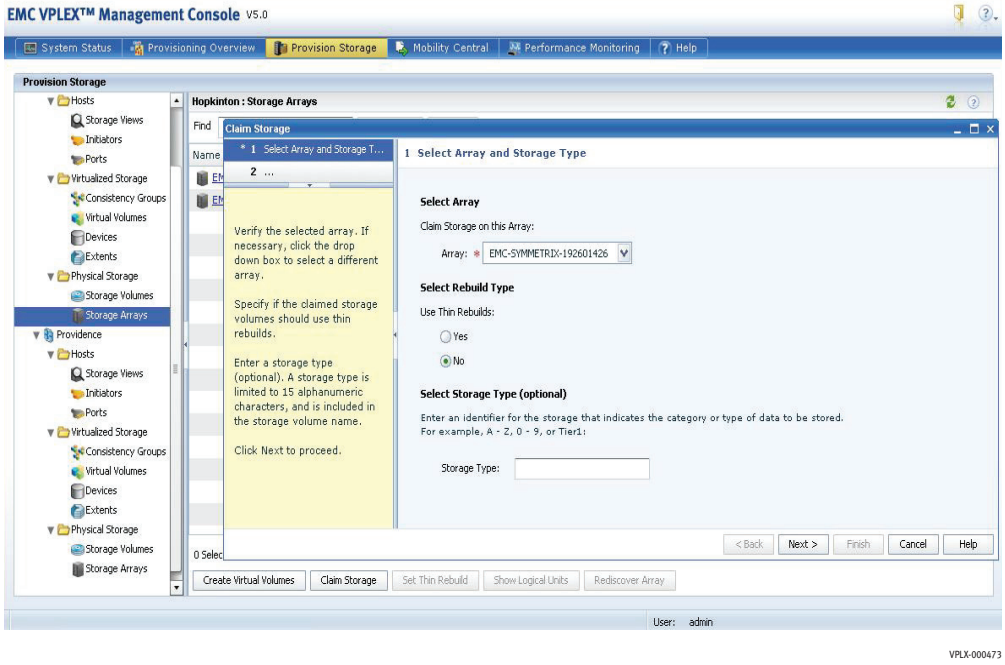

Figure 15 Using the GUI to claim storage

To perform other VPLEX operations that are not available in the GUI, refer to the CLI, which supports full functionality.

VPLEX CLI

VPLEX CLI is a command line interface (CLI) to configure and operate VPLEX systems. The CLI is divided into command contexts. Some commands are accessible from all contexts, and are referred to as *global commands*. The remaining commands are arranged in a hierarchical context tree that can only be executed from the appropriate location in the context tree. Example 1 shows a CLI session that performs the same tasks as shown in Figure 15.

Example 1 Find unclaimed storage volumes, claim them as thin storage, and assign names from a CLARiiON hints file:

```
VPlexcli:/clusters/cluster-1/storage-elements/storage-vol
umes> claimingwizard --file /home/service/clar.txt
--thin-rebuild

Found unclaimed storage-volume
VPD83T3:6006016091c50e004f57534d0c17e011 vendor DGC :
claiming and naming clar_LUN82.
```

```
Found unclaimed storage-volume
VPD83T3:6006016091c50e005157534d0c17e011 vendor DGC :
claiming and naming clar_LUN84.

Claimed 2 storage-volumes in storage array clar

Claimed 2 storage-volumes in total.

VPlexcli:/clusters/cluster-1/storage-elements/storage-vol
umes>
```

The *EMC VPLEX with GeoSynchrony 5.0 and Point Releases CLI Guide* provides a comprehensive list of VPLEX commands and detailed instructions on using those commands.

SNMP support for performance statistics

The VPLEX snmpv2c SNMP agent provides performance statistics as follows:

- Supports retrieval of performance-related statistics as published in the VPLEX-MIB.mib.
- Runs on the management server and fetches performance related data from individual directors using a firmware-specific interface.
- Provides SNMP MIB data for directors for the local cluster only.

LDAP / AD Support

VPLEX offers Lightweight Directory Access Protocol (LDAP) or Active Directory as an authentication directory service.

VPLEX Element Manager API

VPLEX Element Manager API uses the Representational State Transfer (REST) software architecture for distributed systems such as the World Wide Web. It allows software developers and other users to use the API to create scripts to run VPLEX CLI commands.

The VPLEX Element Manager API supports all VPLEX CLI commands that can be executed from the root context on a director.

Call home

The Call Home feature in GeoSynchrony is a leading technology that alerts EMC support personnel of warnings in VPLEX so they can arrange for proactive remote or on-site service. Certain events trigger the Call Home feature. Once a call-home event is triggered, all informational events are blocked from calling home for 8 hours.

Provisioning

VPLEX allows easy storage provisioning among heterogeneous storage arrays. After a storage array LUN volume is encapsulated within VPLEX, all of its block-level storage is available in a global directory and coherent cache. Any front-end device that is zoned properly can access the storage blocks.

Table 5 describes the methods available for provisioning.

Table 5 Provisioning methods

EZ provisioning	EZ provisioning capitalizes on a Create Virtual Volumes wizard that claims storage, creates extents, devices, and then virtual volumes on those devices. EZ provisioning uses the entire capacity of the selected storage volume to create a device, and then creates a virtual volume on top of the device.
Advanced provisioning	Advanced provisioning allows you to slice storage volumes into portions or *extents*. Extents are then available to create devices, and then virtual volumes on these devices. Advanced provisioning requires manual configuration of each of the provisioning steps and the method is useful for tasks such as creating complex devices.

Thick and thin storage volumes

Thin provisioning optimizes the efficiency with which available storage space is used in the network. Unlike traditional (thick) provisioning where storage space is allocated beyond the current requirement in anticipation of a growing need, *thin provisioning* allocates disk storage capacity only as the application needs it — when it writes. Thinly provisioned volumes are expanded dynamically depending on the amount of data written to them, and they do not consume physical space until written to.

VPLEX automatically discovers storage arrays that are connected to its back-end ports. By default, VPLEX treats all storage volumes as if they were thickly provisioned on the array.

Storage volumes that are thinly provisioned on the array should be claimed with the **`thin-rebuild`** parameter in VPLEX.This provides thin to thin copies in VPLEX using a different type of rebuild. Unlike a traditional rebuild that copies all the data from the source to the target, in this case, VPLEX first reads the storage volume, and if the target is thinly provisioned, it does not write unallocated blocks to the target. Writing unallocated blocks to the target would result in VPLEX converting a thin target to thick, eliminating the efficiency of the thin volume.

About extents

An extent is a portion of a disk. The ability to provision extents allows you to break a storage volume into smaller pieces. This feature is useful when LUNs from a back-end storage array are larger than the desired LUN size for a host. Extents provide a convenient means of allocating what is needed while taking advantage of the dynamic thin allocation capabilities of the back-end array. Extents can then be combined into devices.

About devices

Devices combine extents or other devices into one large device with specific RAID techniques such as mirroring or striping. Devices can only be created from extents or other devices. A device's storage capacity is not available until you create a virtual volume on the device and export that virtual volume to a host. You can create only one virtual volume per device.

There are two types of devices:

- A *simple device* is configured by using one component — an extent.

- A *complex device* has more than one component, combined by using a specific RAID type. The components can be extents or other devices (both simple and complex).

Device visibility

You can create a virtual volume from extents at one VPLEX Metro cluster that are available for I/O at the other VPLEX Metro cluster. This is done by making the virtual volume, or the consistency group containing the volume, globally visible. A virtual volume on a top-level device that has global visibility can be exported in storage views on any cluster in a VPLEX Metro. Consistency groups aggregate volumes together to ensure the common application of a set of properties to the entire group. Figure 16 shows a local consistency group with global visibility.

Remote virtual volumes suspend I/O during inter-cluster link outages. As a result, the availability of the data on remote virtual volumes is directly related to the reliability of the inter-cluster link.

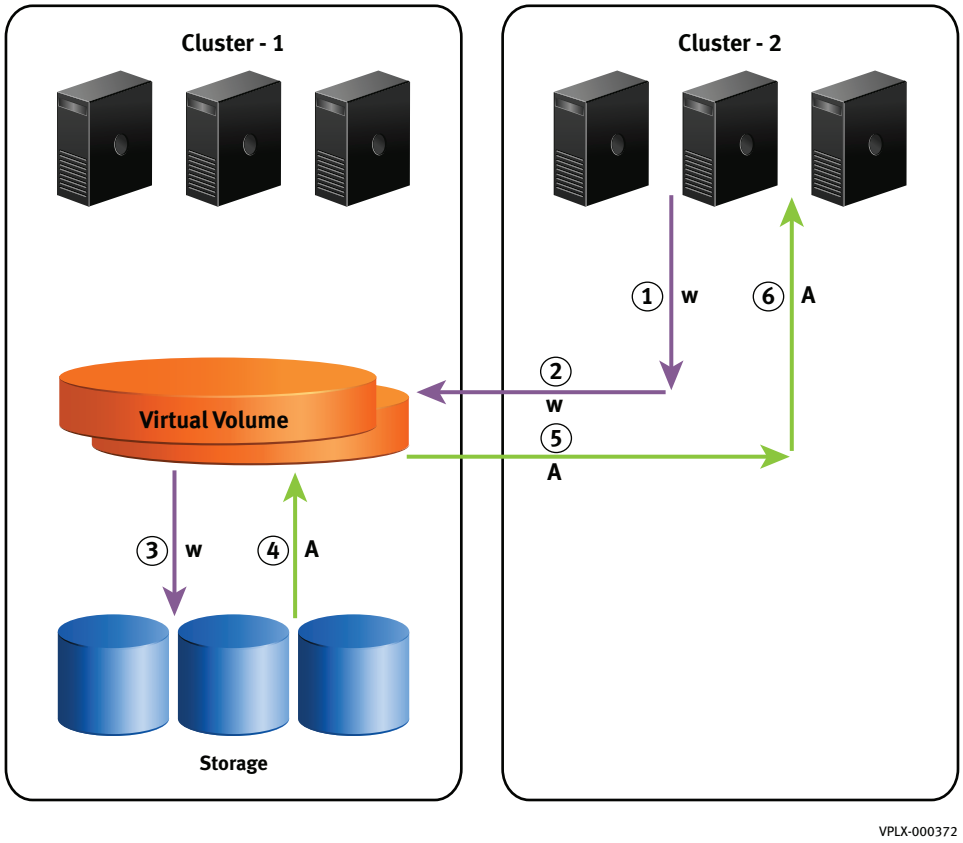

Figure 16 Local consistency group with global visibility[1]

Distributed devices

Distributed devices are present at both clusters for simultaneous active/active read/write access using AccessAnywhere to ensure consistency of the data between the clusters. Each distributed virtual volume looks to the hosts as if it is a centralized single volume served by a single array located in a centralized location; except that

1. In this figure, **W** indicates the write path and **A** indicates the acknowledgement path.

everything is distributed and nothing is centralized. Because distributed devices are configured using storage from both clusters, they are used only in a VPLEX Metro or VPLEX Geo configuration. Figure 17 shows distributed devices.

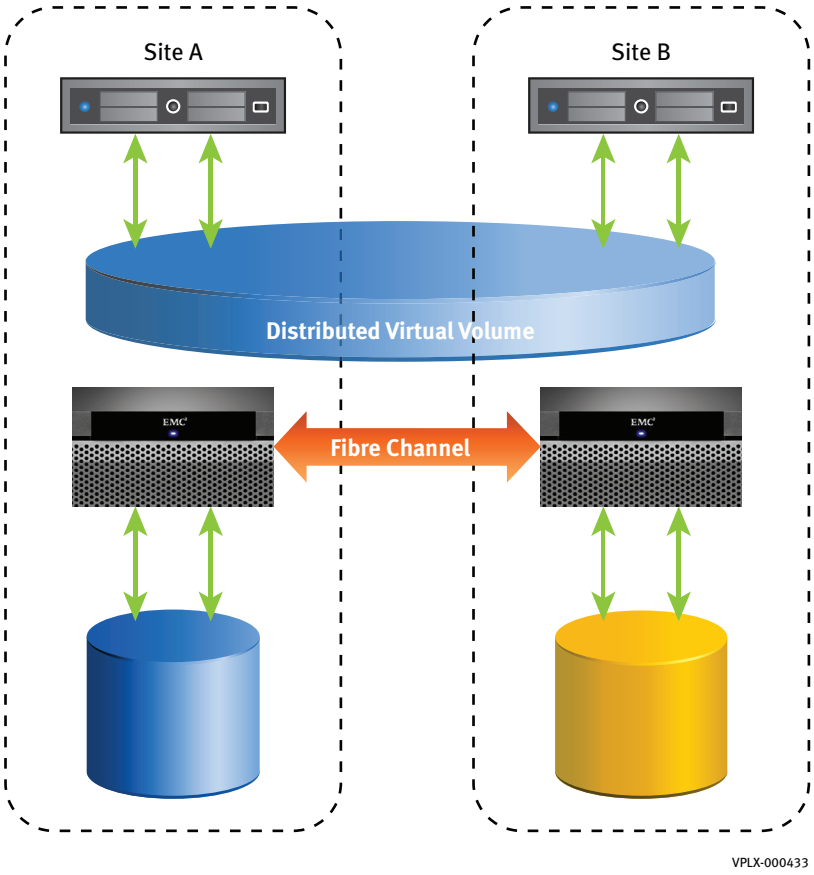

Figure 17 Distributed devices

Virtual volumes

A virtual volume is created on a device or a distributed device, and is presented to a host through a storage view. Virtual volumes are created on top-level devices only, and always use the full capacity of the device or distributed device.

You can non-disruptively expand a virtual volume up to 945 times as necessary. However, expanding virtual volumes created on distributed RAID-1 devices is not supported in the current release.

Logging volumes

During initial system setup, GeoSynchrony requires you to create logging volumes (sometimes referred to as dirty region logging volumes or DRLs) for VPLEX Metro and VPLEX Geo configurations to keep track of any blocks changed during a loss of connectivity between clusters. After an inter-cluster link is restored or when a peer cluster recovers, VPLEX uses the resultant bitmap on the logging volume to synchronize distributed devices by sending only the contents of the changed blocks over the inter-cluster link. I/O to the distributed volume is allowed in both clusters while the resynchronization works in the background.

Back-end load balancing

VPLEX uses all paths to a LUN in a round robin fashion thus balancing the load across all paths.

Slower storage hardware can be dedicated for less frequently accessed data and optimized hardware can be dedicated to applications that require the highest storage response.

Data mobility

Data mobility allows you to non-disruptively move data between extents or devices.

Extent migrations move data between extents in the same cluster. Use extent migrations to:

- Move extents from a "hot" storage volume shared by other busy extents
- Defragment a storage volume to create more contiguous free space
- Migrate data between dissimilar arrays

Device migrations move data between devices (RAID 0, RAID 1, or RAID C devices built on extents or on other devices) on the same cluster or between devices on different clusters. Use device migrations to:

- Migrate data between dissimilar arrays
- Relocate a *hot* volume to a faster array
- Relocate devices to new arrays in a different cluster

Figure 18 shows an example of data mobility.

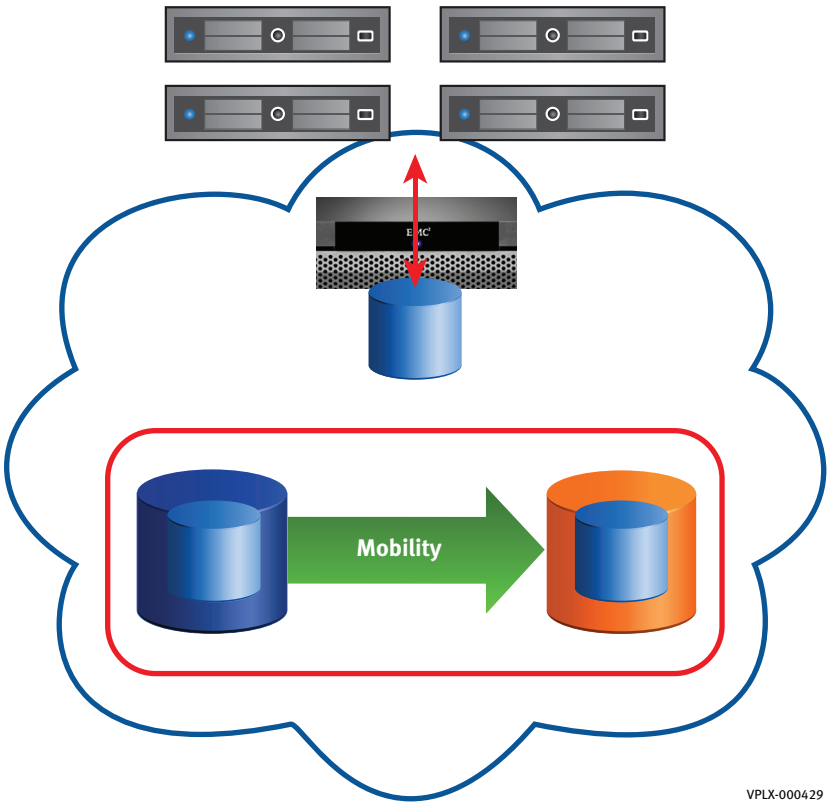

Figure 18 Data mobility

Mirroring

Mirroring is the writing of data to two or more disks simultaneously. If one of the disk drives fails, the system automatically leverages one of the other disks without losing data or service. RAID 1 provides mirroring. VPLEX manages mirroring in the following ways:

- Local mirroring on VPLEX Local for RAID 1 virtual volumes within a data center
- Distributed RAID 1 volumes on VPLEX Metro and VPLEX Geo configurations across data centers

Mirroring is supported between heterogenous storage platforms.

Local mirroring

VPLEX RAID 1 devices provide a local full-copy RAID 1 mirror of a device independent of the host and operating system, application, and database. This mirroring capability allows VPLEX to transparently protect applications from back-end storage array failure and maintenance operations.

RAID 1 data is mirrored using at least two extents to duplicate the data. Read performance is improved because either extent can be read at the same time. Writing to RAID-1 devices requires one write to each extent in the mirror. Use RAID-1 for applications that require high fault tolerance.

Remote mirroring

VPLEX Metro and VPLEX Geo support distributed mirroring that protects the data of a virtual volume by mirroring it between the two VPLEX clusters. There are two types of caching used for consistency groups:

- Write-through caching
- Write-back caching

Distributed volumes with write-through caching

Write-through caching performs a write to back-end storage in both clusters before acknowledging the write to the host. Write-through caching maintains a real-time synchronized mirror of a virtual volume between the two clusters of the VPLEX system providing an RPO of zero data loss and concurrent access to the volume through either cluster. This form of caching is performed on VPLEX Local configurations and on VPLEX Metro configurations. Write-through caching is known as synchronous cache mode in the VPLEX user interface.

Distributed volumes with write-back caching

In *write-back caching*, a director processing the write, stores the data in its cache and also protects it at another director in the local cluster before acknowledging the write to the host. At a later time, the data is written to back end storage. Because the write to back-end storage is not performed immediately, the data in cache is known as dirty data. Write-back caching provides a RPO that could be as short as a few seconds. This type of caching is performed on VPLEX Geo configurations, where the latency is greater than 5ms. Write-back caching is known as asynchronous cache mode in the VPLEX user interface.

Consistency groups

CAUTION

All references to detach rules or preferences in this section describe the behavior of detach rules and preferences in Release 5.0 only.

Consistency groups aggregate volumes together to ensure the common application of a set of properties to the entire group. Create consistency groups for sets of volumes that require the same I/O behavior in the event of a link failure. In the event of a director, cluster, or inter-cluster link failure, consistency groups prevent possible data corruption.

There are two types of consistency groups:

Synchronous consistency groups — provide a convenient way to apply rule sets and other properties to a group of volumes in a VPLEX Local or VPLEX Metro configuration, simplifying system configuration and administration on large systems. Volumes in a synchronous group have global or local visibility. A synchronous consistency group contains either local or distributed volumes. It cannot contain a mixture of local and distributed volumes. Synchronous consistency groups use write-through caching (known as synchronous cache mode in the VPLEX user interface) and are supported on clusters separated by 5ms of latency or less. Synchronous means that VPLEX sends writes to the back-end storage volumes, and acknowledges the writes to the application as soon as the back-end storage volumes in both clusters acknowledge the writes. You can configure up to 1024 synchronous consistency groups. Each synchronous consistency group can contain up to 1000 virtual volumes. The optional VPLEX Witness failure recovery semantics apply *only* to volumes in synchronous consistency groups and only if a rule identifying specific preference is configured.

Asynchronous consistency groups — used for distributed volumes in a VPLEX Geo, separated by up to 50ms of latency. All volumes in an asynchronous consistency group share the same detach rule and cache mode, and behave the same way in the event of an inter-cluster link failure. Detach rules define how each cluster should proceed in the event of an inter-cluster link failure or cluster link failure. Only distributed volumes can be included in an asynchronous consistency group. Asynchronous consistency groups use write-back caching (known as asynchronous cache mode in the VPLEX user interface). This means the director caches each write and then protects the write on a second director in the same cluster. Writes are acknowledged to the host once the write to both directors is complete. These writes are then grouped into deltas. Deltas are exchanged and combined between both clusters before the data is committed to back end storage. Writes to the virtual volumes in an asynchronous consistency group are ordered such that all the writes in a given delta are written before writes from the next delta. However, the writes within an individual data are not ordered. Therefore, if access to the back end array is lost while the system is writing a delta, the data on disk is no longer consistent and requires automatic recovery when access is restored. Asynchronous cache mode can give better performance, but there is a higher risk that data will be lost if:

- Multiple directors fail at the same time
- There is an inter-cluster link partition and both clusters are actively writing and instead of waiting for the link to be restored, the user chooses to accept a data rollback in order to reduce the RTO
- The cluster that is actively writing fails

VPLEX supports a maximum of 16 asynchronous consistency groups. Each asynchronous consistency group can contain up to 1000 virtual volumes.

Synchronous consistency groups

A set of LUNs for a database application requires a consistent image on disk at all times. Figure 19 on page 59 shows a synchronous consistency group that spans two clusters in a VPLEX Metro configuration. The hosts at both clusters write to the VPLEX distributed volumes in the consistency group. VPLEX writes data to the back-end storage on both clusters before acknowledgment is returned to the host issuing the write. This guarantees that the image on the back end storage is an exact copy on both sides.

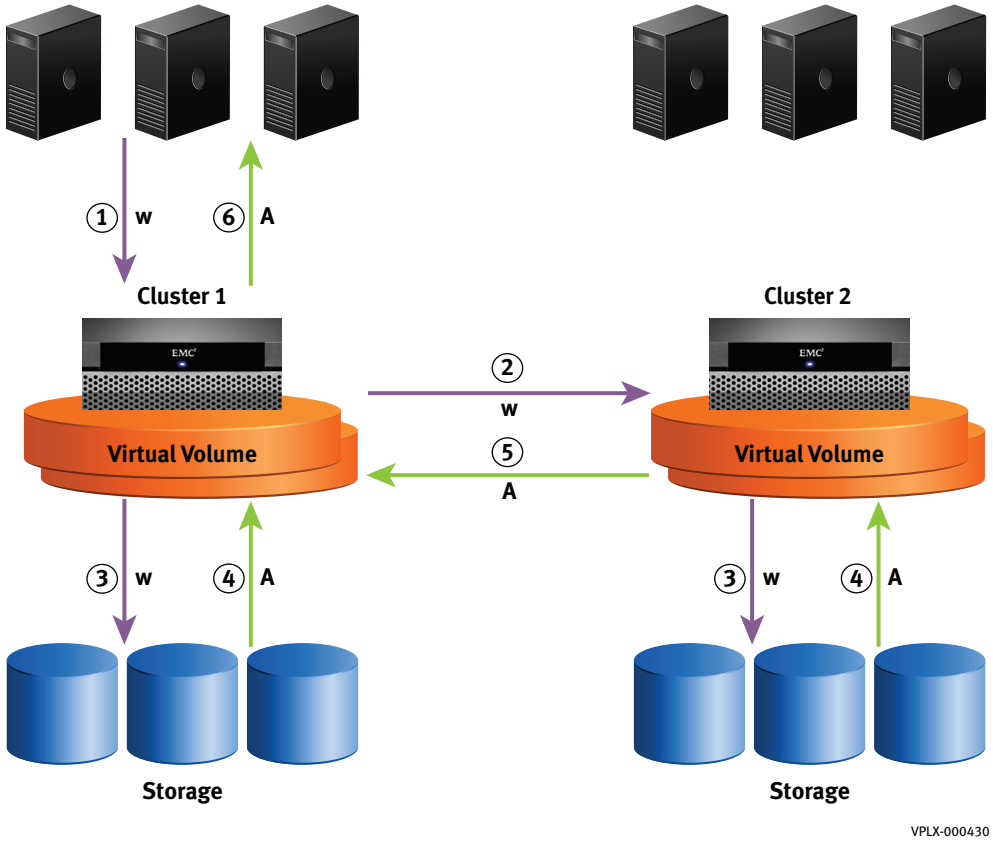

Figure 19 Synchronous consistency group

Local consistency group

A local consistency group is a logical grouping of volumes grouped together to organize the volumes. I/O performed on local consistency groups is always synchronous. This means that I/O must be written through VPLEX to the back-end cache of the storage array prior to sending the acknowledgement to the host. Local consistency groups can only be read and written by their local cluster unless they are local consistency groups with global visibility. Figure 20 on page 60 shows a local consistency group.

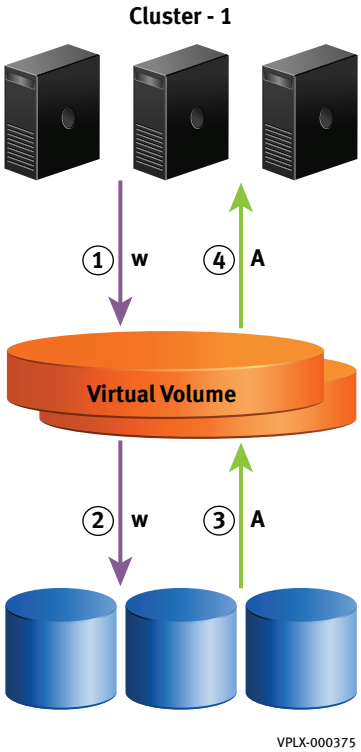

Figure 20 **Local consistency groups**

Local consistency groups with global visibility

Volumes in local consistency groups that are visible at both clusters can receive I/O from the cluster that does not have a local copy. However, all writes from that remote cluster must pass over the inter-cluster WAN link before they are acknowledged. Any reads that cannot be serviced from local cache must also be transferred across the link. This allows the remote cluster to have instant on-demand access to the consistency group, but also adds additional latency for the remote cluster. Local consistency groups with global visibility are supported in VPLEX Metro environments. Only local volumes can be placed into the local consistency group with global visibility. Local consistency groups with global visibility always use write-through cache mode (known as synchronous cache mode in VPLEXcli). I/O that goes to local consistency groups with global visibility will always be synchronous.

Figure 21 on page 61 shows a local consistency group with global visibility.

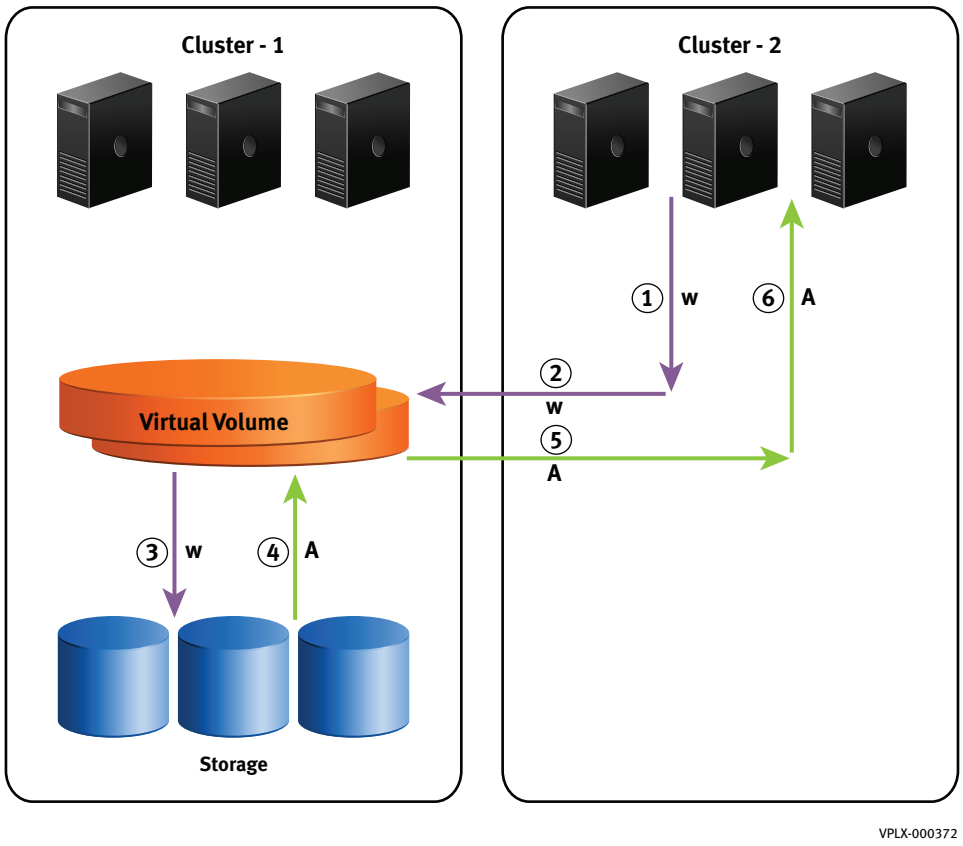

Figure 21 Local consistency groups with global visibility

Distributed synchronous consistency group

A distributed synchronous consistency group in VPLEX Metro differs from a local consistency group in that the physical storage resides at both clusters. When writing to this type of synchronous consistency group, I/O must be written to the array at both clusters prior to sending the acknowledgement to the host. In this example, the host writes to the VPLEX at Cluster 1 and the write passes through VPLEX at Cluster 1 to the back-end array. It is also sent across the inter-cluster WAN link to the remote cluster where it is sent to the back-end storage at the remote cluster. The acknowledgement is sent back to both VPLEX clusters from the arrays. It is also sent from the remote VPLEX to the local cluster where the local cluster sends an acknowledgement to the host. Synchronous consistency groups can be created in VPLEX Metro. Only distributed volumes can be added to distributed synchronous consistency groups.

Asynchronous consistency groups

A VPLEX Metro allows synchronous I/O only, with a round-trip delay (RTT) of up to 5ms between clusters. In a VPLEX Geo where clusters can be further apart, a higher round-trip delay would significantly affect synchronous I/O to distributed volumes since host I/O must be acknowledged at both clusters before being written to the back end. Because a VPLEX Geo allows up to 50 ms of round-trip delay between clusters, asynchronous I/O is used for distributed volumes.

An asynchronous consistency group differs from a synchronous consistency group in that the host receives acknowledgment after the write reaches the VPLEX cache and has been protected to another director in the local cluster. VPLEX collects writes at the cluster in the form of a *delta*. At a later point in time, the clusters exchange deltas

creating a globally merged delta. The clusters send communication messages back and forth between each other to facilitate the exchange. Once the exchange is complete, the clusters write a global merged delta to the back-end arrays. Asynchronous consistency groups are only supported in VPLEX Geo deployments and the maximum round trip latency between clusters must be 50 ms or less.

Figure 22 shows a asynchronous consistency group configuration in an active/passive phase.

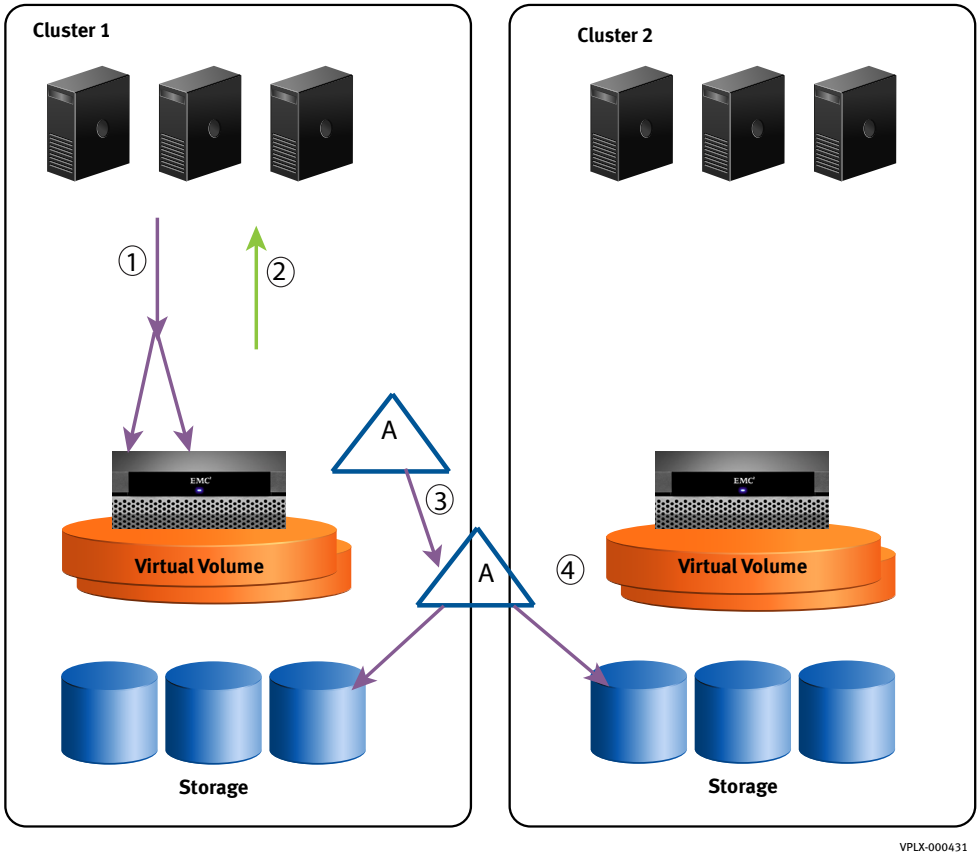

Figure 22 Asynchronous consistency group active/passive

In Figure 22, only one cluster is actively reading and writing. This simplifies the view of asynchronous I/O. In this case application data is written to the director in Cluster 1 and protected in another director of Cluster 1. When the write to the director is complete, the host application is acknowledged. VPLEX collects these writes into a delta of a fixed size. Once that delta is filled or when a set time period has elapsed, the two clusters of the VPLEX Geo begin a communication process to exchange deltas. The combination of the deltas is referred to as a global delta. In Figure 22, the global delta only includes the writes that occurred on Cluster 1 because Cluster 2 was inactive. This data is then written to the back-end storage at Cluster 1 and Cluster 2.

Figure 23 on page 63 shows the same process when both clusters are active.

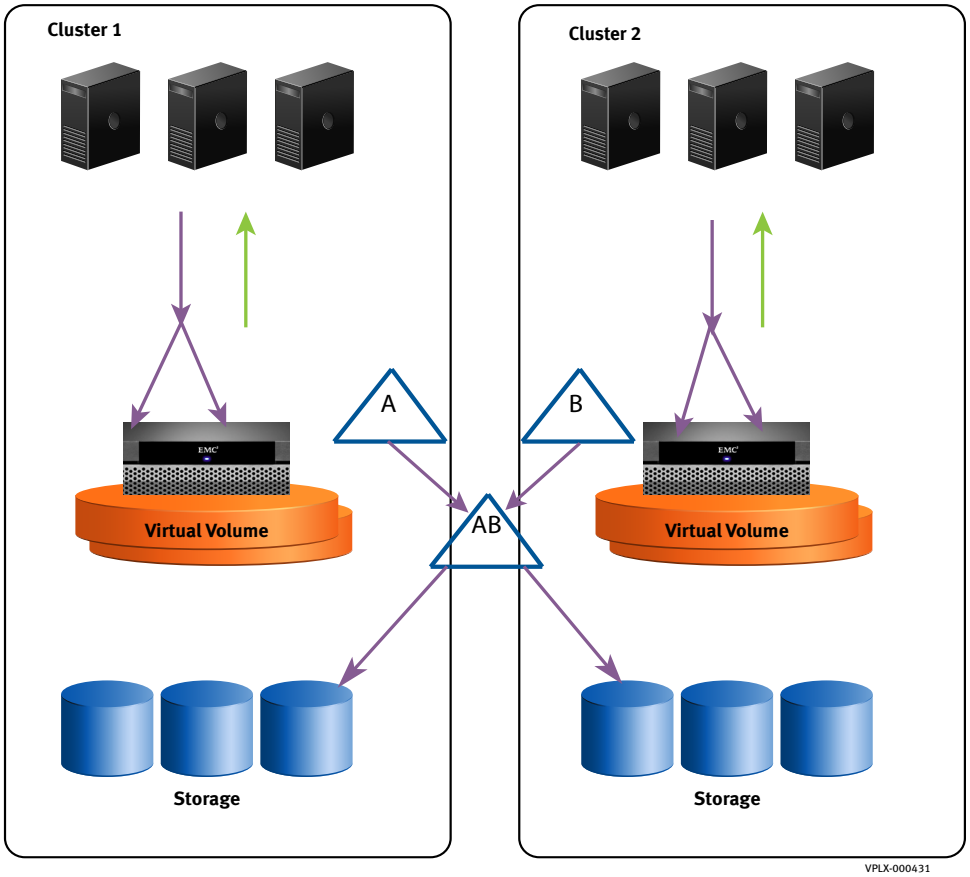

Figure 23 **Asynchronous consistency group active/active**

Figure 23 shows asynchronous I/O when both clusters are actively reading and writing. The applications at Cluster 1 and Cluster 2 are both writing to their local VPLEX cluster. Again, the application is acknowledged once each cluster caches the data in two directors. The VPLEX collects the data in deltas at each cluster. After a time period, or after one of these deltas becomes full, the clusters begin an exchange of deltas. VPLEX then writes the resulting delta to back end storage at each cluster.

This process coordinates data written to the storage at each cluster. At any given time there is only one open delta, one exchanging delta, and one global delta being written to back-end storage. There can be multiple closed deltas waiting to be exchanged.

There are additional properties associated with asynchronous groups that are used to configure deltas to control the recovery point objective in the event of a failure. The *recovery point objective (RPO)* is the maximum acceptable level of data loss resulting from a failure, and represents the point in time (prior to the failure) to which lost data can be recovered. For example, by setting the maximum queue depth property and the closure timer you can control the RPO. However both must be set sufficiently high to accommodate the latency and usable bandwidth between clusters.

Detach rule

Detach rules are predefined rules that determine which cluster continues I/O during an inter-cluster link failure or cluster failure. In these situations, until communication is restored, most I/O workloads require specific sets of virtual volumes to resume on one cluster and remain suspended on the other cluster unless the no-active-winner rule is used, in which case both clusters will suspend.

In the event of connectivity loss with the remote cluster, the detach rule defined for each consistency group identifies a preferred cluster (if there is one) that can resume I/O to the volumes in the consistency group. In a VPLEX Metro configuration, I/O proceeds on the preferred cluster and is suspended on the non-preferred cluster. In a VPLEX Geo configuration, I/O proceeds on the active cluster only when the remote cluster has no dirty data in cache.

Refer to the "Consistency Groups" chapter in the *VPLEX with GeoSynchrony 5.0 and Point Releases CLI Guide* for the available detach rules for synchronous and asynchronous consistency groups.

Active and passive clusters

Asynchronous consistency groups have the *active-cluster-wins rule*. When using the active-cluster-wins rule, I/O continues at the cluster where the application was actively writing last (provided there was only one such cluster). The active cluster is the preferred cluster.

An *active cluster* has data in cache that has yet to be written to the back-end storage. This data is referred to as *dirty data*. A *passive* cluster refers to a cluster that has no dirty data in its cache. If both clusters were active during the failure, I/O must suspend at both clusters. *I/O suspends because the cache image is inconsistent on both clusters and must be rolled back to a point where both clusters had a consistent image to continue I/O*. Application restart is required after roll back. If both clusters were passive and have no dirty data at the time of the failure, the cluster that was the last active one (before it became passive) will proceed with I/O after failure. Regardless of the detach rules in Asynchronous consistency groups, as long as the remote cluster has dirty data, the local cluster suspends I/O if it observes loss of connectivity with the remote cluster regardless of preference. This is done to allow the administrator for the application time to stop or restart the application prior to exposing the application to the rolled back, time consistent, data image. It might also be possible for you to recover the inter-cluster link and recover without having to perform a rollback.

Note: VPLEX Witness has no bearing on the failover semantics of asynchronous consistency groups. VPLEX Witness still provides its guidance (which can be used for diagnostic purposes later on) but it does not affect actual failover.

The GUI shows which clusters are *active* in an asynchronous consistency group, or which clusters have data in cache that has not been written to the back-end array. It also provides information on the state of the cluster and if you need to run a recovery command to enable access to the consistency group. You can also query these states through the CLI.

Data loss failure mode (DLFM)

VPLEX Metro and VPLEX Local have been designed to leverage write-through caching in order to ensure a crash consistent data image on disk in the presence of various failures. On the other hand, VPLEX Geo leverages write-back caching, which means that the image on disk is not crash consistent during the write phase when the cluster is writing out the exchanged deltas. In some extremely rare system-wide multi-failure scenarios the system may enter Data Loss Failure Mode, which indicates data loss and potential loss of crash consistency on disk in both clusters. This state applies to individual asynchronous consistency groups; that is, while one asynchronous consistency group may be in Data Loss Failure Mode, others on the same VPLEX Geo system may be unaffected.

Best practices to avoid DLFM

To avoid DLFM, follow these best practices when configuring your system:

- Mirror local DR1 legs to two local arrays with a RAID-1. This would minimize the risk of back-end visibility issues in the presence of array failures

- Follow best practices for high availability when deploying VPLEX Geo See "Path redundancy" on page 76 for more information.

Recovering from DLFM

Refer to the troubleshooting procedure for recovering from DLFM in the Procedure Generator.

Cache vaulting

Cache Vaulting is necessary in VPLEX Geo configurations to safeguard the dirty cache data under emergency conditions. *Dirty cache pages* are pages in a director's memory that have not been written to back-end storage but were acknowledged to the host. Dirty cache pages also include the copies protected on a second director in the cluster. These pages must be preserved in the presence of power outages to avoid loss of data already acknowledged to the host.

Note: Cache vault is requires releases of GeoSynchrony 5.0.1 or greater.

CAUTION

Although there is no dirty cache data in VPLEX Local or VPLEX Metro configurations, vaulting is still necessary to quiesce all I/O when power failure has been detected. This is done to minimize the risk of metadata corruption.

When the system recovers, VPLEX can unvault this vaulted data, avoiding any data loss.

For information on distributed cache protection, refer to Chapter 4, "System Integrity and Resiliency." For information on conditions that cause a vault see Chapter 2, "VPLEX Hardware Overview."

Vaulting can be used in two scenarios:

- *Cluster-wide power failure*: VPLEX monitors all components that provide power to the individual engines. If it detects AC power loss, in any component, in both power zones simultaneously, it takes a conservative approach to avoid data loss; it promotes the power loss condition to a cluster wide power failure, which triggers cluster wide vaulting if the power loss exceeds 30 seconds. This condition is triggered if power is lost in power zone A from engine X and power zone B from engine Y (X and Y would be the same in a single engine configuration but they may or may not be the same in dual or quad engine configurations).

Note: Power failure of the UPS (in dual and quad engine configurations) does not currently trigger any vaulting actions. Because all monitored components across all engines are fed by the same input power source in each power zone, with the exception of hardware failure or physically disconnecting a component from the power feed, power loss detected in any component implies power loss in all components within the same power zone.

WARNING

WARNING: When performing maintenance activities, service personnel must not remove the power in one or more engines that would result in the power loss of both power zones (as described above) unless both directors in those engines have been shutdown and are no longer monitoring power. Failure to do so, will lead to data unavailability in the affected cluster.

- *Manual emergency cluster shutdown:* When unforseen circumstances require an unplanned and immediate shutdown, it is known as an *emergency cluster shutdown.* You can use a CLI command to manually start vaulting if an emergency shutdown is required.

For information on the redundant and backup power supplies in VPLEX, refer to Chapter 2, "VPLEX Hardware Overview."

Figure 24 on page 67 shows the process VPLEX uses when it experiences an interruption in power or when an administrator manually initiates a vault.

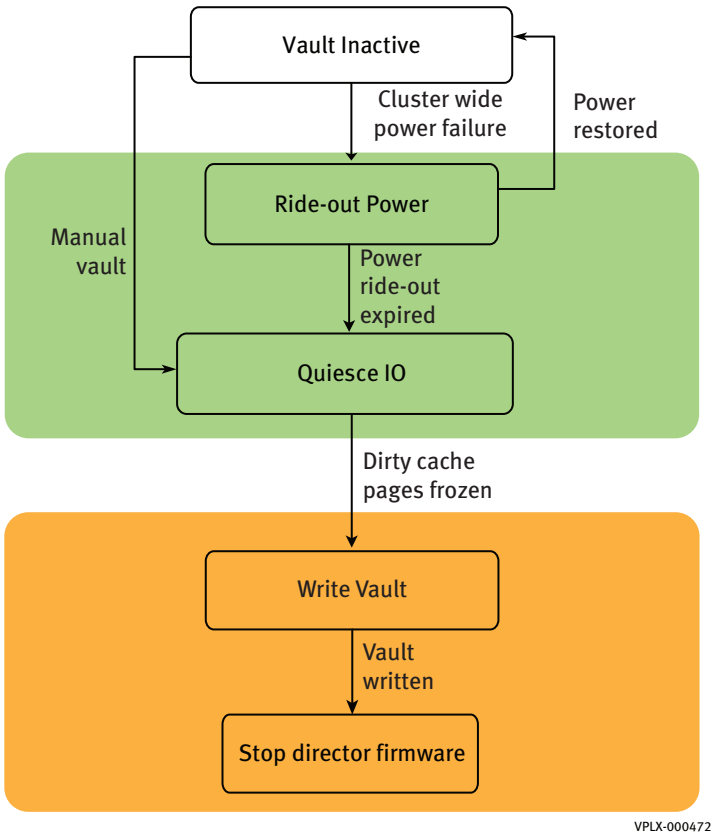

Figure 24 Cache vaulting process flow

When a cluster detects AC power loss in any component, in both power zones simultaneously, VPLEX triggers the cluster to enter a 30 second ride-out phase. This delays the (irreversible) decision to vault, allowing for a timely return of AC input to avoid vaulting altogether. During the ride-out phase, all mirror rebuilds and migrations pause, and the cluster disallows new configuration changes on the local cluster, to prepare for a possible vault.

If the power is restored prior to the 30 second ride-out, all mirror rebuilds and migrations resume, and configuration changes are once again allowed.

If the power is not restored within 30 seconds, the cluster begins vaulting. Power restoration after this time will *not* stop the vaulting process.

Power ride-out is not necessary when a manual vault has been requested. However, similar to the power ride-out phase, manual vaulting stops any mirror rebuilds and migrations and disallows any configuration changes on the local cluster.

Once the cluster has decided to proceed with vaulting the dirty cache, the vaulting cluster quiesces all I/Os and disables inter-cluster links to isolate itself from the remote cluster. These steps are required to freeze the director's dirty cache in preparation for vaulting.

Once the dirty cache (if any) is frozen, each director in the vaulting cluster isolates itself from the other directors and starts writing. When finished writing to its vault, the director stops its firmware.

This entire process is completed within the time parameters supported by the stand-by power supplies.

It is recommended that you follow the cluster shutdown procedure to ensure that the cluster is shutdown in an organized fashion and to save any remaining battery charge so that recharge will complete faster when the cluster is once again powered on. Refer to the Procedure Generator for the shutdown procedure for VPLEX.

Recovery after vault

Once the cluster is ready to recover from the conditions that caused the vault, the cluster is powered up. Figure 25 shows the process used to unvault during a recovery process.

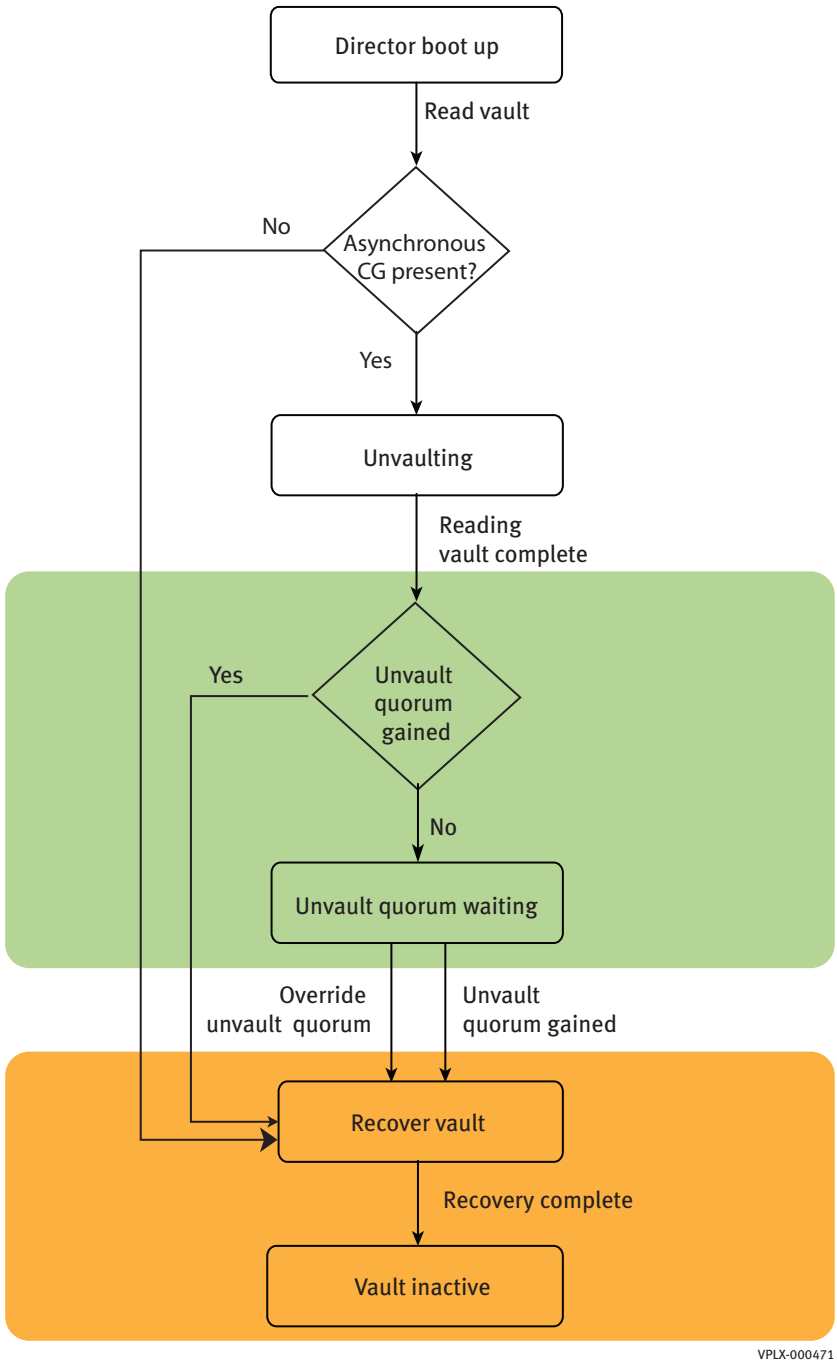

Figure 25 Unvaulting cache process

At the start of cluster recovery, VPLEX checks to see if there are any configured asynchronous consistency groups. If there are none (as would be the case in VPLEX Local and VPLEX Metro configurations), the entire unvault recovery process is skipped.

As the directors boot up, each director reads the vaulted dirty data from its respective vault disk. Once the directors have completed reading the vault, each director evaluates if the unvaulted data can be recovered.

The cluster then evaluates if it has gained unvault quorum. The *unvault quorum* is the set of directors that had vaulted their dirty cache data during the last successful cluster wide vault. In order to recover their vaulted data, which is required to preserve cache coherency and avoid data loss, these directors must boot and rejoin the cluster. If the unvault quorum is achieved the cluster proceeds to recover the vault.

If the cluster determines that it has not gained unvault quorum, it waits indefinitely for the required directors to boot up and join the cluster. During the waiting period, the cluster remains in the *unvault quorum wait state*.

After 30 minutes in the unvault quorum wait state, the cluster generates a call home indicating the current state of the cluster and indicating that manual intervention is needed to allow the cluster to process I/O.

Once the cluster enters the unvault quorum wait state it cannot proceed to the recovery phase until any of the following events happen:

- The directors required to gain unvault quorum become available
- You issue the **override unvault quorum** command and agree to accept a possible data loss

Refer to the VPLEX Procedure Generator troubleshooting procedure for cache vaulting for instructions on how recover in this scenario. See the *VPLEX with GeoSynchrony 5.0 and Point Releases CLI Guide* for details on the use of **override unvault quorum** command.

Successful Recovery

VPLEX Geo can handle one invalid or missing vault because each director has vaulted a copy of each dirty cache page of its protection partner. The cache can be recovered as long as the original dirty cache vault or its protected copy is available.

An *invalid* vault can be caused by:

- A director not successfully completing write of the vault
- A director containing stale (older) vault data

A vault can be *missing* because:

- A director failed during unvault
- A director never came up during cluster power up

If the cluster determines it has sufficient valid vaults, it proceeds with recovery of the vaulted data into the distributed cache. In this scenario the unvaulted cluster looks like a cluster that has recovered after an inter-cluster link failure as no data is lost on the vaulting cluster. VPLEX behavior following this recovery process depends on how the detach rules were configured for each asynchronous consistency group.

Refer to the "Consistency Group" chapter in the *VPLEX with GeoSynchrony 5.0 and Point Releases CLI Guide*.

Unsuccessful Recovery

If the cluster determines that more than one invalid vault is present, the cluster discards the vault and reports a data loss. In this scenario the unvaulted cluster looks like a cluster that has recovered after a cluster failure. The cluster still waits for all

configured directors to boot and rejoin the cluster. The volumes are marked as **recovery-error** and refuse I/O. If one volume of a consistency group is marked **recovery-error**, all other volumes of that consistency group must also refuse I/O.

4

System Integrity and Resiliency

VPLEX provides numerous high availability and redundancy features for servicing I/O. The following features allow robust system integrity, and resiliency.

Overview

VPLEX clusters are capable of surviving any single hardware failure in any subsystem within the overall storage cluster. These include the host connectivity and memory subsystems. A single failure in any subsystem does not affect the availability or integrity of the data. Multiple failures in a single subsystem and certain combinations of single failures in multiple subsystems might affect the availability or integrity of data.

VPLEX features fault tolerance for devices and hardware components to continue operation as long as one device or component survives. This highly available and robust architecture can sustain multiple device and component failures while servicing storage I/O.

VPLEX configurations continue to service I/O in the following classes of faults and service events:

- Unplanned and planned storage outages
- SAN outages
- VPLEX component failures
- VPLEX cluster failures
- Data center outages

This availability requires that you create redundant host connections and supply hosts with multi path drivers. In the event of a front-end port failure or a director failure, hosts without redundant physical connectivity to a VPLEX cluster and without multipathing software installed could be susceptible to data unavailability.

Cluster

A cluster is a collection of one, two, or four engines in a physical cabinet. A cluster serves I/O for one site.

All hardware resources (CPU cycles, I/O ports, and cache memory) are pooled.

Configurations of two clusters in a VPLEX Metro or VPLEX Geo topology provide higher resilience against a site-wide outage.

Path redundancy

The following sections discuss the resilience produced by multiple paths. They include examples of the following paths:

- Path redundancy through different ports
- Path redundancy through different directors
- Path redundancy through different engines
- Path redundancy through site distribution

Different ports

The front-end ports on all directors can provide access to any virtual volume in the cluster. Including multiple front end ports in each storage view protects against port failures. When a director port goes down for any reason, the host multipathing software will seamlessly fail over to another path through a different port, as shown in Figure 26.

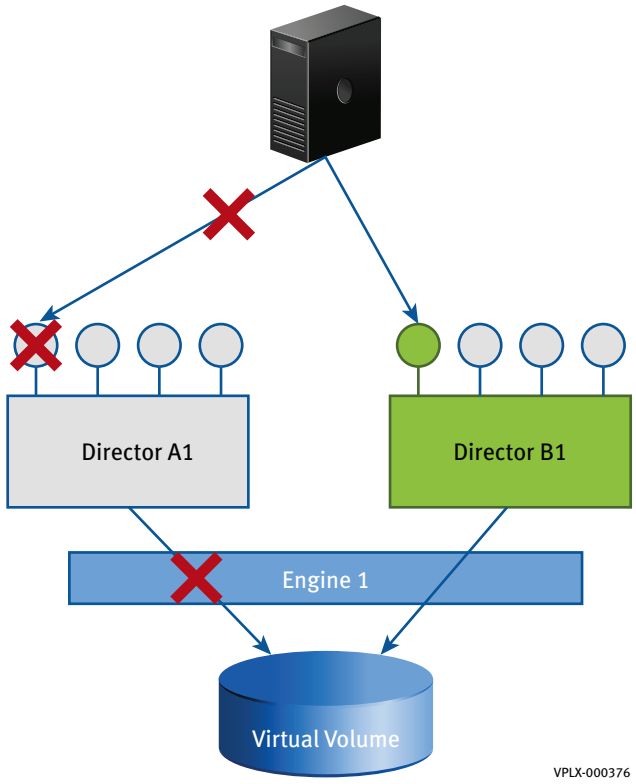

Figure 26 Port redundancy

Multi-pathing software plus redundant volume presentation yields continuous data availability in the presence of port failures.

Different directors

If a director were to go down, the other director can completely take over the I/O processing from the host, as shown in Figure 27.

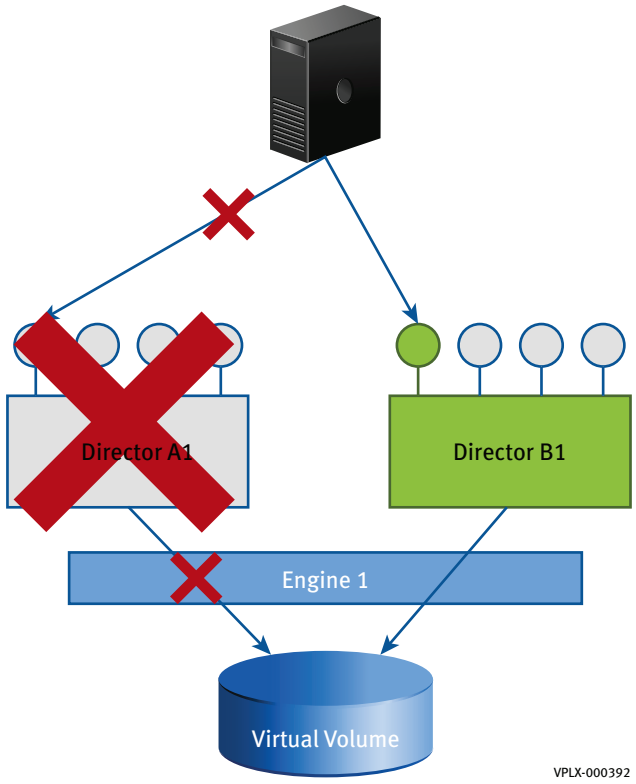

Figure 27 Director redundancy

Multi-pathing software plus volume presentation on different directors yields continuous data availability in the presence of director failures.

Each director can service I/O for any other director in the cluster due to the redundant nature of the global directory and cache coherency.

Best practices

For maximum availability, present virtual volumes through each director so that all directors but one can fail without causing data loss or unavailability. Connect all directors to all storage. To have continuous I/O during a non-disruptive upgrade of VPLEX, it is critical to have a path through an A director and a path through a B director.

When a pair of redundant Fibre Channel fabrics is used with VPLEX, VPLEX directors should be connected to each fabric both for the front-end (host-side) connectivity, as well as for the back-end (storage array side) connectivity. This deployment, along with the isolation of the fabrics, allows the VPLEX system to ride through failures that take out an entire fabric, and allows the system to provide continuous access to data through this type of fault.

Hosts must be connected to both fabrics and use multi-pathing software to ensure continuous data access in the presence of such failures.

It is recommended that I/O modules be connected to redundant fabrics.

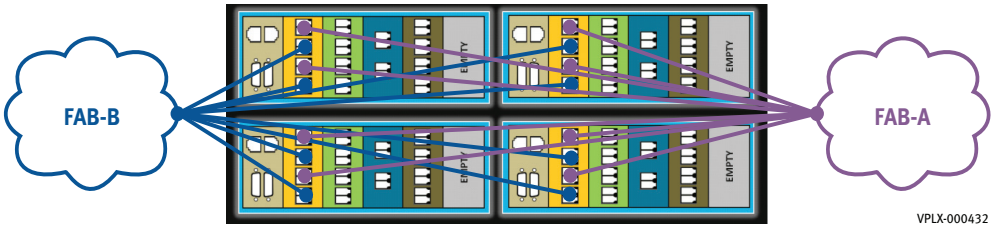

Figure 28 **Recommended fabric assignments for front-end and back-end ports**

Different engines

In a dual- or quad-engine environments on VPLEX Metro, if one engine goes down, another engine completes the host I/O processing, as shown in Figure 29.

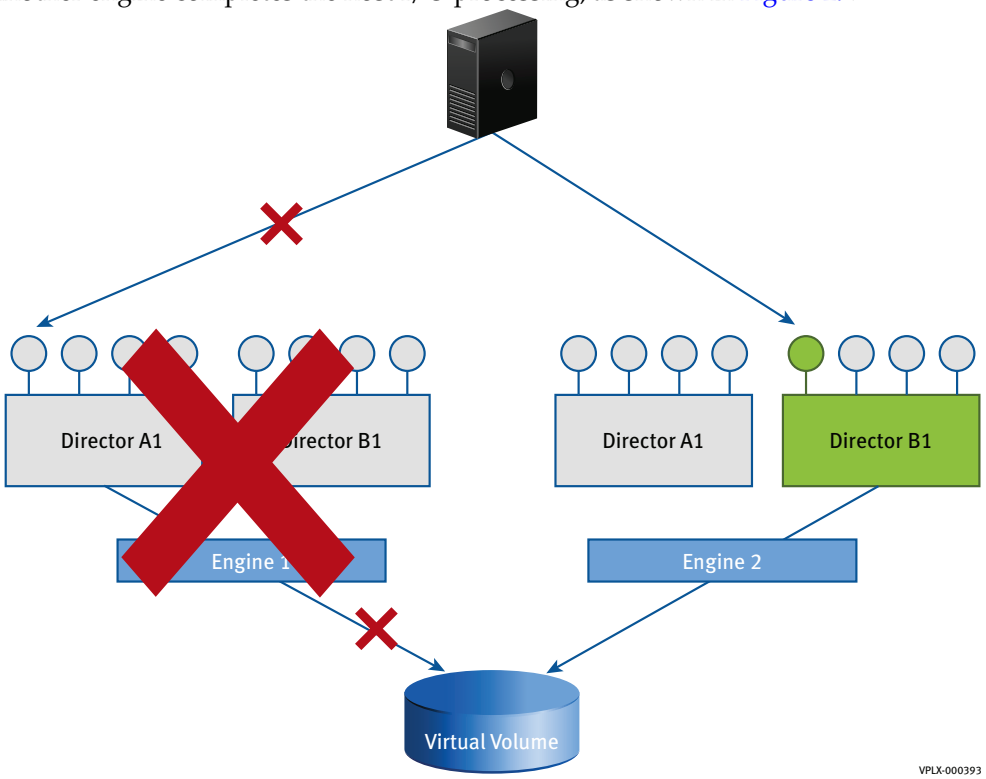

Figure 29 **Engine redundancy**

In VPLEX Geo, directors in the same engine serve as protection targets for each other. If a single director in an engine goes down, the remaining director uses another director in the cluster as its protection pair. Simultaneously losing an engine in an active cluster, though very rare. could result in DLFM. However, the loss of 2 directors in different engines can be handled as long as other directors can serve as protection targets for the failed director. For more information about DLFM, see Chapter 3, "VPLEX Software."

Multi-pathing software plus volume presentation on different engines yields continuous data availability in the presence of engine failures on VPLEX Metro.

Site distribution

VPLEX Metro ensures that if a data center goes down, or even if the link to that data center goes down, the other site can continue processing the host I/O, as shown in Figure 30. On site failure of Data Center B, the I/O continues unhindered in Data Center A.

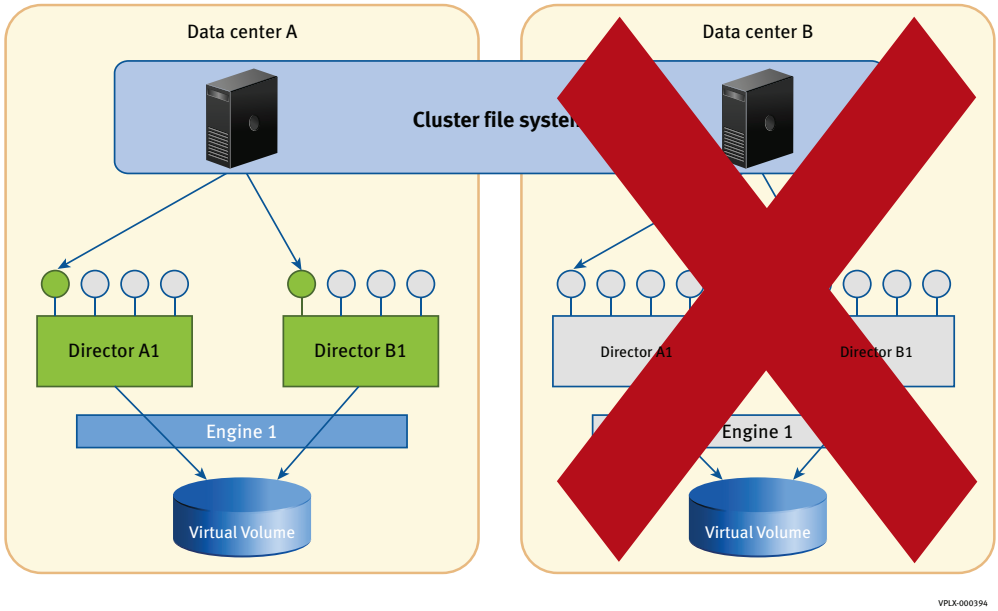

Figure 30 Site redundancy

Optionally, you can install the VPLEX Witness on a server in a separate failure domain to provide further fault tolerance in VPLEX Metro configurations. See See "High Availability through VPLEX Witness" on page 80 for more information.

High Availability through VPLEX Witness

VPLEX GeoSynchrony 5.0 systems running in a VPLEX Metro or VPLEX Geo configuration can now rely on an optional component called VPLEX Witness. VPLEX Witness is designed to be deployed in customer environments where the regular detach rule sets alone provide insufficient recovery time objective fail-over in the presence of VPLEX cluster failures or inter cluster link partitions.

- In aVPLEX Metro configuration, VPLEX Witness provides seamless zero RTO fail-over for synchronous consistency groups (from a storage perspective) in the presence of these failures.
- In a VPLEX Geo configuration, VPLEX Witness can be useful for diagnostic purposes.

CAUTION

VPLEX Witness does not automate any fail-over decisions for asynchronous consistency groups in Release 5.0.1.

VPLEX Witness connects to both VPLEX clusters over the management IP network. By reconciling its own observations with the information reported periodically by the clusters, the VPLEX Witness enables the clusters to distinguish between inter-cluster network partition failures and cluster failures and automatically resume I/O in these situations.

VPLEX Witness installation considerations

The responsibility of VPLEX Witness is to provide improved availability in your storage network. For this reason, you should carefully consider how you install VPLEX Witness with your VPLEX Metro clusters.

WARNING

VPLEX Witness must be deployed in a failure domain independent from either of VPLEX clusters. If this requirement cannot be met, the VPLEX Witness should not be installed.

An external VPLEX Witness is deployed as a virtual machine running on a customer supplied VMware ESX server deployed in a failure domain separate from either of the VPLEX clusters (to eliminate the possibility of a single fault affecting both the cluster and the VPLEX Witness). The VPLEX Witness software is also loaded as a client on each of the clusters of a VPLEX Metro or VPLEX Geo configuration. Each of the clusters in this configuration should also reside in separate failure domains from each other. A failure domain is a set of entities effected by the same set of faults. The scope of the failure domain depends on the set of fault scenarios that must be tolerated in a given environment. For example, if the two clusters of a VPLEX Metro configuration are deployed on two different floors of the same data center, deploy the VPLEX Witness on a separate floor. On the other hand, if the two clusters of a VPLEX Metro configuration are deployed in two different data centers, deploy the VPLEX Witness in the third data center.

VPLEX Metro failure without VPLEX Witness

Without VPLEX Witness, all VPLEX Metro synchronous consistency groups and all distributed volumes rely on configured rule sets to identify the preferred cluster in the presence of cluster partition or cluster failure. However, if the preferred cluster happens to fail (in the result of a disaster event, or similar condition), VPLEX Metro is unable to automatically fail-over and allow the non-preferred cluster to continue I/O

to the affected distributed volumes. VPLEX Witness has been designed specifically to solve this problem for synchronous consistency groups configured with a specific preference rule set[1].

CAUTION

All references to preference rule set or detach rules in this chapter refer to the behavior of Release 5.0.1. Behaviors in Release 5.0 are different than those described here.

In all of the scenarios shown in Figure 31, the synchronous consistency group is configured with rule set that designates Cluster 1 as preferred and allows it to continue if there is a failure in one of the clusters or in the inter-cluster link.

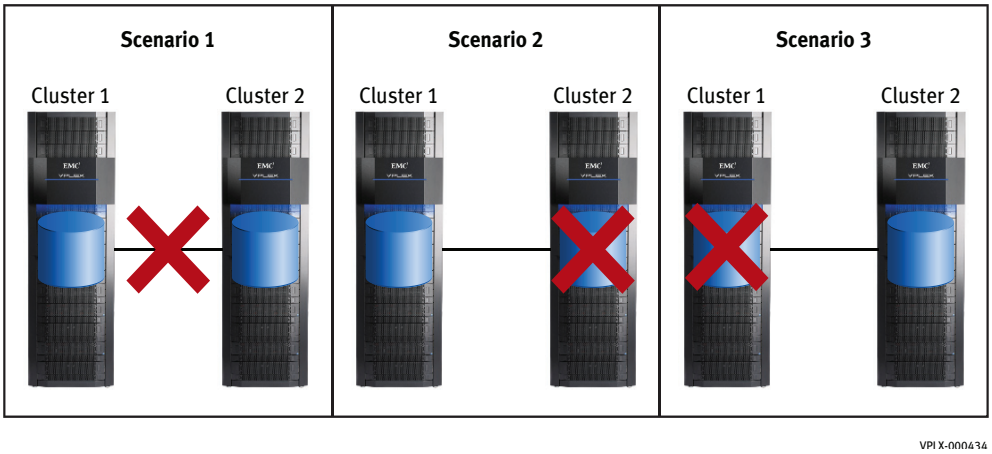

Figure 31 VPLEX failure recovery scenarios in VPLEX Metro configurations

In Scenario 1, the inter cluster link goes down. In this case, Cluster 1 continues I/O and Cluster 2 suspends I/O. This enables Cluster 1 to continue service (to avoid data unavailability) and Cluster 2 to suspend to avoid split brain. Once the communication link is restored, both clusters can continue I/O while Cluster 2 updates its data to match that of Cluster 1 in the background.

In Scenario 2, Cluster 2 fails. Because the rule set was configured to allow Cluster 1 to continue I/O anyway, the rule set is effective in this case.

In Scenario 3, Cluster 1, the preferred cluster fails. The rule set is configured to suspend I/O at Cluster 2 in the event of a cluster failure. Without VPLEX Witness, VPLEX has no automatic way of recovering from this failure and it suspends I/O at the only operating cluster. In some cases, the failed cluster could recover, in which case recovery is actually automatic. Otherwise, manual intervention is required to re-enable I/O on Cluster 2. This is the scenario that VPLEX Witness solves for VPLEX Metro configurations.

Note: VPLEX Witness has no impact on distributed volumes in synchronous consistency groups configured with the no-automatic-winner rule. In that case, manual intervention is required in the presence of any failure scenario described above.

1. If a synchronous consistency group is configured with the no-winner rule, each cluster suspends if it loses contact with its peer cluster regardless of whether VPLEX Witness is deployed or not.

Cluster failures in the presence of VPLEX Witness

Note: The discussion in this section assumes that VPLEX Witness is enabled in your configuration. If VPLEX Witness is disabled, it has no impact on the fail-over semantics.

Figure 32 shows four failure types that could occur.

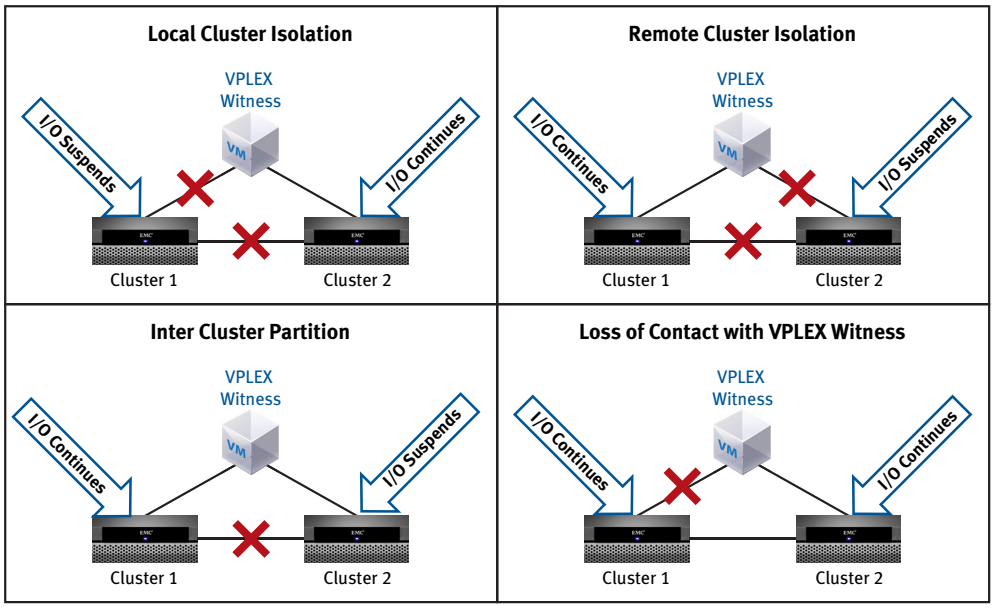

Figure 32 Failures in the presence of VPLEX Witness

Note: In this figure, the terms local and remote are from the perspective of Cluster 1.

Local Cluster Isolation occurs when the local cluster loses contact with both the remote cluster and VPLEX Witness. If the example shown is a VPLEX Metro configuration, Cluster 1 (the local cluster) unilaterally suspends I/O and the VPLEX Witness guides Cluster 2 to continue I/O, unless the synchronous consistency group is configured with the no-automatic-winner rule set. VPLEX Witness has no impact on synchronous consistency groups configured with the no-automatic-winner rule set. If the example shown is a VPLEX Geo configuration with asynchronous consistency groups, the remote cluster disregards the guidance of VPLEX Witness and failover is handled according to the configured rule set.

In the Remote Cluster Isolation scenario, the local cluster (Cluster 1) has lost contact with the remote cluster *and* with the VPLEX Witness. However, the Cluster 1 still has access to the VPLEX Witness. If the example shown is a VPLEX Metro configuration, Cluster 1 continues I/O as it is still in contact with the VPLEX Witness. Cluster 2 suspends I/O. VPLEX Witness has no impact on synchronous consistency groups configured with the no-automatic-winner rule set. If the example shown is a VPLEX Geo configuration with asynchronous consistency groups and there is no data loss, VPLEX disregards the guidance of VPLEX Witness and failover is handled according to the configured rule set.

In the case of an Inter-Cluster Partition where both clusters lose contact with each other but still have access to the VPLEX Witness the action taken by the clusters depends on the type of VPLEX configuration and the detach rule configured.

CAUTION

In a VPLEX Metro, if the consistency group is configured with the no-automatic-winner rule set, VPLEX Witness has no impact.

- In a VPLEX Metro configuration with synchronous consistency groups and any other rule set, I/O continues on the preferred cluster.
- In a VPLEX Metro with synchronous consistency groups or VPLEX Geo configuration with asynchronous consistency group, if the preferred cluster can not proceed because it has not fully synchronized, the cluster suspends I/O.
- In a VPLEX Geo configuration with asynchronous consistency groups and active writer detach rules configured, I/O continues on the active writer cluster provided the other cluster has no dirty data and the local storage is not out of date. VPLEX disregards the guidance of VPLEX Witness and the failover semantics.

VPLEX Witness always preserves detach rule semantics in case of inter-cluster network partition.

In the scenario Loss of Contact with the VPLEX Witness, if the clusters are still in contact with each other, there is no change in I/O. The cluster that lost connectivity with VPLEX Witness issues a Call Home.

In the case of an inter-cluster partition, VPLEX Witness preserves the semantics of the detach rule. In the case of a cluster failure, VPLEX Witness on a VPLEX Metro configuration might override the detach rule. Overriding the detach rule results in a zero RTO policy for VPLEX Metro.

If the VPLEX Witness fails, both clusters Call Home. As long as both clusters remain connected with each other, there is no impact on I/O. However, if either of the clusters fails or if the inter-cluster link were to fail while the VPLEX Witness is down, VPLEX experiences data unavailability in all surviving clusters.

CAUTION

Its is important to evaluate the failure of VPLEX Witness or its connectivity carefully. If the VPLEX Witness is expected to be unreachable for a prolonged interval of time, disable the VPLEX Witness functionality (refer to *EMC VPLEX with GeoSynchrony 5.0 and Point Releases CLI Guide*) and use preconfigured detach rules in the interim.

Once the VPLEX Witness or its connectivity recovers, you can re-enable it.

When the VPLEX Witness observes a failure and provides its guidance it sticks to this guidance until both clusters report complete recovery. This is crucial in order to avoid split-brain and data corruption. As a result you may have a scenario where Cluster 1 becomes isolated and the VPLEX Witness tells Cluster 2 to continue I/O and then Cluster 2 becomes isolated. However, because Cluster 2 has previously received guidance to proceed from the VPLEX Witness, it proceeds even while it is isolated. In the meantime, if Cluster 1 were to reconnect with the VPLEX Witness server, the VPLEX Witness server tells it to suspend. In this case, because of event timing, Cluster 1 is connected to VPLEX Witness but it is suspended, while Cluster 2 is isolated but it is proceeding.

VPLEX Witness in VPLEX Geo configurations

In case of a VPLEX Geo cluster failure, in this release of GeoSynchrony, the VPLEX Witness provides its guidance, but the cluster ignores this guidance for all asynchronous consistency groups. Instead the clusters continues leveraging the

preference rules set up for each asynchronous consistency group. For asynchronous consistency groups, the surviving cluster might require a data rollback before I/O can proceed. In this situation, the consistency groups continue suspending I/O regardless of the guidance of Witness and regardless of the pre-configured rule set. Manual intervention is required to force the rollback in order to change (roll back) the current data image to the last crash-consistent image preserved on disk. Since this is done without the application's knowledge, you will likely need to manually restart the application to ensure that it does not continue using the stale data that has been discarded by VPLEX (but still remains in the application's cache).

Similar semantics apply in the presence of cluster partition. For each consistency group, the VPLEX Witness guides a preferred cluster (because it is the active writer) to continue I/O. However, the cluster ignores this guidance for all asynchronous consistency groups. Instead the clusters continues leveraging the preference rules set up for each asynchronous consistency group. If data rollback is required, the cluster continues suspending, waiting for manual intervention.

Dynamically enabling VPLEX Witness

Once installed, you can enable or disable VPLEX Witness. When you disable VPLEX Witness, all distributed volumes in consistency groups leverage their pre-existing detach rule set semantics. On the other hand, if you enable VPLEX Witness, all distributed volumes placed in the synchronous consistency groups begin using the VPLEX Witness failure semantics.

Note: VPLEX Witness has no impact on either local volumes or distributed volumes outside of configured synchronous consistency groups.

Value of VPLEX Witness in VPLEX Geo

The value of the VPLEX Witness is different with VPLEX Geo than it is with VPLEX Metro. With VPLEX Metro, VPLEX Witness provides a zero-RTO and zero-RPO storage solution in the presence of cluster or inter-cluster connectivity failure. With VPLEX Geo, VPLEX Witness *does not* automate any failure scenarios. The data presented by VPLEX Witness CLI context may be helpful to facilitate the manual fail-over. See the *VPLEX with GeoSynchrony 5.0 CLI Reference* for details on the commands used to determine state of the VPLEX Witness and the clusters attached to the Witness.

Higher availability — VPLEX Metro and VPLEX Witness

The use cases detailed in "VPLEX Metro HA in a campus" on page 101 describe the combination of VMware and cross cluster connection with configurations that contain a VPLEX Witness. Refer to Chapter 5, "VPLEX Use Cases" for more information on the use of VPLEX Witness withVPLEX Metro and VPLEX Geo configurations.

Recovery

VPLEX stores configuration and metadata on system volumes created from storage devices. The two types of system volumes are metadata volumes and logging volumes.

Metadata volume failure

VPLEX maintains its configuration state, referred to as metadata, on storage volumes provided by storage arrays. Each VPLEX cluster maintains its own metadata, which describes the local configuration information for this cluster as well as any distributed configuration information shared between clusters.

VPLEX uses this persistent metadata on a full system boot and loads the configuration information onto each director. When you make changes to the system configuration, VPLEX writes these changes to the metadata volume. Should VPLEX lose access to the metadata volume, the VPLEX directors continue to provide their virtualization services using the in-memory copy of the configuration information. Should the storage supporting the metadata device remain unavailable, configure a new metadata device. Once you assign a new device, VPLEX records its in-memory copy of the metadata device maintained by the cluster on the new metadata device.

VPLEX suspends the ability to perform configuration changes when access to the persistent metadata device is not available.

During normal operations and while configuration changes are taking place, metadata volumes experience light I/O activity. During boot operations and non-disruptive upgrade activities, metadata volumes experience high read I/O.

Best practices for metadata volumes

Follow these best practices to provide the highest resilience in your federated storage area network:

- Allocate storage volumes of 78GB for the metadata volume
- Configure the metadata volume for each cluster with multiple back-end storage volumes provided by different storage arrays of the same type
- Use the data protection capabilities provided by these storage arrays, such as RAID 1 and RAID 5 to ensure the integrity of the system's metadata
- Create backup copies of the metadata whenever configuration changes are made to the system
- Perform regular backups of the metadata volumes on storage arrays that are separate from the arrays used for the metadata volume

Dirty region logging volumes

In the event of a link outage between clusters, and if you have chosen to continue I/O to distributed volumes at one of the clusters, the updated pages are synchronized with the other cluster after the link is restored. To minimize the traffic across the link after an outage, VPLEX maintains a record of which pages are written at each cluster during the outage so that only the changed pages need be exchanged. This area is referred to as a dirty region log and is stored on the logging volume. Should this volume become inaccessible, the directors record the entire leg as out-of-date and require a full synchronization of this leg once it is reattached to the mirror.

During normal operations, when both legs of the RAID 1 volume are active and accessible, there is no I/O to the logging volume. When a loss of connectivity occurs and the distributed mirror is fractured, there is high write I/O activity on the cluster that continues operation. When the detached leg of the mirror is reattached, VPLEX

performs an incremental synchronization where it reads this logging volume to determine what writes are necessary to synchronize the reattached volume. During that synchronization, the logging volume experiences high read I/O activities.

Best practices for logging volumes

Use these best practices when configuring logging volumes on a VPLEX Metro or VPLEX Geo configuration.

- Create one logging volume for each cluster
- Use the data protection capabilities provided by the storage array, such as RAID 1 and RAID 5 to ensure the integrity of the system's logging volume
- Configure 1 GB of logging volume space for every 32TB of distributed device space

VPLEX distributed cache protection and redundancy

VPLEX utilizes the individual director's memory systems to ensure durability of user and critical system configuration data. User data is made durable in one of two ways depending on the cache mode used for the data.

- Write-through cache mode leverages the durability properties of a back-end array by writing user data to the array and obtaining an acknowledgement for the written data before it acknowledges the write back to the host.
- Write-back cache mode ensures data durability by storing user data into the cache memory of the director that received the I/O, then placing a protection copy of this data on another director in the local cluster before acknowledging the write to the host. This ensures the data is protected in two independent memories. The data is later destaged to back-end storage arrays that provide the physical storage media.

Global distributed cache protection from power failure

In Release 5.0.1, in the event of a cluster-wide power failure in a VPLEX Geo configuration, each VPLEX director copies its dirty cache data to the local solid state storage devices (SSDs). This process, known as *cache vaulting*, protects user data in cache if power is lost. After each director vaults its dirty cache pages, VPLEX then shuts down the director's firmware.

When you resume operation of the cluster, if any condition is not safe, the system does not resume normal status and calls home for diagnosis and repair. This allows EMC Customer Support to communicate with the VPLEX system and restore normal system operations.

Once power is restored, the VPLEX system startup program initializes the hardware and the environmental system, checks the data integrity of each vault, and restores the cache memory contents from the solid state vault devices.

Under normal conditions, the SPS batteries can support two consecutive vaults; this ensures after the first power failure, that the system can resume I/O immediately, and that it can still vault if there is a second power failure, without risking data loss.

Chapter 3, "VPLEX Software," includes details of the cache vaulting and recovery process.

VPLEX security features

The VPLEX management server operating system (OS) is based on a Novell SUSE Linux Enterprise Server 10 distribution. The operating system has been configured to meet EMC security standards by disabling or removing unused services, and protecting access to network services through a firewall.

Security features include:

- Using SSH to access the management server shell
- Using HTTPS to access the VPLEX GUI
- Using an IPsec VPN in a VPLEX Metro and VPLEX Geo configurations
- Using an IPSEC VPN to connect each cluster of a VPLEX Metro or VPLEX Geo to the VPLEX Witness server
- Using SCP to copy files
- Using a tunneled VNC connection to access the management server desktop
- Separate networks for all VPLEX cluster communication
- Defined user accounts and roles
- Defined port usage for cluster communication
- Network encryption
- Certificate Authority (CA) certificate (default expiration 5 years)
- Two host certificates (default expiration 2 years)
- Third host certificate for optional VPLEX Witness

CAUTION

The inter cluster link carries unencrypted user data. To ensure privacy of the data, establish an encrypted VPN tunnel between the two sites.

5 VPLEX Use Cases

This section provides examples of VPLEX configurations and how they can be used.

Technology refresh

In current IT environments, servers are often connected to redundant front-end fabrics and storage is connected to redundant back-end fabrics. Current models do not facilitate federation among storage arrays. Migrations between heterogeneous arrays are often complicated and necessitate the purchase of additional software or functionality. Integrating heterogeneous arrays in a single environment is difficult and requires a staff with a diverse skill set.

Figure 33 shows the traditional view of storage arrays with servers attached at the redundant front end and storage (Array 1 and Array 2) connected to a redundant fabric at the back end.

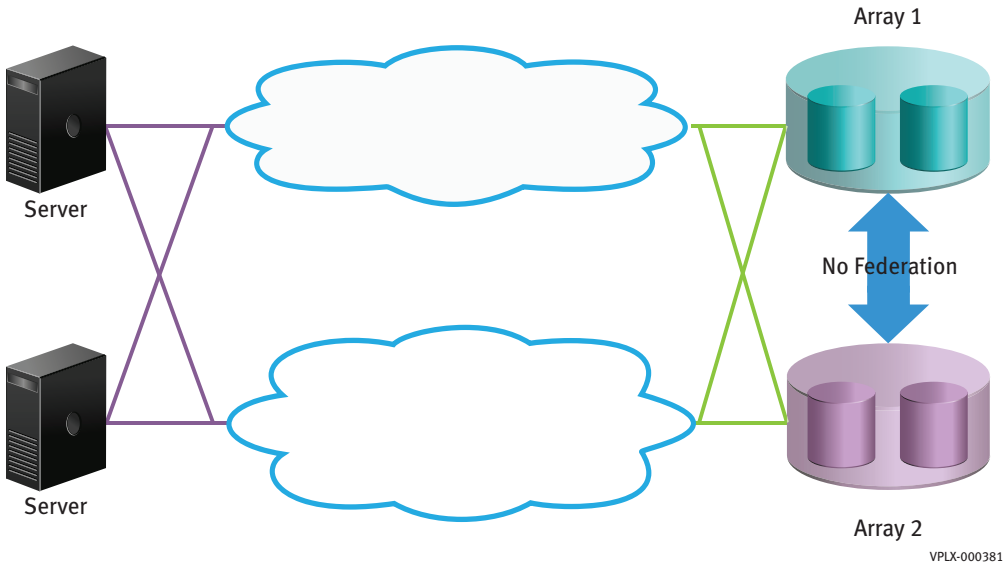

Figure 33 Traditional view of storage arrays

VPLEX introduces a virtualization layer to the current IT environment. VPLEX is inserted between the front-end and back-end redundant fabrics. VPLEX appears to be the target to hosts and the initiator to storage. This allows you to change the back-end storage transparently. VPLEX makes it easier to integrate heterogeneous storage arrays on the back-end. Migration between storage arrays becomes much simpler with VPLEX as well.

Figure 34 shows the vision of storage when VPLEX is presenting an abstract storage configuration.

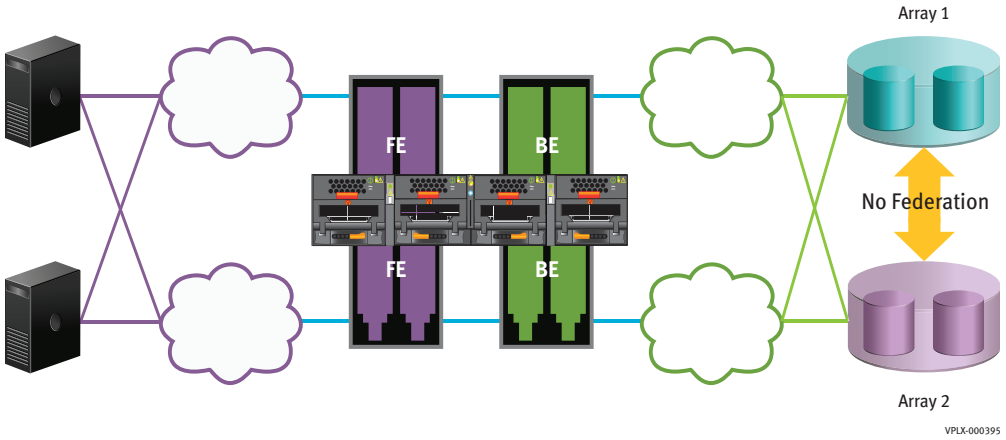

Figure 34 VPLEX virtualization layer

This abstract view of storage becomes very powerful when it comes time to replace the physical array that is providing storage to applications. Historically, the data used by the host was copied to a new volume on the new array and the host was reconfigured to access the new volume. This process requires downtime for the host.

With VPLEX, because the data is in virtual volumes, it can be copied nondisruptively from one array to another without any downtime for the host. The host does not need to be reconfigured; the physical data relocation is performed by VPLEX transparently and the virtual volumes retain the same identities and the same access points to the host.

Figure 35 shows an example of a technology refresh.

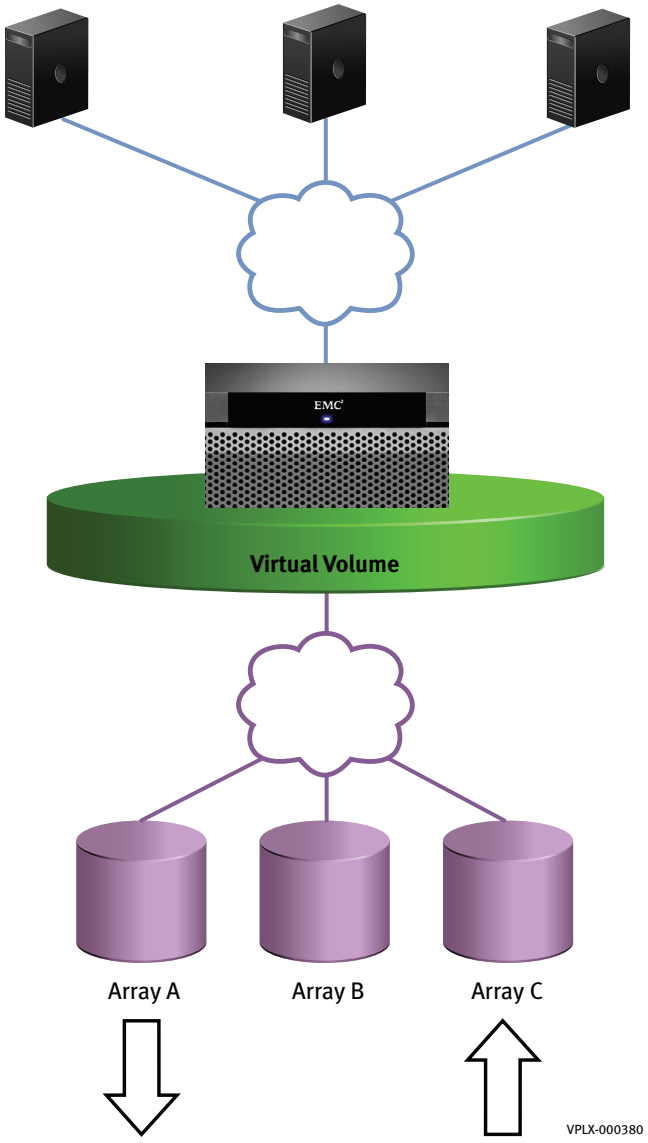

Figure 35 VPLEX technology refresh

In Figure 35, the virtual disk is made up of the disks of Array A and Array B. The site administrator has determined that Array A has become obsolete and should be replaced with a new array. Array C is the new storage array. The administrator adds this array into the VPLEX cluster and using the Mobility Central functionality in the GUI, assigns a target extent from the new array to each extent from the old array and instructs VPLEX to perform the migration. Copying the data from Array A to Array C occurs while the host continues its access to the virtual volume without disruption. After the copy of Array A to Array C is complete, the administrator can decommission Array A. Because the virtual machine is addressing its data to the abstracted virtual volume, its data continues to flow to the virtual volume with no need to change the address of the data store. Although this example uses virtual machines, the same is true for traditional hosts. Using VPLEX the, administrator can move data used by an application to a different storage array without the application or server being aware of the change.

Data mobility

VPLEX provides direct support for data mobility both within and between data centers near and far and enables application mobility, data center relocation, and consolidation.

Data mobility is the relocation of data from one location (the source) to another (the target), after which the data is subsequently accessed only through the target. By contrast, data replication enables applications to continue to access the source data after the target copy is created. Similarly, data mobility is different from data mirroring, which transparently maintains multiple copies of the data for the purposes of data protection and access.

During and after a data mobility operation, applications continue to access the data using its original VPLEX volume identifier. This avoids the need to point applications to a new data location or change the configuration of their storage settings, effectively eliminating the need for application cut over.

There are many types and reasons for data mobility:

- Moving data from one storage device to another. For example, if a device has been deemed "Hot" the data can be moved to a less utilized storage device.
- Moving applications from one storage device to another.
- Moving operating system files from one storage device to another.
- Consolidating data or database instances.
- Moving database instances.
- Moving storage infrastructure from one physical location to another.

The non-disruptive nature of VPLEX data mobility operations helps to simplify the planning and execution factors that would normally be considered when performing a disruptive migration.

It is still important to consider some of these factors, however, when performing data mobility between data centers and increasing the distance between an application and its data. Considerations include the business impact and the type of data to be moved, site locations, and total amount of data, as well as time considerations and schedules.

Mobility with the VPLEX migration wizard

The VPLEX GUI supports the ability to easily move the physical location of virtual storage while VPLEX provides continuous access to this storage by the host. Using this wizard, you first display and select the extents (for extent mobility) or devices (for device mobility) to move. The wizard then displays a collection of candidate storage volumes. Once virtual storage is selected, VPLEX automates the process of moving the data to its new location. Throughout the process the volume retains it volume identity, and continuous access is maintained to the data from the host.

There are three types of mobility jobs:

Table 6 Types of data mobility operations

Extent	Moves data from one extent to another extent (within a cluster).
Device	Moves data from one device to another device (within a cluster).
Batch	Groups extent or device mobility jobs into a batch job, which is then executed as a single job.

Best practices for data mobility

Data mobility must be a planned activity.

All components of the system (virtual machine, software, volumes) must be available and in a running state.

Data mobility can be used for disaster avoidance or planned upgrade or physical movement of facilities.

How data mobility works

When a mobility job begins, VPLEX creates a temporary RAID 1device above each device or extent to be migrated. The target extent or device becomes a mirror leg of the temporary device, and synchronization between the source and the target begins. Once synchronization completes, you can commit (or cancel) the mobility job.

Because the data mobility operation is non-disruptive, the application continues to write to the volumes during a mobility operation. The new I/Os are written to both legs of the device.

The following rules apply to mobility operations:

- The source device can have active I/O.
- The target device cannot be in use (no virtual volumes created on it).
- The target extent/device must be the same size or larger than the source extent/device.
- The target extent cannot be in use (no devices created on it).

When creating a mobility job, you can control the transfer speed. The higher the speed, the greater the impact on host I/O. A slower transfer speed results in the mobility job taking longer to complete, but has a lower impact on host I/O. You can change the transfer speed of a job while the job is in the queue or in progress. The change takes effect immediately.

In GeoSynchrony 5.0, the thinness of a thinly provisioned storage volume is retained through a mobility operation. By default, VPLEX moves all volumes as thick storage volumes unless you specify that rebuilds should be thin at the time you provision the thin volume. Refer to the *EMC VPLEX with GeoSynchrony 5.0 and Point Releases CLI Guide* or the online help for more information on thin provisioning of volumes.

Redundancy with RecoverPoint

CLARiiON-RecoverPoint integration can be deployed along with VPLEX, in VPLEX Local and VPLEX Metro configurations, to complement VPLEX storage federation technology with any-point-in-time data recovery volumes implemented on those CLARiiON arrays.

Note: RecoverPoint integration is not supported in VPLEX Geo configurations.

The RecoverPoint utilizes RecoverPoint Appliances (RPAs) to replicate any write operations to a data recovery site.The RecoverPoint appliance runs the RecoverPoint software on top of a custom 64-bit Linux Kernel inside a secure environment built from an industry standard server platform. An RPA manages all aspects of data protection for a storage group including capturing changes, maintaining the images in the journaling volumes, and performing image recovery. A single appliance can manage multiple storage groups, each with differing policies.

Figure 36 shows a RAID 1 device that has legs sourced from two different local back-end arrays. The host accesses the virtual volume that resides on top of the RAID 1 device. The individual legs of the RAID 1 device must use one-for-one encapsulation. One-for-one encapsulation requires that there can only be one extent created on the storage volume and the extent must utilize all of the capacity of the storage volume. The device that uses the extent also uses the entire capacity of the extent. The second leg of the device comes from a CLARiiON CX4 storage volume that is utilizing the CLARiiON CX4 splitter. The data recovery site contains a third array that is acting as a remote copy target for RecoverPoint CRR. The third leg hosted at the data recovery site is for replication and recovery only; a host should not access the volume directly. The distance between Site 1 and the data recovery site varies based on the recovery objectives, inter-site bandwidth, latency, and other limitations outlined in the EMC Simple Support Matrix (ESSM) for RecoverPoint.

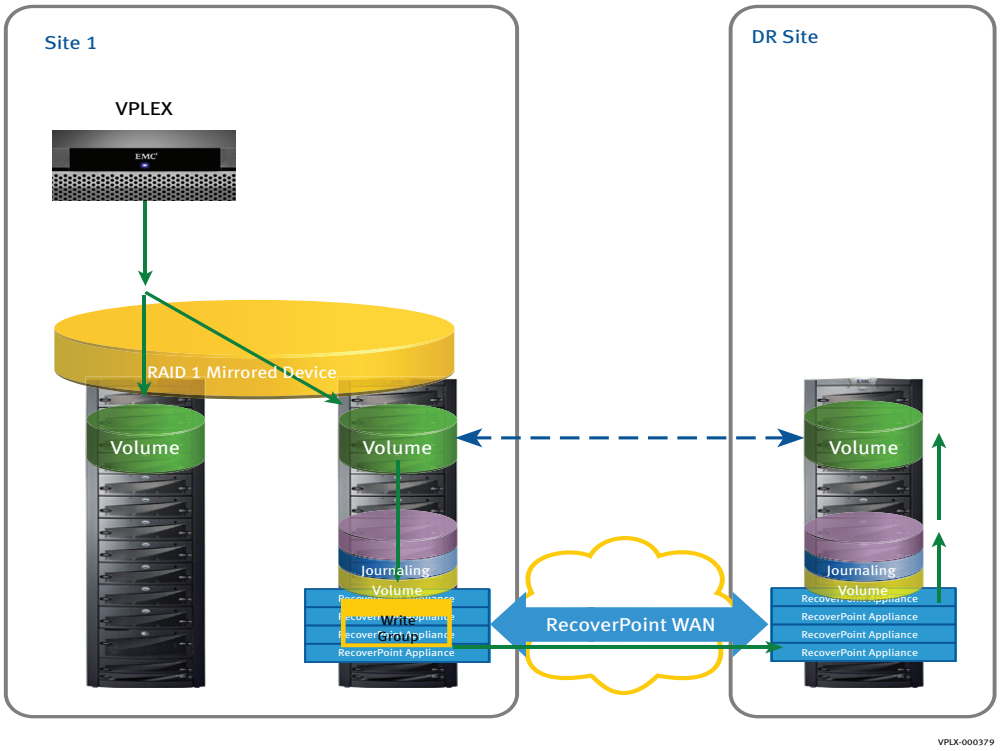

Figure 36 RecoverPoint used with a mirrored device

RecoverPoint uses two volumes to manage its replication efforts:

The *repository volume* is a SAN-attached volume provisioned only to RecoverPoint. RecoverPoint uses this volume to maintain the configuration and communication between RPAs in a cluster. Similar to a cluster server's quorum volume, the repository volume contains the status of the overall RecoverPoint system and acts as a resource arbitrator during RPA failover and recovery operations. There is no user-accessible information stored on the repository volume.

The journal volume holds data waiting to be distributed to target replica volumes and also retains copies of the data previously distributed to the target volumes to facilitate operational recovery to any point in time that is retained in the journal history. Each consistency group has its own journal volumes, which allows for differing retention periods across consistency groups. Each consistency group has two or three journal volumes, one assigned to the local copy volumes, one assigned to the remote copy volumes, and one assigned to the production or source volumes. Journal volumes are used for the source and both copies in order to support production failover from the current active source volume to either the local or remote copy volume. A given copy journal can consist of one or more storage devices. RecoverPoint will stripe the data across the number of devices provisioned for a given journal.

A write originates from a host at Site 1 and, is written to both legs of the RAID 1 device. VPLEX acknowledges the write back to the host. A *splitter* resides in the CLARiiON array's I/O stack and intercepts writes to send them to the RPA. While writes enter the CLARiiON CX4, the CX4 splitter splits writes and compresses them into write groups. Eventually, the entire compressed write group is sent across the RecoverPoint WAN link to the remote data recovery site. RecoverPoint uncompresses the write group and writes to journaling volumes. Writes are distributed to the appropriate target volumes from the journaling volumes.

RecoverPoint maintains transactional consistent journals for each application defined within a RecoverPoint system. The journal allows convenient rollback to any point in time, enabling instantaneous recovery for application environments.

RecoverPoint can be combined with VPLEX Metro to facilitate three-site redundancy. Site 1 and Site 2 VPLEX clusters, allowing the creation of distributed devices and facilitating RAID 1 mirroring across a distance. The DR Site serves as the data recovery site. Site 1 could host RecoverPoint Appliances with CX4 splitters to perform the write-splitting

.

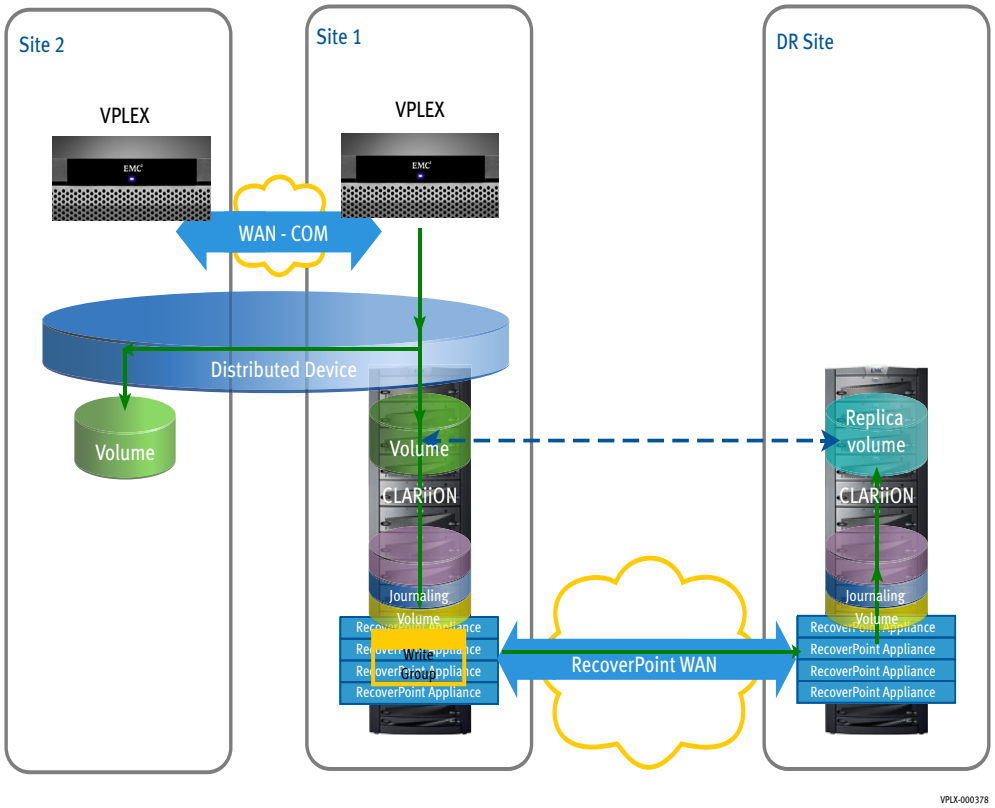

Figure 37 RecoverPoint used with a VPLEX Metro distributed device

Figure 37 shows a distributed RAID 1 device that spans two arrays at two separate sites. The host accesses the volume that resides on top of the RAID 1 local device. The individual legs of the RAID 1 device must use one-for-one encapsulation. The first leg of the device comes from a CLARiiON CX4 storage volume that is using the CLARiiON CX4 splitter. The data recovery site contains a third array that is acting as a remote copy target for RecoverPoint CRR. The third leg hosted at the data recovery site is for replication and recovery only; a host should not access the volume directly. The distance between Site 1, Site 2, and DR site will vary based on the recovery objectives, inter-site bandwidth, latency, and other limitations outlined in the EMC Simple Support Matrix (ESSM) for RecoverPoint.

A write originates from a host at Site 1 and, is replicated across the VPLEX inter-cluster WAN link to Site 2. The write is then written to back-end storage at both sites and the write operation is acknowledged back to the host. While writes enter the CLARiiON CX4, the CX4 splitter is splitting writes and compressing them into a write group. Eventually, the entire compressed write group is sent across the

RecoverPoint WAN link to the remote data recovery site. The write group is then uncompressed and written to journaling volumes. Writes are distributed to the appropriate target volumes from the journaling volumes.

Restoring VPLEX virtual volumes with RecoverPoint CRR

During a restore process, RecoverPoint writes to the back-end storage volumes outside of the VPLEX I/O path. When writes occur outside of the VPLEX I/O path, VPLEX read cache may not match the data on the storage volume. To avoid this situation, VPLEX read cache invalidation for each restored storage volume is required. To perform per-volume read cache invalidation, remove the corresponding VPLEX virtual volume from the storage view and the read cache for that volume is discarded. The VPLEX virtual volumes being restored must remain removed from all storage views until RecoverPoint completes its writes. Once the restore activities finish, then the virtual volumes can be added back into a storage view and accessed normally.

Data mobility with RecoverPoint

As described in "Data mobility," VPLEX provides nondisruptive data mobility within and between storage arrays. Because the RecoverPoint array-based CX4 splitter is based on writes being sent to a specific CLARiiON source LUN, if VPLEX performs a mobility job from that source LUN to another LUN, then the RecoverPoint no longer performs replication for that source volume. In order for RecoverPoint replication to restart you must perform the following tasks:

- Remove the RecoverPoint relationship between the old source volume and a target
- Configure the new source volume appropriately
- Perform a full sweep on the device

The *EMC RecoverPoint Administrator's Guide* explains the necessary steps and the impact of this process. When mobility jobs are planned for VPLEX virtual volumes, it is vital that the planning also takes into account any corresponding RecoverPoint CRR relationships.

Distributed data collaboration

With VPLEX, the same data can be accessible from either VPLEX cluster at all times — even if they are at different sites. The data is literally shared, not copied, so that a change made in one site are replicated at the other site. Current applications for the sharing of information over distance are not suitable for collaboration or for BigData environments. For example, transfer of hundreds of GB or event TB of data across WAN using FTP is extremely slow, results in duplication of data, and is extremely inefficient if you need to modify a small portion of a huge data set or use it for analysis. With VPLEX, the problem of unnecessarily copying the data is eliminated. Instead, VPLEX makes the data available in both locations, and because VPLEX is smart, it doesn't need to ship the entire file back and forth like other solutions — it only sends the information that is being accessed but is not available locally, thus greatly reducing bandwidth costs and offering significant savings over other solutions. Deploying VPLEX in conjunction with third-party WAN optimization solutions can deliver even greater benefits. And with VPLEX AccessAnywhere, the data remains online, and available.

This is a huge benefit for customers who have always had to rely on shipping large log files and data sets back and forth across sites, then wait for updates to be made in another location before they could resume working on it again.

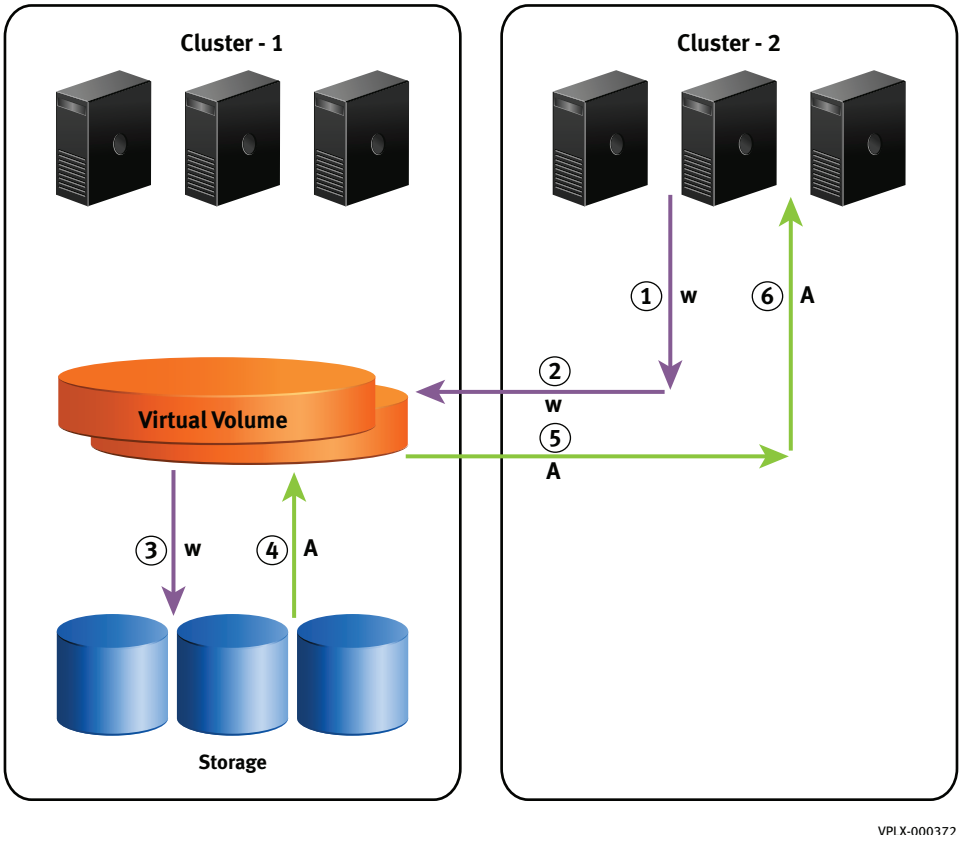

Figure 38 Data shared with global visibility

One of the ways in which distributed data collaboration could be implemented is in the form of local consistency groups with global visibility. Local consistency groups with global visibility allow the remote cluster to read and write to the consistency group. However, all reads and writes to the consistency group from the remote

cluster must pass over the WAN-COM link to access the consistency group. This allows the remote cluster to have instant on-demand access to the consistency group, but also adds additional latency for the remote cluster. Local consistency groups with global visibility are supported in both VPLEX Metro and VPLEX Geo environments. However, the round-trip latency in both cases must be 5ms RTT latency or less. Only local volumes can be placed into the local consistency group with global visibility. Local consistency groups with global visibility cannot be set to asynchronous cache mode. I/O that goes to local consistency groups with global visibility will always be synchronous.

For distributed data collaboration over greater distances, an asynchronous consistency group could provide mirrored volumes at both locations. In a configuration where clusters are further apart, a higher network latency would significantly affect I/O to distributed volumes since host I/O must be acknowledged at both clusters before being written to the back end. This can result in data loss in the event of an engine, cluster, or link failure. To enable this configuration to allow more than 5ms of latency between clusters, asynchronous I/O is used. Figure 39 shows how an asynchronous consistency group can support the distributed data collaboration use case.

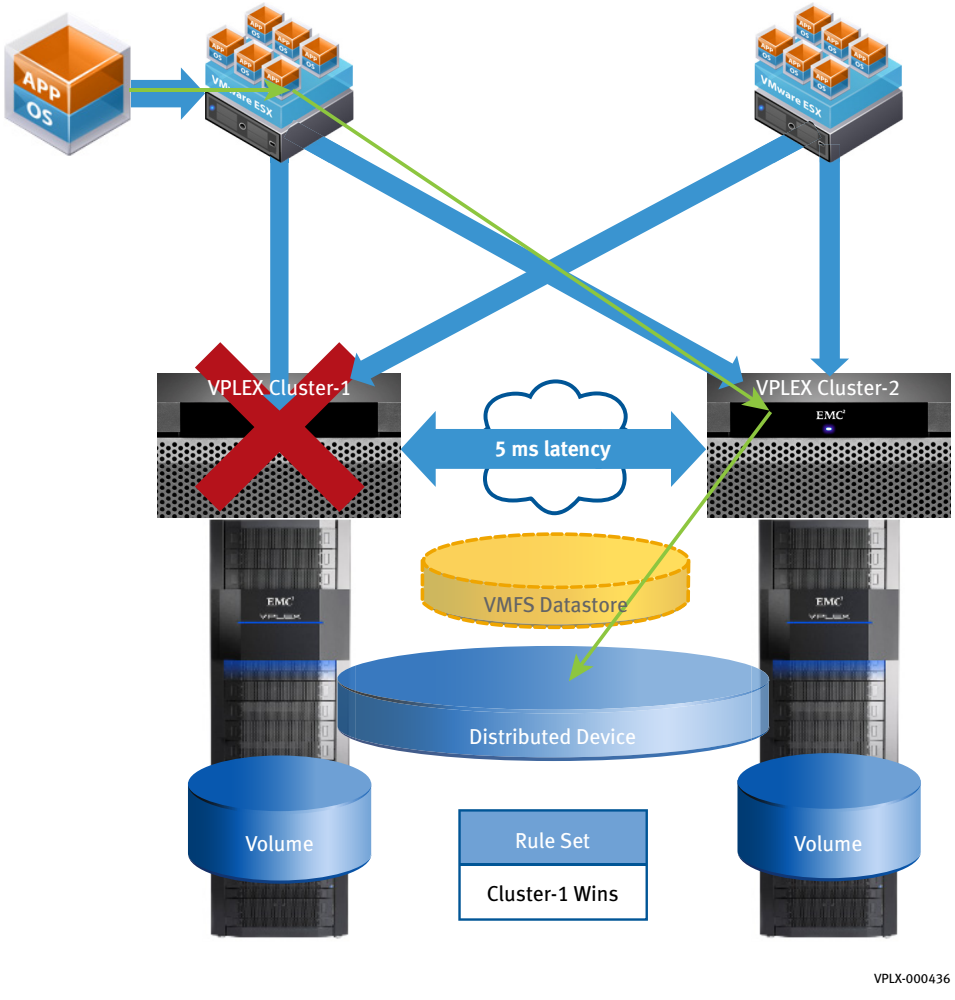

Figure 39 Asynchronous consistency group for distributed data collaboration

VPLEX Metro HA in a campus

VPLEX Metro HA configurations consist of a VPLEX Metro system deployed in conjunction with VPLEX Witness. There are two types of Metro HA configurations; a generic one that can be stretched up to 5 ms of latency between data centers, and one using Cross Connect between VPLEX clusters and hosts, which provides a higher level of availability but is constrained to distances with up to 1ms round trip time latency. The key to these environments is AccessAnywhere. It allows both clusters to provide simultaneous 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 during normal operation and not only after failover. When you combine this with host failover clustering technologies such as VMware HA, this provides you with 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. VPLEX Metro HA configurations:

- Ride through any single component failures within the storage subsystem
- Provide automatic restart in case of any failure in the environment
- Optionally, with DRS enabled, workload spikes can be distributed between data centers alleviating the need to purchase more storage
- No requirement to stretch the Fiber Channel fabric between sites. You can maintain fabric isolation between the two sites

VMware ESX can be deployed at both clusters in a Metro environment to create a high availability environment. VMware can be deployed with or without Cross Connect. Figure 40 on page 102 shows the Metro HA configuration without Cross Connect. Notice that the two clusters must have less than 5 ms of WAN-COM latency.

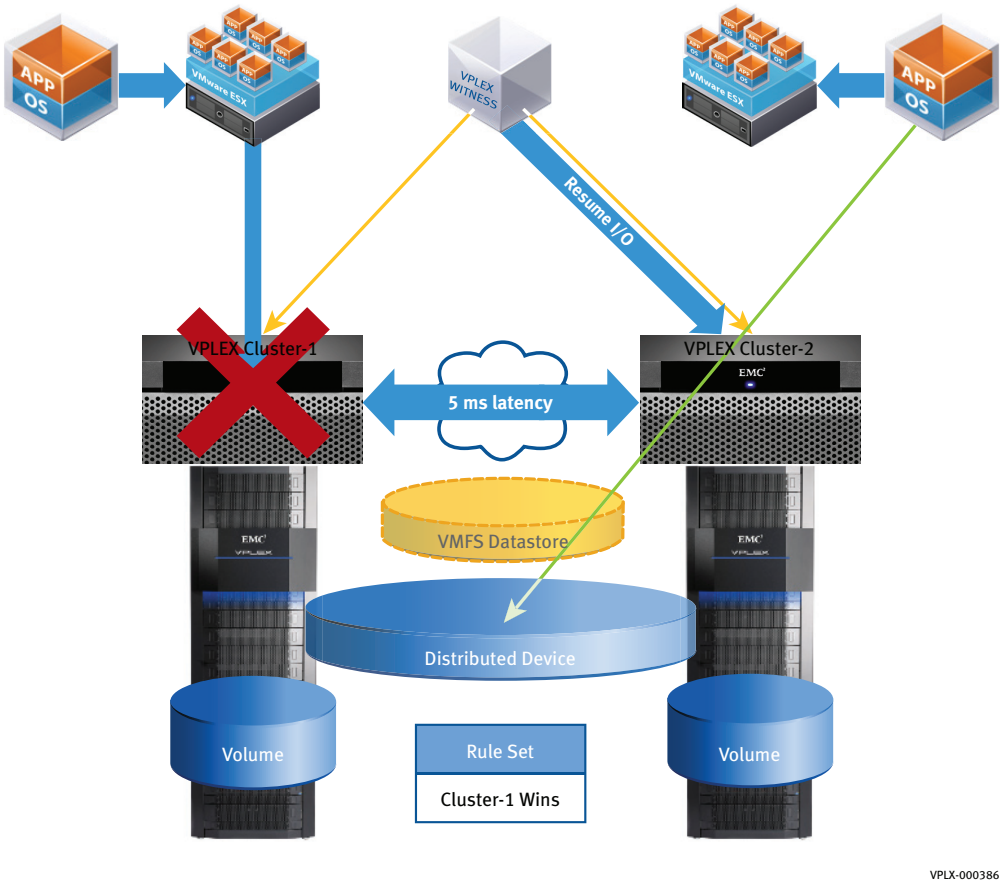

Figure 40 VMware Metro HA without Cross Connect

In this scenario, a virtual machine can write to the same distributed device from either cluster. In other words, if the customer is using VMware Distributed Resource Scheduler (DRS), which allows the automatic load distribution on virtual machines across multiple ESX servers, a virtual machine can be moved from an ESX server attached to Cluster-1 to an ESX server attached to Cluster-2 without losing access to the underlying storage. This configuration allows virtual machines to move between two geographically disparate locations with up to 5ms of latency.

If the Distributed Resource Scheduler moves a virtual machine to the Cluster-2 ESX server, the virtual machine continues to write to its distributed device. VPLEX, in this way, supports the mobility of virtual machines across geographic locations.

A data unavailability event can occur when there is not a full site outage, but there is a VPLEX outage on Cluster-1 and the virtual machine is currently running on the ESX server attached to Cluster-1. If this configuration also contains a VPLEX Witness, the VPLEX Witness recognizes the outage and recommends that Cluster-2 resume I/O rather than following the rule set. The virtual machine running against storage exported at Cluster-1 does not fail, despite all paths to storage having gone away. Because it doesn't fail, VMware HA has no reason to try to restart it on a server attached to Cluster-2. Cluster-2, in this case, is ready to process reads and writes from the virtual machine, but the virtual machine is hung at Cluster-1. Manual intervention is required at this point to move the virtual machine to the Cluster-2 ESX server to provide access to the data.

Best practices with Metro HA

Follow these best practices when configuring a Metro HA in a campus:

- Configure the front end with a stretched layer-2 network so that when a virtual machine moves between sites, its IP address can stay the same.
- Use DRS host affinity rules if DRS is enabled. Host affinity rules keep virtual machines running in the preferred site as long as the virtual machines can, and only moves the virtual machines to the non-preferred site if they can not run in the preferred site.
- Deploy the VPLEX Witness with this type of solution to avert some system-wide data unavailability events.

The VMware Metro HA cross-connect environment is very similar to the deployment shown in Figure 40 on page 102. However, each ESX server is connected and zoned to both VPLEX clusters. By cross-connecting ESX servers to VPLEX, the failure of one VPLEX cluster does not result in data unavailability.

Note: The VMware Metro HA cross connect use case is supported at latencies of up to 1ms RTT.

In Figure 41 on page 104, the virtual machine is deployed at Cluster-1. If Cluster-1 fails, the VPLEX Witness recommends that VPLEX ignore the rule set and use Cluster-2. Immediately after Cluster-2 resumes I/O, the virtual machine can write through the remote VPLEX cluster without being migrated to Cluster-2.

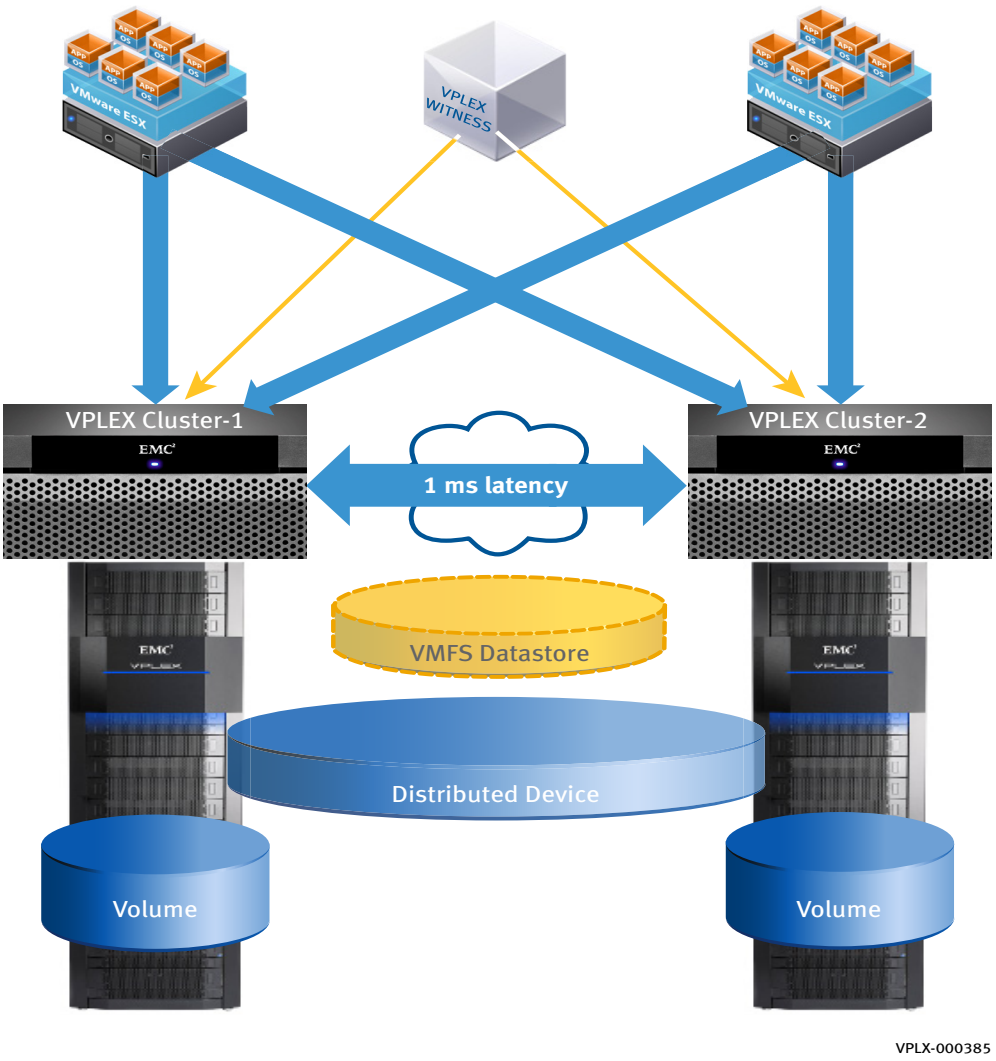

Figure 41 **VMware Metro HA with Cross-Connect**

Again, it is very important that the VPLEX Witness be deployed in this type of environment to help avert data unavailable events.

Refer to the *EMC Simple Support Matrix, EMC VPLEX and GeoSynchrony*, available at http://elabnavigator.EMC.com under the Simple Support Matrix tab for the latest list of software and hardware versions supported by VPLEX with GeoSynchrony 5.0 and point releases.

VPLEX MetroHA failure handling

Figure 42 shows areas in which a typical VPLEX Metro configuration that could (though unlikely) fail.

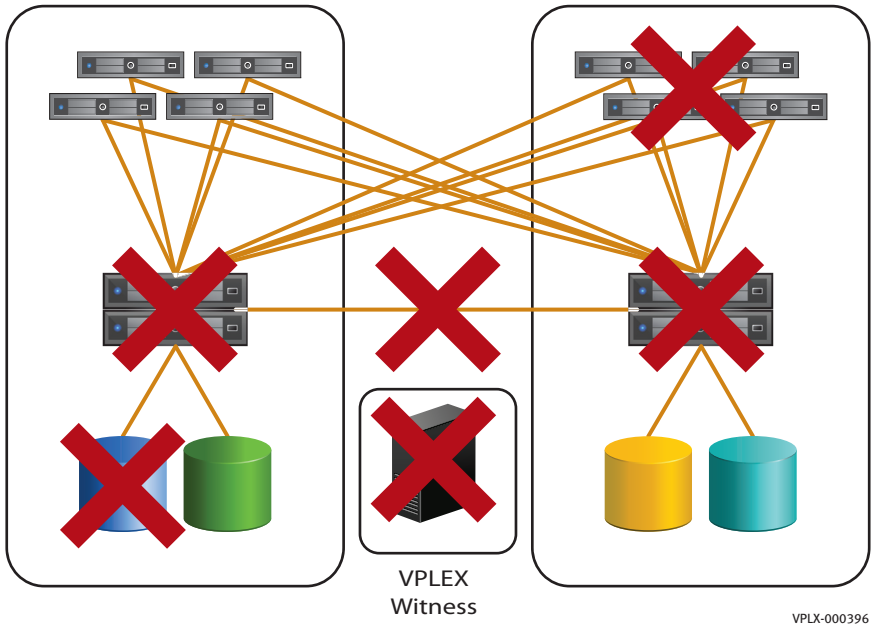

Figure 42 VPLEX Metro HA failure handling

Table 7 describes how VPLEX Metro HA handles each failure shown in Figure 42.

Table 7 How VPLEX Metro HA recovers from failure

Failure description	Failure handling
Server failure in Data Center 1	VMware HA software restarts the affected applications in data center 2 automatically.
VPLEX cluster failure in Data Center 2	VPLEX Witness detects the failure and enables all volumes on surviving cluster.
Inter site link failure	If the cross-connect links leverage a different physical link from that used by the inter-cluster WAN Com, applications are unaffected. Every volume continues to be made available in one data center or the other. However, if the cross-connect links leverage the same physical link as the inter-cluster WAN Com, application restart is required.
Storage array failure	Applications are unaffected. VPLEX dynamically redirects I/O to mirrored copy on surviving array. **Note:** This example assumes that all distributed volumes are also mirrored on the local cluster. If a distributed volume is not mirrored on the local cluster, then the application still remains available (because the data can be fetched/sent from/to the remote cluster). However, each read/write operation now incurs a small performance cost.
Failure of VPLEX Witness	After recognizing a loss of connectivity with the VPLEX Witness, both clusters call home. As long as both clusters continue to operate and there is no inter-cluster link partition, applications are unaffected. If either cluster fails or if there is an inter-cluster link partition, the system is in jeopardy of data unavailability. Therefore, it is recommended that if the VPLEX Witness outage is expected to be long, the VPLEX Witness functionality should be disabled to prevent the possible data unavailability.