Best Practices

VPLEX™ Overview and General Best Practices

Implementation Planning and Best Practices

Abstract

This White Paper provides an overview of VPLEX and general best practices. It provides guidance for VS2 and VS6 hardware platforms.

March 2019

H13545

Revisions

Date	Description
March 2019	Version 4

Acknowledgements

This paper was produced by the following:

Author: VPLEX CSE Team

VPLEX_CSE_TEAM@emc.com

Table of contents

Executive summary

EMC VPLEX represents the next-generation architecture for data mobility and continuous availability. This architecture is based on EMC's 20+years of expertise in designing; implementing and perfecting enterprise class intelligent cache and distributed data protection solutions.

VPLEX addresses two distinct use cases:

- Data Mobility: The ability to move application and data across different storage installations – within the same datacenter, across a campus, or within a geographical region.
- Continuous Availability: The ability to create a continuously available storage infrastructure across the same varied geographies with unmatched resiliency

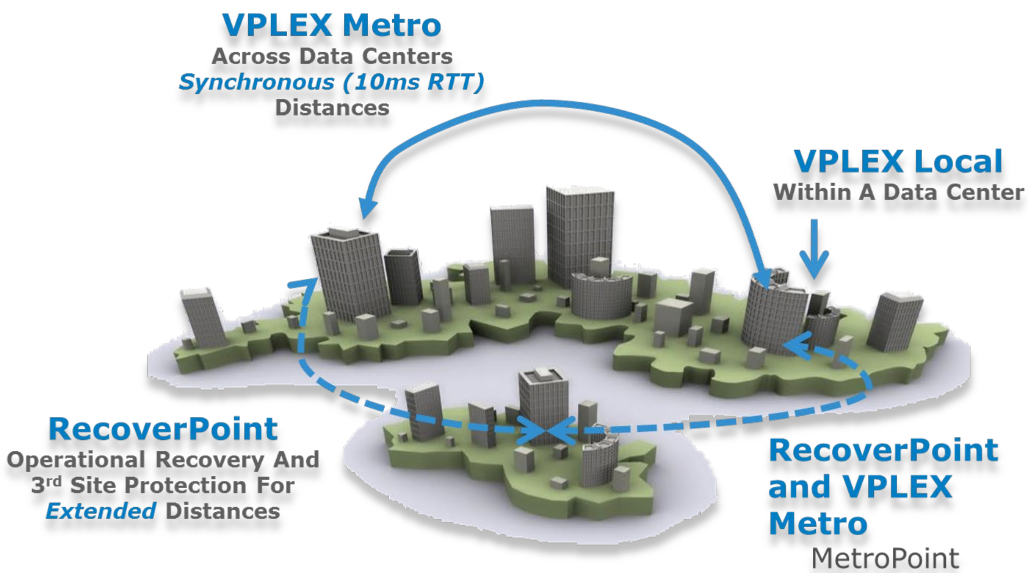

Figure 1 Access anywhere, protect everywhere

1 VPLEX Overview

1.1 VPLEX Platform Availability and Scaling Summary

VPLEX addresses continuous availability and data mobility requirements and scales to the I/O throughput required for the front-end applications and back-end storage.

Continuous availability and data mobility features are characteristics of both VPLEX Local and VPLEX Metro.

A VPLEX VS2 cluster consists of one, two, or four engines (each containing two directors), and a management server. A dual-engine or quad-engine cluster also contains a pair of Fibre Channel switches for communication between directors within the cluster.

A VPLEX VS6 cluster consists of one, two, or four engines (each containing two directors), and dual integrated MMCS modules in Engine 1 that replace the management server. A dual-engine or quad-engine cluster also contains a pair of Infiniband switches for communication between directors within the cluster.

Each engine is protected by a standby power supply (SPS) on the VS2 or integrated battery backup modules (BBU) on the VS6, and each internal switch gets its power through an uninterruptible power supply (UPS). (In a dual-engine or quad-engine cluster, the management server on VS2 also gets power from a UPS.)

The management server has a public Ethernet port, which provides cluster management services when connected to the customer network. This Ethernet port also provides the point of access for communications with the VPLEX Witness. The MMCS modules on VS6 provide the same service as the management server on VS2. Both MMCS modules on engine 1 of the VS6 must be connected via their public Ethernet port to the customer network and allow communication over both ports. This adds the requirement of having to allocate two IP addresses per cluster for the VS6. The GUI and CLI are accesses strictly through MMCS-A and the vplexcli is disabled on MMCS-B.

VPLEX scales both up and out. Upgrades from a single engine to a dual engine cluster as well as from a dual engine to a quad engine are fully supported and are accomplished non-disruptively. This is referred to as scale up. Scale “out” upgrade from a VPLEX Local to a VPLEX Metro is also supported non-disruptively. Generational upgrades from VS2 to VS6 are also non-disruptive.

1.2 Data Mobility

Dell EMC VPLEX enables the connectivity to heterogeneous storage arrays providing seamless data mobility and the ability to manage storage provisioned from multiple heterogeneous arrays from a single interface within a data center. Data Mobility and Mirroring are supported across different array types and vendors.

VPLEX Metro configurations enable migrations between locations over synchronous distances. In combination with, for example, VMware and Distance vMotion or Microsoft Hyper-V, it allows you to transparently relocate Virtual Machines and their corresponding applications and data over synchronous distance. This provides you with the ability to relocate, share and balance infrastructure resources between data centers.

All Directors in a VPLEX cluster have access to all Storage Volumes making this solution what is referred to as an N -1 architecture. This type of architecture allows for multiple director failures without loss of access to data down to a single director.

During a VPLEX Mobility operation any jobs in progress can be paused or stopped without affecting data integrity. Data Mobility creates a mirror of the source and target devices allowing the user to commit or cancel the job without affecting the actual data. A record of all mobility jobs are maintained until the user purges the list for organizational purposes.

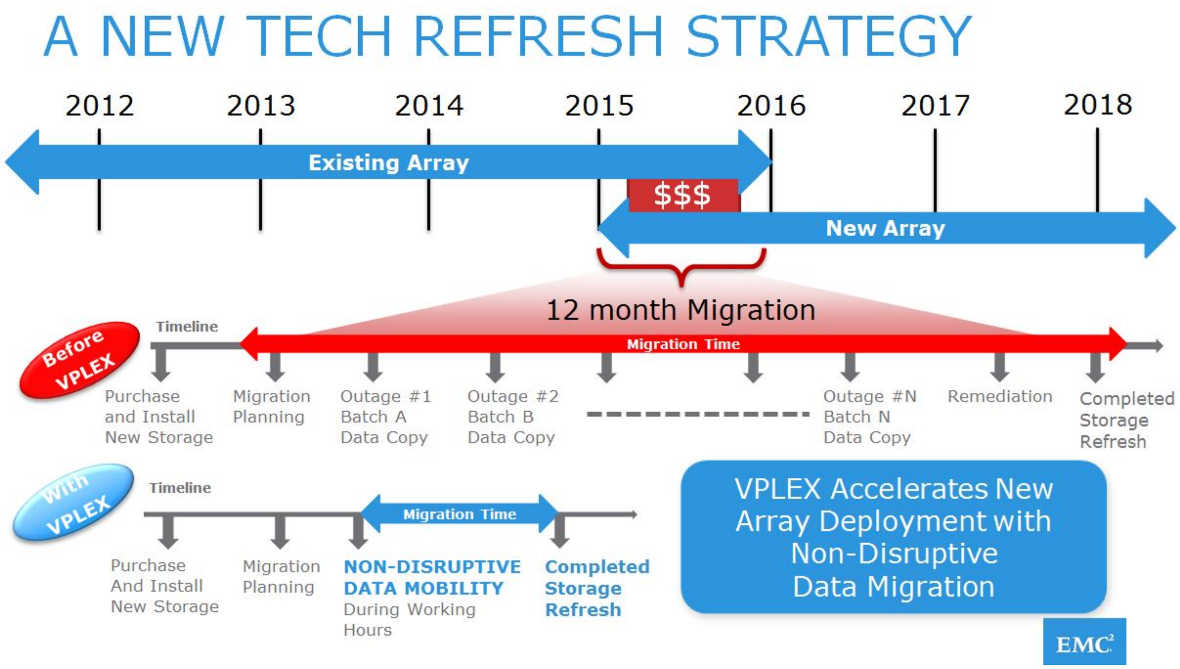

Figure 2 Migration Comparison

One of the first and most common use cases for storage virtualization in general is that it provides a simple transparent approach for array replacement. Standard migrations off an array are time consuming due to the requirement of coordinating planned outages with all necessary applications that don't inherently have the ability to have new devices provisioned and copied to without taking an outage. Additional host remediation may be required for support of the new array which may also require an outage.

VPLEX eliminates all these problems and makes the array replacement completely seamless and transparent to the servers. The applications continue to operate uninterrupted during the entire process. Host remediation is not necessary as the host continues to operate off the Virtual Volumes provisioned from VPLEX and is not aware of the change in the backend array. All host level support requirements apply only to VPLEX and there are no necessary considerations for the backend arrays as that is handled through VPLEX.

If the solution incorporates RecoverPoint and the RecoverPoint Repository, Journal and Replica volumes reside on VPLEX virtual volumes then array replacement is also completely transparent to even to RecoverPoint. This solution results in no interruption in the replication so there is no requirement to reconfigure or resynchronize the replication volumes.

1.3 Continuous Availability

1.3.1 Virtualization Architecture

Built on a foundation of scalable and continuously available multi-processor engines, EMC VPLEX is designed to seamlessly scale from small to large configurations. VPLEX resides between the servers and heterogeneous storage assets and uses a unique clustering architecture that allows servers at multiple data centers to have read/write access to shared block storage devices.

Unique characteristics of this architecture include:

- Scale-out clustering hardware lets you start small and grow big with predictable service levels
- Advanced data caching utilizes large-scale cache to improve performance and reduce I/O latency and array contention
- Distributed cache coherence for automatic sharing, balancing, and failover of I/O across the cluster
- Consistent view of one or more LUNs across VPLEX clusters (within a data center or across synchronous distances) enabling new models of continuous availability and workload relocation

With a unique scale-up and scale-out architecture, VPLEX advanced data caching and distributed cache coherency provide workload resiliency, automatic sharing, balancing, and failover of storage domains, and enables both local and remote data access with predictable service levels.

EMC VPLEX has been architected for multi-site virtualization enabling federation across VPLEX Clusters. VPLEX Metro supports max 10ms RTT, FC or 10 GigE connectivity. The nature of the architecture will enable more than two sites to be connected in the future.

EMC VPLEX uses a VMware Virtual machine located within a separate failure domain to provide a VPLEX Witness between VPLEX Clusters that are part of a distributed/federated solution. This third site needs only IP connectivity to the VPLEX sites and a 3-way VPN will be established between the VPLEX management servers and the VPLEX Witness.

Many solutions require a third site, with a FC LUN acting as the quorum disk. This must be accessible from the solution’s node in each site resulting in additional storage and link costs.

1.3.2 Storage/Service Availability

Each VPLEX site has a local VPLEX Cluster and physical storage and hosts are connected to that VPLEX Cluster. The VPLEX Clusters themselves are interconnected across the sites to enable continuous availability. A device is taken from each of the VPLEX Clusters to create a distributed RAID 1 virtual volume. Hosts connected in Site A actively use the storage I/O capability of the storage in Site A, Hosts in Site B actively use the storage I/O capability of the storage in Site B.

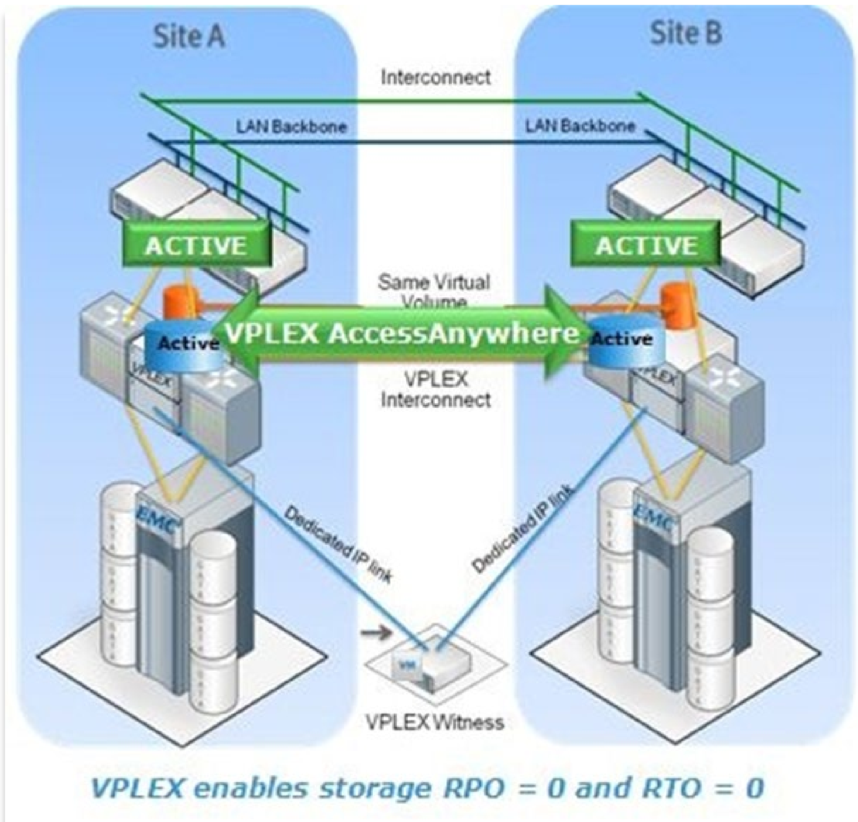

Figure 3 Continuous Availability Architecture

VPLEX distributed volumes are available from either VPLEX cluster and have the same LUN and storage identifiers when exposed from each cluster, enabling true concurrent read/write access across sites. Essentially the distributed device seen at each site is the same volume with a global visibility.

When using a distributed virtual volume across two VPLEX Clusters, if the storage in one of the sites is lost, all hosts continue to have access to the distributed virtual volume, with no disruption. VPLEX services all read/write traffic through the remote mirror leg at the other site.

2 VPLEX Components

2.1 VPLEX Engine (VS2 and VS6)

The VS2 hardware is on 2U engines and will be detailed below. The VS6 hardware is on 4U engines. The VS2 Directors are positioned side by side and the VS6 Directors are one above the other with Director A located at the bottom and Director B located above and inverted.

The following figure shows the front and rear views of both VPLEX VS2 and VS6 Engines.

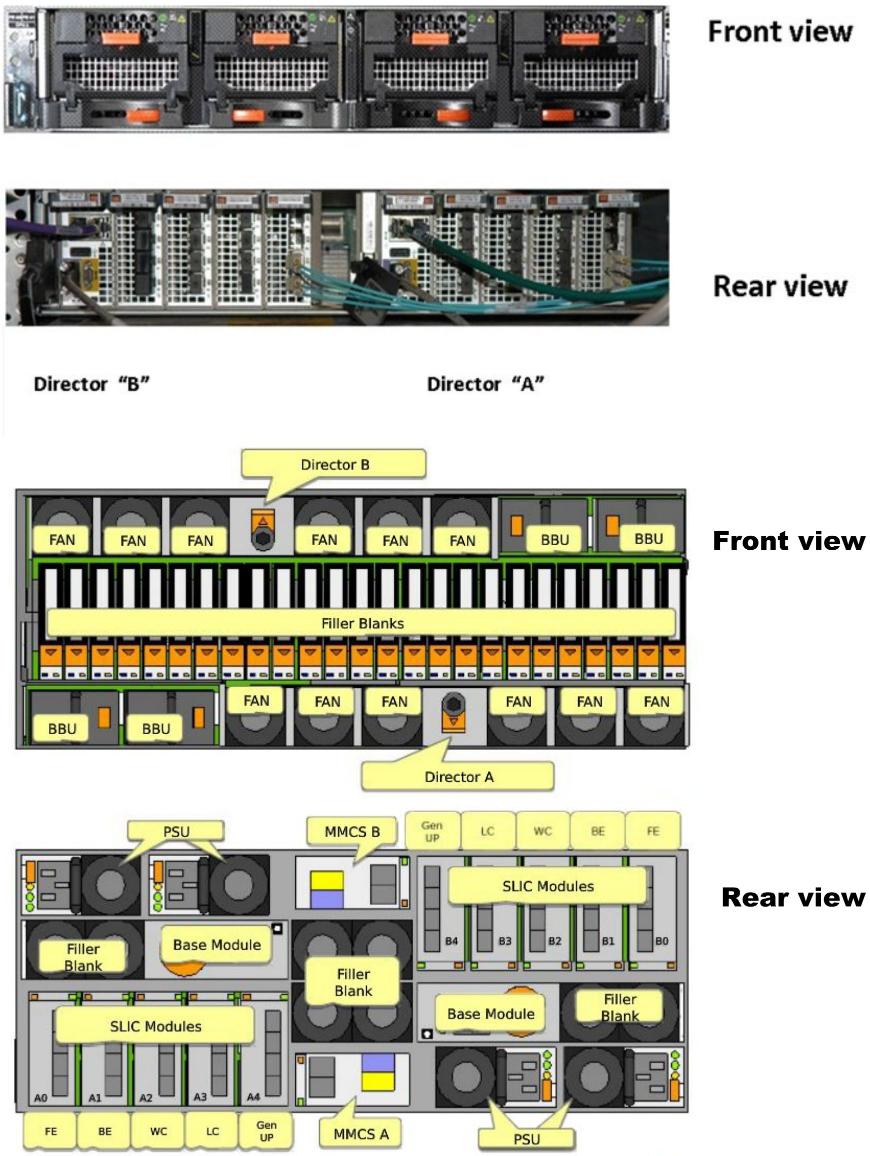

Figure 4 Front and rear view of a VPLEX VS2 and VS6 Engines

2.2 Connectivity and I/O Paths

This section covers the hardware connectivity best practices for connecting to the SAN. The practices mentioned below are based on Dual Fabric SAN, which is Industry best practice. We'll discuss Host and Array connectivity. The VPLEX hardware is designed with a standard preconfigured port arrangement that is not reconfigurable with the exception of WAN COM. The VPLEX hardware must be ordered as a Local or Metro. VPLEX Metro hardware is pre-configured with FC or 10 Gigabit Ethernet WAN connectivity from the factory and does not offer both solutions in the same configuration. The initial release of the VS6 Metro hardware only supports Fibre Channel WAN COM. A subsequent service pack release will add support for 10 Gigabit Ethernet WAN. A VPLEX Local configuration will be shipped without WAN COM modules and can be added after if the customer chooses to scale out non-disruptively to a VPLEX Metro configuration.

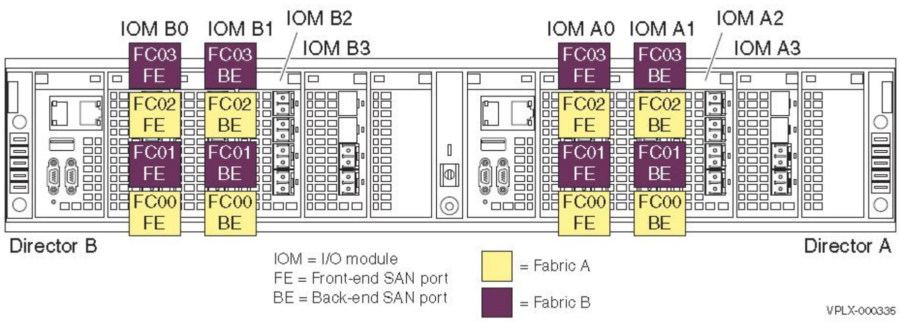

Figure 5 Preconfigured port assignments - VS2

Director A and Director B each have four I/O modules. I/O modules A0 and B0 are configured for host connectivity and are identified as frontend while the A1 and B1 are configured for array connectivity identified as backend. The frontend ports will log in to the fabrics and present themselves as targets for zoning to the host initiators and the backend ports will log in to the fabrics as initiators to be used for zoning to the array targets. Each director will connect to both SAN fabrics with both frontend and backend ports. Array direct connect is also supported however limiting. Special consideration must be used if this option is required.

The I/O modules in A2 and B2 are for WAN connectivity. This slot may be populated by a four port FC module or a two port 10 GigE for VPLEX Metro configurations. VPLEX Local configurations will ship with filler blanks in slots A2 and B2 and may be added in the field for connecting to another net new cluster for Metro upgrades. The I/O modules in slots A3 and B3 are populated with FC modules for Local Com and will only use the bottom two ports.

Note: VS6 hardware have the WAN COM and Local COM modules opposite from the VS2 hardware.

The FC WAN Com ports will be connected to dual separate backbone fabrics or networks that span the two sites. This allows for data flow between the two VPLEX clusters in a Metro configuration without requiring a merged fabric between the two sites. Dual fabrics that currently span the two sites are also supported but not required. The 10 GigE I/O modules will be connected to dual networks consisting of the same QoS. All A2-FC00 and B2-FC00 ports or A2-XG00 and B2-XG00 ports from both clusters will connect to one fabric or network and all A2-FC01 and B2-FC01 ports for or A2-XG01 and B2-XG01 ports will connect to the other fabric or network (A3 and B3 for VS6). This provides a redundant network capability where each director on one cluster communicates with all the directors on the other site even in the event of a fabric or network failure. For VPLEX Metro, each director's WAN COM ports on one cluster must see all of the director's WAN COM ports within the same port group on the other cluster across two different pipes. This applies in both directions.

When configuring the VPLEX Cluster cabling and zoning, the general rule is to use a configuration that provides the best combination of simplicity and redundancy. In many instances connectivity can be configured to varying degrees of redundancy. However, there are some minimal requirements that must be adhered to for support of features like NDU. Various requirements and recommendations are outlined below for connectivity with a VPLEX Cluster.

Frontend (FE) ports provide connectivity to the host adapters also known as host initiator ports. Backend (BE) ports provide connectivity to storage arrays ports known as target ports or FA's.

Do not confuse the usage of ports and initiator ports within documentation. Any general reference to a port should be a port on a VPLEX director. All references to HBA ports on a host should use the term “initiator port”. VPLEX Metro section has more specific discussion of cluster-to-cluster connectivity.

2.3 General Information

Official documents as may be found in the VPLEX Procedure Generator refer to a “minimal config” and describe how to connect it to bring up the least possible “host” connectivity. While this is adequate to demonstrate the features within VPLEX for a Proof of Concept or use within a Test/Dev environment it should not be implemented in a full production environment. As clearly stated within the documentation for minimal config this is not a “Continuously Available” solution. Solutions should not be introduced into production environments that are not HA. Also, this “minimal config” documentation is specific to host connectivity. Please do not try to apply this concept to backend array connectivity. The requirements for backend must allow for connectivity to both fabrics for dual path connectivity to all backend storage volumes from each director. Direct connect for array connectivity is supported however not recommended as it limits scalability, tech refresh and hardware generational non-disruptive upgrades.

Frontend direct connect is not supported.

The following are recommended:

- Dual fabric designs for fabric redundancy and HA should be implemented to avoid a single point of failure. This provides data access even in the event of a full fabric outage.
- Each VPLEX director will physically connect to both fabrics for both host (front-end) and storage (back-end) connectivity. Hosts will connect to both an A director and B director from both fabrics and across engines for the supported HA level of connectivity as required with the NDU pre-checks.

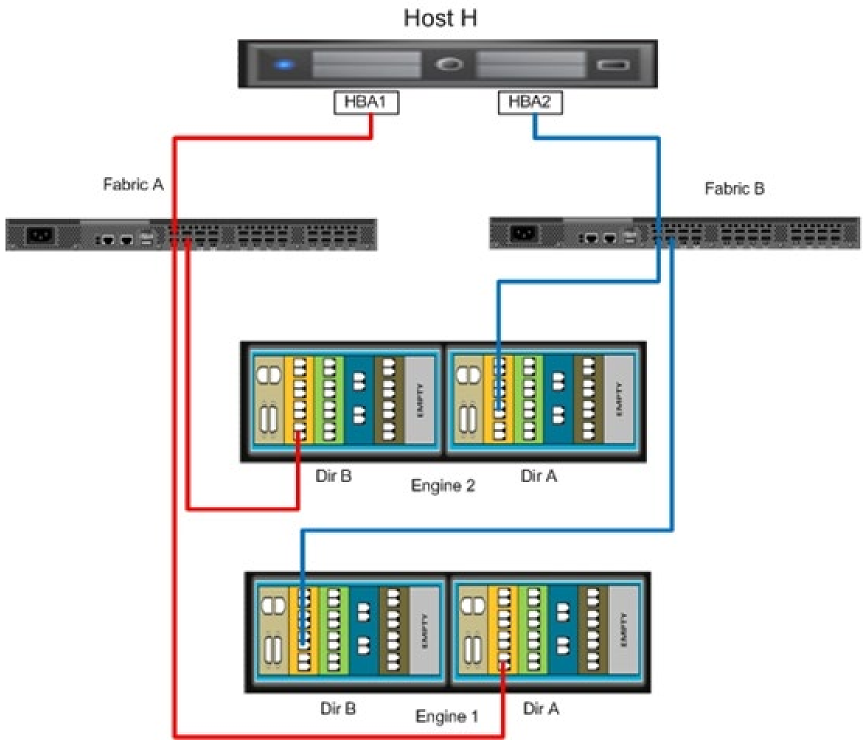

Figure 6 Continuous availability front-end config

- Back-end connectivity checks verify that there are two paths to each LUN from each director. This assures that the number of active paths (reported by the ndu pre-check command) for the LUN is greater than or equal to two active paths and less than or equal to four active paths. This check assures that there are at least two unique initiators and two unique targets in the set of paths to a LUN from each director. These backend paths must be configured across both fabrics as well. No volume is to be presented over a single fabric to any director as this is a single point of failure.
- Fabric zoning should consist of a set of zones, each with a single initiator and up to 16 targets.
- Avoid incorrect FC port speed between the fabric and VPLEX. Use highest possible bandwidth to match the VPLEX maximum port speed and use dedicated port speeds i.e. do not use oversubscribed ports on SAN switches.
- Each VPLEX director has the capability of connecting both FE and BE I/O modules to both fabrics with multiple ports. The ports connected to on the SAN should be on different blades or switches so a single blade or switch failure won’t cause loss of access on that fabric overall. A good design will group VPLEX BE ports with Array ports that will be provisioning groups of devices to those VPLEX BE ports in such a way as to minimize traffic across blades.
- Each VPLEX director will physically connect to both fabrics for both host (front-end) and storage (back-end) connectivity. Hosts will connect to both an A director and B director from both fabrics and across engines for the supported HA level of connectivity as required with the NDU pre-checks.

3 System Volumes

3.1 Metadata Volumes

A metadata volume contains information specific to a VPLEX Cluster such as virtual-to-physical mapping information, data about the devices and virtual volumes, system configuration settings, and other system information. Metadata volumes are created during system installation. However, you may need to create a metadata volume if, for example, you are migrating to a new array.

Metadata volumes consistency should be periodically checked using the 'meta-volume verify-on-disk-consistency –style short -c <cluster>' command

---

An active meta-volume with an inconsistent on-disk state can lead to a data unavailability (DU) during NDU.
Best practice is to upgrade immediately after passing this meta-volume consistency

---

check.

Note the following:

- A metadata volume is the only object you create on a storage volume without claiming it first
- You create metadata volumes directly on storage volumes, not extents or devices
- Metadata volumes should be on storage volumes that have underlying redundant properties such as RAID 1 or RAID 5
- Metadata volumes must be mirrored between different storage arrays whenever more than one array is configured to a VPLEX Cluster – NDU requirement
- Two Metadata volume backups must be configured for every VPLEX Cluster and must be placed on different arrays whenever more than one array is configured to a VPLEX Cluster – NDU requirement
- Installations that were initially configured with a single array per VPLEX cluster must move a Metadata mirror leg and a backup copy to the second array once provisioned
- Metadata volumes are written to at the time of a configuration change and read from only during the boot of each director
- Metadata volumes are supported on Thin/Virtually Provisioned devices. It is required for those devices to be fully allocated
- Meta volume names can be a max of 39 characters
- Metadata volumes provisioned on Thin devices have the additional requirement to fully allocate the device so that over allocation of the thin pool won't have adverse effects to the metadata volume should additional updates be required.

3.2 Backup Policies and Planning

You need to have a metadata backup you can recover from.

Plan on the following:

- Spare volumes for each cluster to hold backups — You need to rotate a minimum of two backups per VPLEX Cluster
- A system-wide scheduled backup done at regular times — A single cluster backup for a VPLEX Metro is not useful
- On-demand backups before/after major reconfigurations and/or migrations

---

Note: Renaming Metadata Backup volumes is not supported.

Metadata volumes are required to be zeroed out. Follow the procedure outlined in VPLEX Procedure Generator to zero the disks to be used for the Metadata Volumes and Metadata volume backups. Additionally, writing zeros to thin devices will force them to be fully allocated which is necessary for VPLEX supportability with both Metadata volumes and logging volumes.

For additional information, please refer to VPLEX with GeoSynchrony CLI Guide found on SolVe Desktop.

3.3 Logging Volumes

A prerequisite to creating a distributed device, or a remote device, is that you must have a logging volume at each cluster. Logging volumes keep track of any blocks written during an inter-cluster link failure. After a link is restored, the system uses the information in logging volumes to synchronize the distributed devices by sending only changed block regions across the link. Each DR1 requires 2 bitmap logs at each site. One for writes initiated remotely that fail locally (gives array death protection) and one for writes succeed locally but fail remotely. Upon restore (of array or link) one log is used to push data and one is used to pull data.

1. The first bitmap log is used if the clusters partition, for writes on the winning site. You will use this log to 'send' the blocks for the log rebuild
2. The 2nd bitmap log is used if your local array fails. Since writes to the local leg will fail, you mark the 2nd bitmap log. You then use this in a log rebuild to 'pull' data from the remote cluster.

If a logging volume is not created, every inter-cluster link-failure could cause a full resynchronization of every distributed device in the system. The logging volume must be large enough to contain one bit for every page of distributed storage space. The calculation for determining logging volume size is: 10 GB protects 320TB/N [n is the number of clusters]. Consequently, you need approximately 10 GB of logging volume space for every 160 TB of distributed devices in a VPLEX Metro. Logging volumes are not used for VPLEX Local configurations and are not used for local mirrors.

The logging volume receives a large amount of I/O during and after link outages, therefore it must be able to handle I/O quickly and efficiently. EMC recommends that you stripe it across several disks to accommodate the I/O volume, and that you also mirror it, since this is important data. EMC also recommends placing the logging volume on separate physical spindles than the storage volumes that it is logging against.

Because logging volumes are critical to the functioning of the system, the best practice is to mirror them across two or more back-end arrays to eliminate the possibility of data loss on these volumes. In addition, they can be striped internally on the back-end arrays. The RAID level within the array should be considered for high performance. When a failure happens that invoke the use of the logging volume then they will experience very high read/write I/O activity during the outage. During normal healthy system activity, logging volumes are not used.

If one array's data may, in the future, be migrated to another array, then the arrays used to mirror the logging volumes should be chosen such that they will not be required to migrate at the same time.

You can have more than one logging volume and can select which logging volume is used for which distributed device.

Logging volumes are supported on thin devices however it is a requirement that the thin device be fully allocated. For similar reasons as the metadata volume requirement, logging volumes are active during critical times and must have the highest level of availability during that time.

4 Requirements vs. Recommendations Table

The following table represents recommendations and requirements for production and Proof of Concept environments. Be sure to adhere to the requirements noted for the type of environment the VPLEX is deployed in.

Table 1

	Production Environment	**Test or Proof of Concept Environment**	**Notes**
Dual Fabric	Requirement for High Availability	Requirement if the tests involve High Availability	Dual Fabrics are a general best practice.
Dual HBA	Required	Required	Single initiator hosts are not supported and a dual port HBA is a single point of failure also.
Initiator connected to both an A and B Director	Required	Recommended	For a Production Environment, it is also required that the connectivity for each initiator span engines in a dual or quad engine VPLEX Cluster.
Four "active" backend paths per Director per Storage Volume	Recommended but also It's a requirement to not have more than 4 active paths per director per storage volume	Recommended	This is the maximum number of "active" paths. An active/passive or ALUA array will have a maximum of four active and four passive or non-preferred paths making eight in all.

	Production Environment	Test or Proof of Concept Environment	Notes
Two "active" backend paths per Director per Storage Volume	Required	Required	This is a minimum requirement in NDU which dictates that two VPLEX Director backend ports will be connected to two array ports per Storage Volume. Depending on workload, size of environment and array type, four "active" path configurations have proven to alleviate performance issues and therefore are recommended over the minimum of two active paths per director per storage volume. Try to avoid only two path connectivity in production environments. Each director in turn should be connected to a different set of array ports. The director connectivity should span array components such as engines or storage controllers to avoid single points of failure.
Host Direct Connected to VPLEX Directors	Not Supported	Not Supported	Host direct connect to a VPLEX Director is never supported.
Arrays direct connected to VPLEX Directors	Not Recommended but Supported	Supported	Array direct connect is supported but extremely limited in scale which is why it is not recommended for a production environment.

	Production Environment	Test or Proof of Concept Environment	Notes
WAN COM single port connectivity	Not Supported	Not Supported	Two ports on the WAN COM for each Director must be configured each in their separate port groups. Fibre Channel WAN COM (Metro Only) is also supported with all four ports each in their own port group.
Metadata and Logging Volume	It is required that metadata and metadata backups are configured across arrays at the local site if more than one array is present. If the site were built originally with a single array and another array were to be added at a later time then it is required to move one leg of the metadata volume and one backup to the new array.	It is required that metadata and metadata backups are configured across arrays at the local site if more than one array is present. A standard test during a POC is to perform an NDU. It would be undesirable to have to use the --skip option if not needed.	While it is a requirement for metadata and metadata backups to be configured across arrays, it is highly recommended to mirror logging volumes across arrays as well. Loss of the array that contains the logging volumes would result in additional overhead of full rebuilds after the array came back up.
Metadata and Logging Volumes on thin provisioned devices	Supported but required that the thin provisioned device be fully allocated	supported	Devices not fully allocated present the opportunity of causing VPLEX to fail should the array be oversubscribed and reach full capacity preventing any further expansion of the Metadata and logging volumes.
Host Cross-Cluster Connect	Supported with VPLEX Witness required	Supported with VPLEX Witness required	VPLEX Witness is a hard requirement for Host Cross-Cluster Connect regardless of the type of environment. The auto resume attribute on all consistency groups must be set to true as an additional requirement.

	Production Environment	Test or Proof of Concept Environment	Notes
Array Cross-Cluster Connect	Only supported if both sites are within 1ms latency from each other and strictly for the purpose of adding protection to metadata and logging volumes. Storage Volumes are not supported connected to both clusters	Only supported if both sites are within 1ms latency from each other and strictly for the purpose of adding protection to metadata and logging volumes. Storage Volumes are not supported connected to both clusters	Connecting an array to both sides of a VPLEX Metro is not supported if the sites exceed 1ms latency from each other. If done then extreme caution must be taken not to share the same devices to both clusters. Also, be cautious if evaluating performance or fault injection tests with such a configuration.
VPLEX Witness	Requirement for High Availability.	Optional but should mirror what will be in production	VPLEX Witness is designed to work with a VPLEX Metro. It is not implemented with a VPLEX Local. VPLEX Witness has proven to be such a valuable enhancement that it should be considered a requirement. VPLEX Witness must never be co-located with either of the VPLEX Clusters that it is monitoring. VPLEX Witness is required to have direct access to both sites. Never configure witness to access both sites through a single site in an “L” configuration.

This page intentionally left blank.