EMC® VPLEX™METRO CONTINUOUS AVAILABILITY AND CLUSTER WITNESS

IMPLEMENTATION AND DESIGN GUIDE

ABSTRACT

This technical note is targeted for EMC field personnel, partners, and customers who will be configuring, installing, supporting and managing EMC VPLEX Metro for Continuous Availability. This document is designed to show all users how VPLEX Metro is deployed with and without VPLEX Witness and explains how to achieve seven 9's availability through proper configuration.

February 2015

REDEFINE

EMC WHITE PAPER

EMC²

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

Part Number H13879

TABLE OF CONTENTS

Preface

This EMC Engineering technical note describes and provides an insightful discussion on how implementation of VPLEX Metro with the inclusion of the VPLEX Witness will give customers seven 9's availability and assurance of business continuity they expect.

As part of an effort to improve and enhance the performance and capabilities of its product lines, EMC periodically releases revisions of its hardware and software. Therefore, some functions described in this document may not be supported by all versions of the software or hardware currently in use. For the most up-to-date information on product features, refer to your product release notes. If a product does not function properly or does not function as described in this document, please contact your EMC representative.

AUDIENCE

This white paper is intended for the following readers:

- EMC Pre-Sales Organization for outlining and describing the architecture for their customers prior to purchase.
- EMC Global Services Application Support for effectively introducing the product into the environment and assuring that the implementation is specifically oriented to the customers' needs and negates any possible DU/DL and/or application failure or misunderstanding of such failures.
- EMC customers interested in deploying VPLEX or have deployed VPLEX and need a solid understanding of how VPLEX Metro and VPLEX Witness behaves under different conditions
- EMC Support reference when issues do get reported, so that they can be quickly triaged under normal conditions as described in the technical playbook.

It is expected this document be shared with customers as a tool for guiding right decisions while implementing EMC VPLEX Metro for Continuous Availability. Readers should be familiar with VPLEX as a product, has previous training and are responsible for the success of product implementation.

RELATED PUBLICATIONS FOR REFERENCE

The following documents are located on EMC.COM and should be used as additional reference depending on environment and application focus:

- EMC VPLEX AND VMWARE TECHNICAL CONTINUOUS AVAILABILITY FAILURE HANDLING
- EMC OVERVIEW AND GENERAL BEST PRACTICES
- EMC VPLEX SAN CONNECTIVITY
- EMC VPLEX HOST MULTIPATHING
- EMC VPLEX METRO CROSS-CONNECT HOST CLUSTERING
- ORACLE REAL APPLICATION CLUSTERS (RAC) ON EXTENDED DISTANCE CLUSTERS WITH EMC VPLEX METRO BEST PRACTICES PLANNING

VPLEX Metro Deployment Scenarios

VPLEX Metro 5.0 (and above) introduced high availability concepts beyond what is traditionally known as physical high availability. Introduction of the "VPLEX Witness" to a high availability environment, allows the VPLEX solution to increase the overall availability of the environment by arbitrating a pure communication failure between two primary sites and a true site failure in a multi-site architecture. EMC VPLEX is the first product to bring to market the features and functionality provided by VPLEX Witness prevents failures and asserts the activity between clusters in a multi-site architecture.

Through this technical note, administrators and customers gain an understanding of the high availability solution that VPLEX provides them:

- VPLEX witness is a game changer to the way continuous availability is achieved
- Active/active use of both of their data centers
- Increased availability for their applications (no single points of storage failure, auto-restart)
- Fully automatic failure handling
- Better resource utilization
- Lower CapEx and lower OpEx as a result

Broadly speaking, when one considers legacy environments one typically sees "highly" available designs or active/active applications implemented within a data center, and disaster recovery or replication type functionality deployed between data centers. One of the main reasons for this is that within data centers components generally operate in active/active with automatic failover whereas between data centers legacy replication technologies use active/passive techniques which 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 now combined since it enables High Availability (HA) applications to be stretched beyond the data center walls.

Since the original publication of this technical note (formerly, TechBook), VPLEX Metro implementations have become the most implemented option for customers using VPLEX. As customers deploy Metro, it is indicative and required the VPLEX Witness be installed to enable all out continuous availability consumers to achieve seven 9's availability.

PLANNED APPLICATION MOBILITY

An online planned application mobility event is defined as when clustered applications or virtual machines can be moved fully online without disruption from one location to another in either the same or remote data center. This type of movement can only be performed when all components that participate in this movement are available (e.g., the running state of the application or VM exists in volatile memory which would not be the case if an active site has failed) and if all participating hosts have read/write access at both location to the same block storage. Additional a mechanism is required to transition volatile memory data from one system/host to another. When performing planned online mobility jobs over distance a prerequisite is the use of an active/active underlying storage replication solution (VPLEX Metro only at this publication). An example of this online application mobility would be VMware vMotion where a virtual machine would need to be fully operational before it can be moved. It may sound obvious but if the VM was offline then movement could not be performed online (This is important to understand and is the key difference over application restart). When vMotion is executed all live components that are required to make the VM function are copied elsewhere in the background before cutting the VM over.

Since these types of mobility tasks are totally seamless to the user some of the use cases associated are for disaster avoidance where an application or VM can be moved ahead of a disaster (such as, Hurricane, Tsunami, etc.) as the running state is available to be copied, or in other cases it can be used to enable the ability to load balance across multiple systems or even data centers.
Due to the need for the running state to be available for these types of relocations these movements are always deemed planned activities.

DISASTER RESTART

Disaster restart is where an application or service is re-started in another location after a failure (be it on a different server or data center) and will typically interrupt the service/application during the failover.

A good example of this technology would be a VMware HA Cluster configured over two geographically dispersed sites using VPLEX Metro where a cluster will be formed over a number of ESX servers and either single or multiple virtual machines can run on any of the ESX servers within the cluster.
If for some reason an active ESX server were to fail (perhaps due to site failure) then the VM can be re-started on a remaining ESX server within the cluster at the remote site as the data store where it was running spans the two locations since it is configured on a VPLEX Metro distributed volume. This would be deemed an unplanned failover which will incur a small outage of the application since the running state of the VM was lost when the ESX server failed meaning the service will be unavailable until the VM has restarted elsewhere. Although comparing a planned application mobility event to an unplanned disaster restart will result in the same outcome (i.e., a service relocating elsewhere) it can now be seen that there is a big difference since the planned mobility job keeps the application online during the relocation whereas the disaster restart will result in the application being offline during the relocation as a restart is conducted.

Compared to active/active technologies the use of legacy active/passive type solutions in these restart scenarios would typically require an extra step over and above standard application failover since a storage failover would also be required (i.e. changing the status of write disabled remote copy to read/write and reversing replication direction flow). This is where VPLEX can assist greatly since it is active/active therefore, in most cases, no manual intervention at the storage layer is required, this greatly reduces the complexity of a DR failover solution. If best practices for physical high available and redundant hardware connectivity are followed the value of VPLEX Witness will truly provide customers with "Absolute" availability!

Other applications aside from VMware that benefit from planned and unplanned events with VPLEX Metro are Oracle RAC, Microsoft Hyper-V, RedHat Linux, Power HA to name some; but all applications benefit while storage is accessible. In addition refer to the VPLEX EMC Simple Support Matrix for up to date supported applications, storage, etc.

VPLEX METRO ARCHITECTURE

VPLEX METRO

VPLEX Metro systems contain two clusters, each cluster having one, two, or four engines. The clusters in a VPLEX Metro deployment need not have the same number of engines. For example, a VPLEX Metro system could be composed of one cluster with two engines and the other with four.

The two clusters of a VPLEX Metro must be deployed within synchronous communication distance of each other (about 5 -10ms of RTT communication latency). VPLEX Metro systems are often deployed to span between two data centers that are close together but they can also be deployed within a single data center for applications requiring a high degree of local availability.

VPLEX WITNESS

With VPLEX Metro, VPLEX virtual volumes can be mirrored between the VPLEX clusters, allowing a host to have access to the data through either cluster. This provides added resiliency in the case of an entire cluster failure. In such a deployment, on a per-consistency group basis, one cluster is designated as the preferred cluster for data availability which will be explained in the next few sections in detail. But at a high level; should the redundant communication between the VPLEX clusters be lost and connectivity with the VPLEX Witness retained, the VPLEX Witness will indicate to the clusters that the preferred cluster should continue providing service to the volumes in the consistency group. In this situation, the non-preferred cluster will stop servicing the volumes, until such time as the link is restored, and the mirrors are re-established. Should the preferred cluster of a consistency group fail, the VPLEX Witness will indicate this failure to the non-preferred cluster, which will continue to provide access to the volumes in the group. Likewise, in the event of the failure of the non-preferred cluster, the Witness will direct the preferred cluster to continue to service the volumes. This prevents a partition between the two clusters from allowing the state of the volumes to diverge; this avoids the well-known split-brain problem.

The use of the Witness is required for Continuous Availability since it provides zero need for RTO of data in the presence of these failures. When the VPLEX Witness is not deployed, the system will suspend I/O to a volume when that volume's preferred cluster fail; again to be explained in section [VPLEX METRO WITHOUT VPLEX WITNESS].

VPLEX Witness functionality applies only to distributed volumes that are placed in consistency groups. Distributed volumes that are not placed in a consistency group have their own independent bias settings as determined by the administrator during initial set-up. These volumes will have their I/O suspended when their preferred cluster fails as previously mentioned.

VPLEX Witness consists of two components:

- Cluster Witness Server – a VM installed on a customer's ESX server connected to both clusters in a VPLEX Metro or Geo configuration.
- Cluster Witness CLI – CLI commands to configure, manage, and diagnose VPLEX Witness and its functionality.

VPLEX Witness is installed as a virtual machine (VM) operating in a customer's ESX server deployed in a failure domain separate from either of the VPLEX clusters. This ensures that the VM is not affected by faults that impact the VPLEX clusters. A failure domain is a collection of entities affected by the same fault.

If you are installing the VPLEX Witness on a configuration running GeoSynchrony 5.1 or later, before deciding to install and enable VPLEX Witness, read the VPLEX Administration Guide to understand VPLEX Witness operation in a VPLEX Geo or VPLEX Metro environment.

It is very important to deploy the VPLEX Witness into a failure domain that is independent of each of the failure domains containing the two VPLEX clusters, to ensure that a single failure impacts no more than one of these entities. Customers who have more than (2) failure domains should operate VPLEX Witness on-prem. Customers have and will ask if it is possible to deploy VPLEX Witness at a service provider off-prem or "cloud" based VM. The answer is yes, supported cloud-based deployments will become available over time and supported by EMC in future release cycles.

VPLEX Witness connects to both VPLEX clusters over a VPN tunnel over the IP management network. Low bandwidth health-check heartbeats are used to observe the status of both clusters. VPLEX Witness reconciles the status reported by the clusters with its own observations and provides guidance back to the clusters, if necessary.

VPLEX Witness is applicable only to VPLEX Metro configurations1. Before the Cluster Witness Server VM is deployed, VPLEX software that supports the VPLEX Witness must be installed first. Ensure that you follow VPLEX installation instructions for Release 5.0 or later before deploying and configuring VPLEX Cluster Witness Server VM.

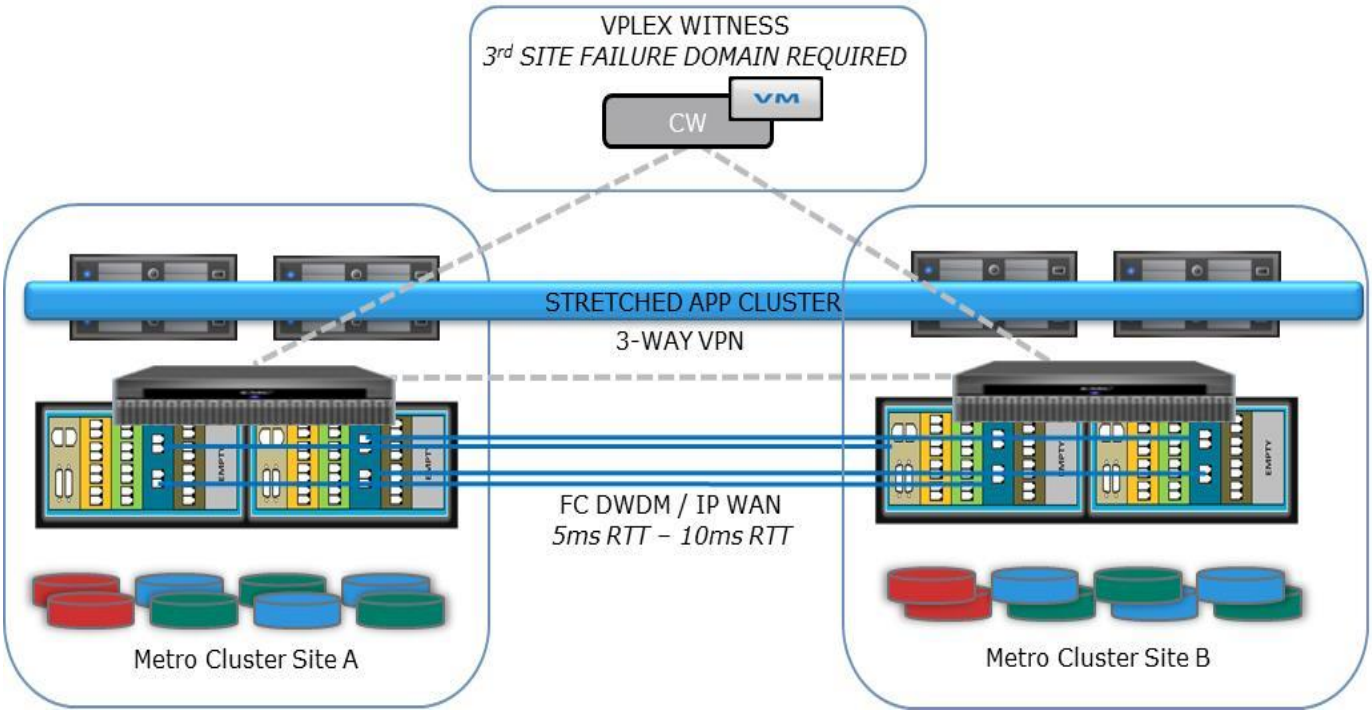


Figure 1. Standard EMC VPLEX Metro Environment

FOUNDATIONS OF VPLEX METRO BEHAVIOR AND CONSISTENCY

Where does consistency begin in the VPLEX Metro environment? It begins with the hardware redundancy, appropriation of paths; the implementation of multipath capabilities and of course application load balancing relation to writes and reads properly through cache coherency intelligence. Before one implements Metro, an understanding of how Metro cache works is essential to optimize the environment and load balance efficiently reads and writes but also highlights why Witness is such a big part of the "whole" entity that is this CA solution.

The individual memory systems of each VPLEX director are combined to form the VPLEX distributed cache. Data structures within these memories in combination with distributed algorithms achieve the coherency and consistency guarantees provided by VPLEX virtual storage. This guarantee ensures that the I/O behavior observed by hosts accessing VPLEX storage is consistent with the behavior of a traditional disk. The VPLEX distributed algorithms are designed to minimize inter-director messaging and take advantage of I/O locality in the placement of key data structures.

[1] VPLEX Geo is supported in addition however for purposes of this technical note, not covered

The design is truly distributed: Any director within a cluster is able to service an I/O request for a virtual volume served by that cluster. Each director within a cluster is exposed to the same set of physical storage volumes from the back-end arrays and has the same virtual-to-physical storage mapping metadata for its volumes. The distributed design extends across a VPLEX Metro system to provided cache coherency and consistency for the global system. This ensures that a host accesses to a distributed volume always receive the most recent consistent data for that volume. This increases speed for read intensive applications and assures that writes are being driven to the nearest path in the Initiator –Target–Lun (ITL) relationship.

VPLEX Metro uses the write-through cache mode. (Local does as well, but we focus this topic on Metro itself) With write-through caching as a write request is received from a host to a virtual volume, the data is written through to the back-end storage volume(s) that map to the volume. When the array(s) acknowledge this data, an acknowledgement is then sent back from VPLEX to the host indicating a successful write. This provides an especially strong guarantee of data durability in the case of a distributed mirror where the back-end storage volumes supporting the mirror can be placed in different data centers.

Business continuity is extended by the VPLEX Local RAID 1 technology, allowing applications to continue processing in the presence of array failures and maintenance operations. The distributed RAID 1 technology extends this protection further, allowing clustered active/active applications to leverage the capabilities of VPLEX to ride through site disasters as previously assured. The distributed RAID 1 features coupled with VPLEX's distributed coherent cache is the core technologies that provides the foundation of distributed mobility, availability, and collaboration across distance. If you think about how distribution with a single cluster behaves and inter-director messaging occurs for coherent messaging based on writes and reads; the same applies when a DR1 is created so global systems as described also occurs across distance. Stability and a level of responsibility *are owed* to this underlying infrastructure.

You the consumer OWE your data a VPLEX Witness to be implemented.

CONSISTENCY GROUPS

As previously mentioned, in order for VPLEX Witness to maintain consistency for VPLEX Metro all virtual volumes need to be placed into consistency groups. These can be carved up however is logical for the customer but in general, groupings accord to applications, IOPs and characteristics of the data should be considered. DR1s will benefit from having like array characteristics in addition as VPLEX response to IO from B/E will always be a fast as the slowest link. Each consistency group will have properties applied to it in the context of preference rules. These rules are:

- Winner (Site A) - any consistency groups that have this marking will retain activity should there be an outage at non-preferred site without witness
- Winner (Site B) – any consistency groups that have this marking will retain activity should there be an outage at non-preferred site without witness
- No Automatic Winner – there is no winner, suspension occurs at both sides without witness

Essentially, the architecture defined by having no VPLEX witness is still an Active/Passive type of environment however the dependency of the rules will dictate if IOs will or will not continue at survival site.

Once VPLEX Witness is implemented, the notion of being at the mercy of the preferred site "rule" or "static bias" no longer applies.

FAILURE HANDLING WITHOUT VPLEX WITNESS

The following section discusses several disruptive scenarios at a high level to a multiple site VPLEX Metro configuration without VPLEX Witness. The purpose of this section is to provide the customer or solutions' architect the ability to understand site failure semantics prior to the deployment of VPLEX Witness. This section is not designed to highlight flaws in high availability but to demonstrate what is ACHIEVABLE by introducing witness technology as opposed to more active/passive approach to availability.

In Figure 2., for purposes of setting example of the following exercises, this shows a Metro solution with Site A and Site B. Currently, there is no VPLEX Witness and IO is flowing actively at both sites.

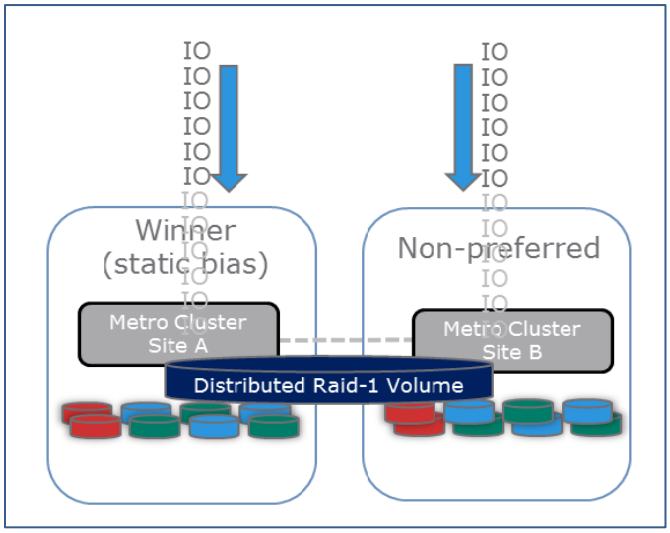


Figure 2. Sample template, no VPLEX Witness

High level Site A failure

Suppose that a clustered application was running only in Site A at the time of the incident it would now need to be restarted at the remaining Site B. Reading this document, you know this since you have an external perspective being able to see the entire diagram. However, if you were looking at this purely from Site B's perspective, all that could be deduced is that communication has been lost to Site A. Without an external independent observer of some kind, it is impossible to distinguish between full Site A failure vs. the inter-cluster link failure. Now, with VPLEX as previously described, there are rule-sets inherently configured to the volumes. In this picture, they are applied to Site A. Therefore, IO is suspended to the "winner or Bias Rules" and B will suspend.

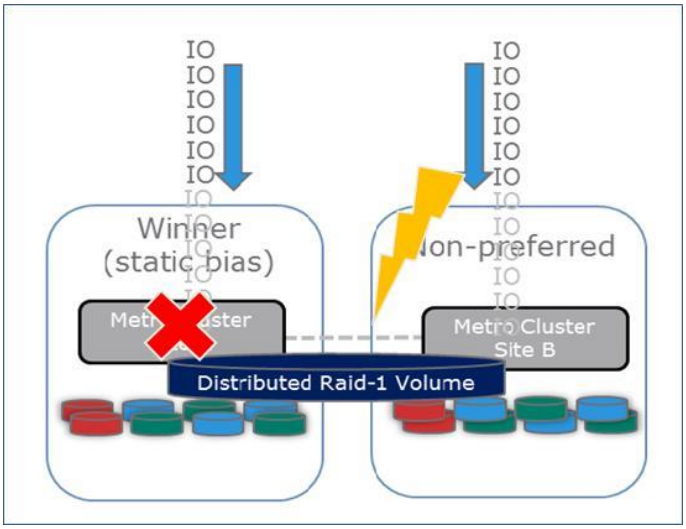


Figure 3. Site outage Site A, Bias Rules Winner A outage, B suspends

High level Site B failure

Suppose that a clustered application was running only in Site B at the time of the incident it would now need to be restarted at the remaining Site A. Reading this document, you know this since you have an external perspective being able to see the entire diagram. However, if you were looking at this purely from Site A's perspective, all that could be deduced is that communication has been lost to Site B. Without an external independent observer of some kind, it is impossible to distinguish between full Site B failure vs. the inter-cluster link failure. However, because preference rules (Bias Rules Winner A) are applied to the Site A volumes, IO will continue in this scenario, Site B will remain suspended.

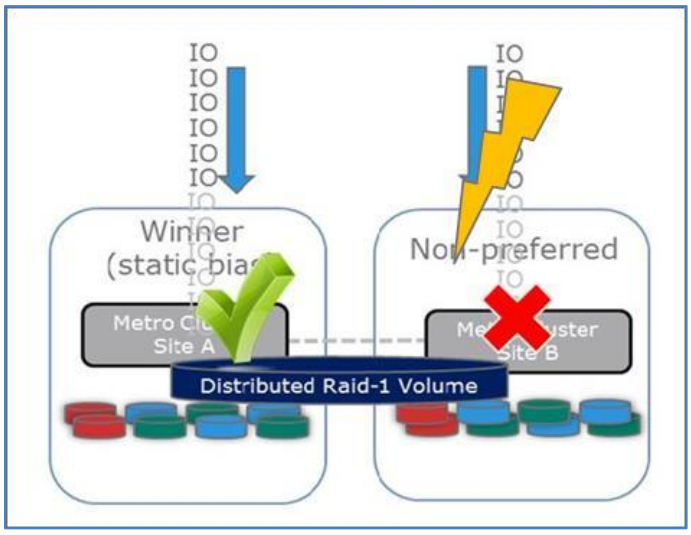


Figure 4. Site outage Site B, Bias Rules Winner A IO continues, B retains outage

Inter-cluster link failure

It is very important to have redundant stable inter-cluster links. In the event that there is an inter-site link outage on the FC DWDM or IP WAN configuration, without witness there will be suspension on the Site B non-preferred site, much like the Site B failure incident. In Site A, as in the Site B failure, it will continue to serve up IO and continue to operate but there will be no more simultaneous writes occurring until restoration of the links.

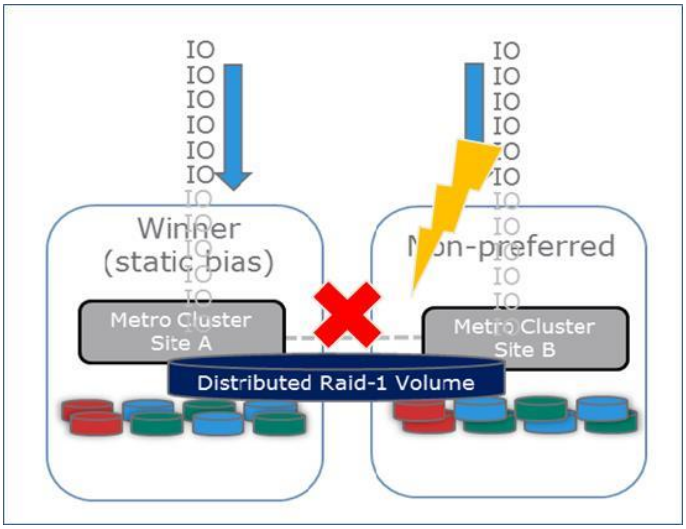


Figure 5. Inter-cluster outage, all non-preferred Site CGs suspend

FAILURE HANDLING WITH VPLEX WITNESS

VPLEX Witness failure semantics

As seen in the previous section VPLEX Witness will operate at the consistency group level for a group of distributed devices and will function in conjunction with the detach rule set (non-preferred rule) within the consistency group. Below represents the architecture of the configuration itself. You can see with the dotted line a 3-way VPN is also represented.

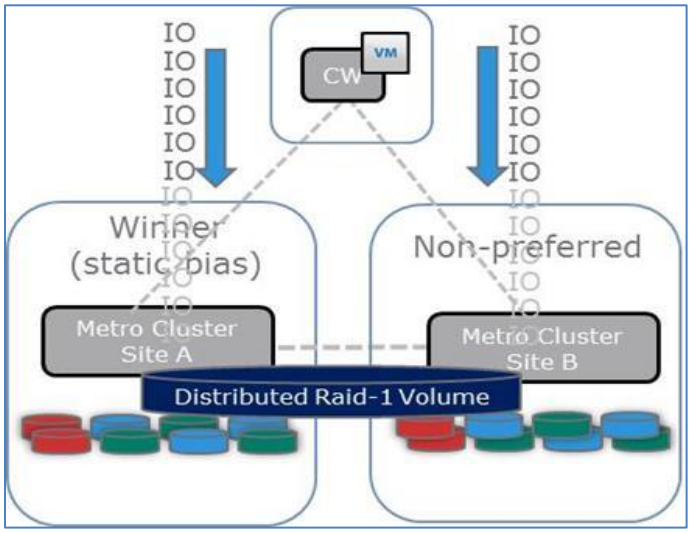


Figure 6. Metro representation with VPLEX Witness

Inter-cluster link failure with Witness

If the inter-cluster link were to fail in this scenario VPLEX Witness would still be able to communicate with both VPLEX clusters since the management network that connects the VPLEX Witness server to both of the VPLEX clusters is still operational. The behavior of the outage in this case doesn't change however from Site B being suspended. The rules will still apply as communication between the two management servers and clusters are no longer available.

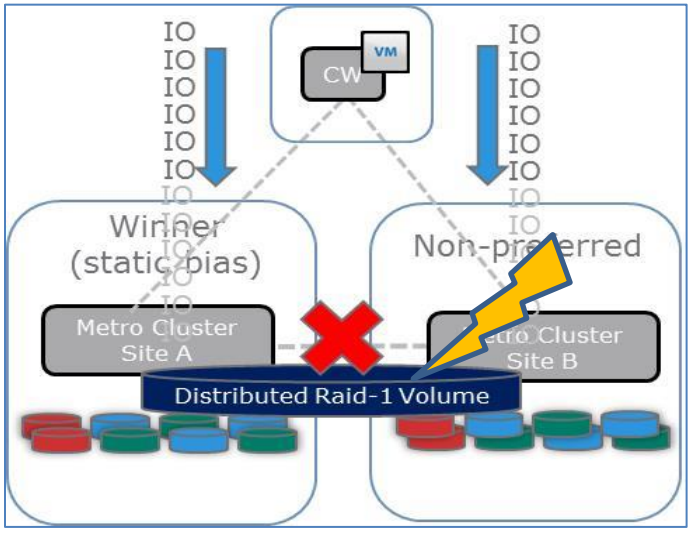


Figure 7. Inter-cluster link failure

VPLEX Witness and static preference after Cluster partition

The next example shows how VPLEX Witness can assist if you have a site failure at the preferred site. As discussed above, this type of failure without VPLEX Witness would cause the volumes in the surviving site to go offline. This is where VPLEX Witness greatly improves the outcome of this event and removes the need for manual intervention.

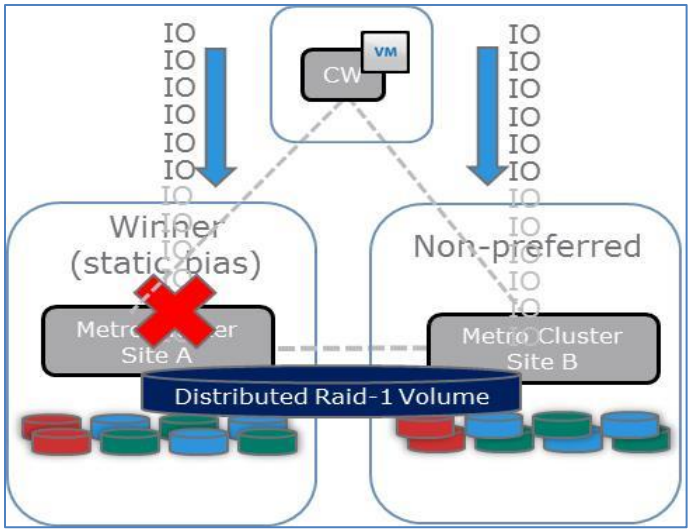


Figure 8. Preferred Site A detaches

VPLEX Witness diagram showing Cluster 2 failure

As discussed in the previous section, when a site has failed then the distributed volumes are now degraded. However, unlike our previous example where there was a site failure at the preferred site and the static preference rule was used forcing volumes into a suspend state at cluster 1, VPLEX Witness will now observe that communication is still possible to cluster 1 (but not cluster 2). Additionally since cluster 1 cannot contact cluster 2, VPLEX Witness can make an informed decision and guide cluster 1 to override the static rule set and proceed with I/O.

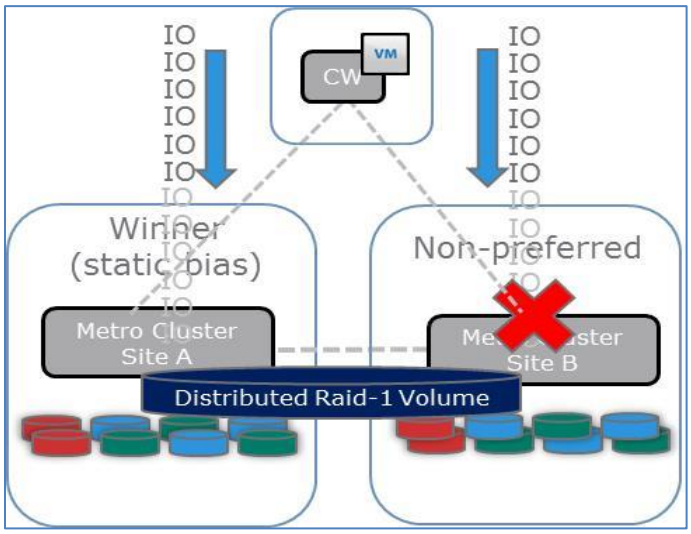


Figure 9. Non-preferred site outage

VPLEX Witness is lost – now what?

Absolutely nothing changes, again the IO will continue to flow as recovery to the witness returns. If however, a failure to either Site A or Site B occurs while Witness is unavailable then VPLEX operates against the normal preference site rules. It is important to monitor the VPLEX witness for availability via email alerts in order to maintain its integrity.

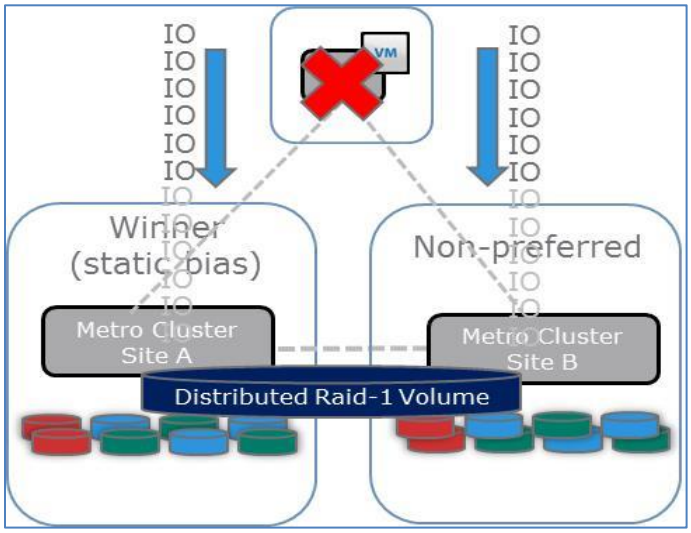
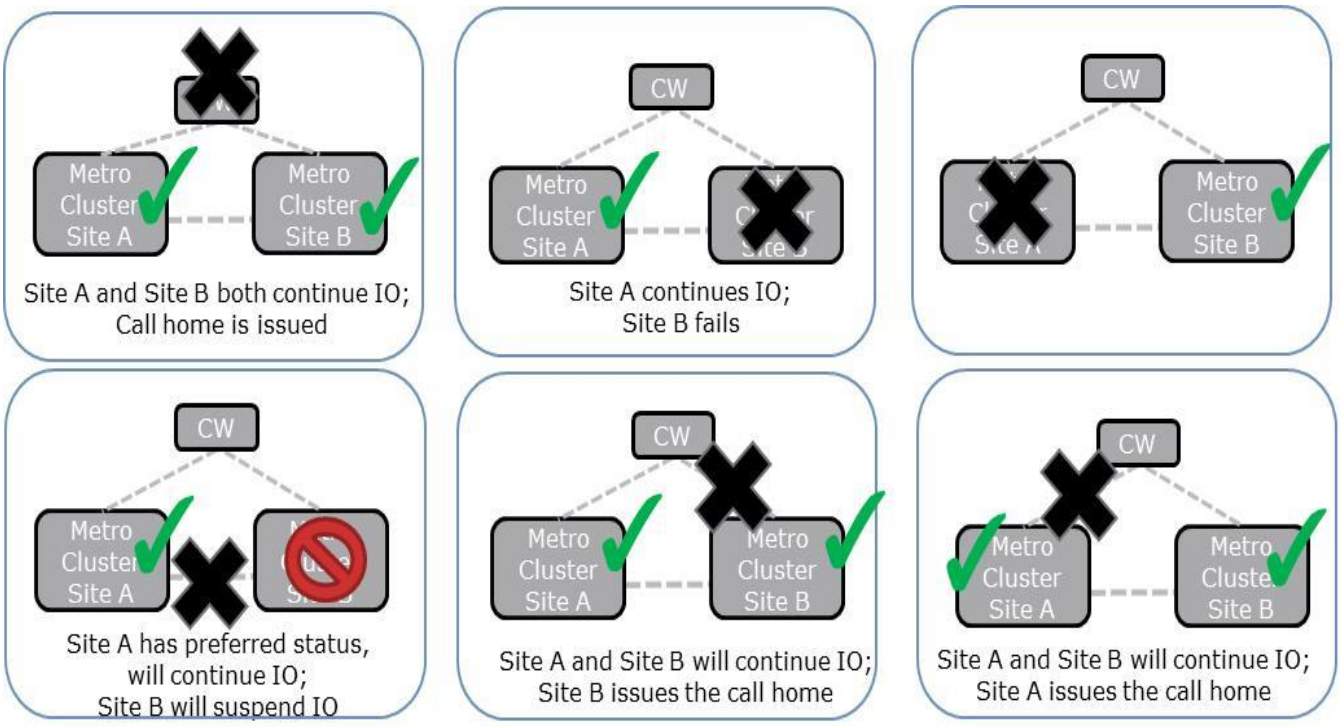


Figure 9. Loss of witness

Failure Scenario Recap with VPLEX Witness



THE A-HA! MOMENT

When it comes to achieving the highest possible of availability for storage environments and respective clustered applications, VPLEX Metro with Cluster Witness is the best option for customer in the industry to date. The amount of physical, logical and controllable redundancy and monitored behavior tied to cache coherency make the product unparalleled. The most important aspect of the implementation as depicted in the foreground of this paper entitles that the consumer is aware of the way VPLEX handles failure and what the best options are when configuring for critical business applications and sustainability of ongoing transactions and uptime.

If you did not get an "a-ha" moment from this note and still need answers, re-read it and draw it out. A fine piece of paper of bar napkin works great. Also, refer to documents mentioned in the beginning of the tech note for more information or contact your local EMC specialists.

APPENDIX

CLI EXAMPLE OUTPUTS

On systems where VPLEX Witness is deployed and configured, the VPLEX Witness CLI context appears under the root context as "cluster-witness." By default, this context is hidden and will not be visible until VPLEX Witness has been deployed by running the **cluster-Witness configure** command. Once the user deploys VPLEX Witness, the VPLEX Witness CLI context becomes visible.

The CLI context typically displays the following information:

```
VPlexcli:/> cd cluster-witness/
VPlexcli:/cluster-witness> ls
Attributes:
Name Value
------------- -------------
admin-state enabled
private-ip-address 128.221.254.3
public-ip-address 10.31.25.45
Contexts:
components
VPlexcli:/cluster-witness> ll components/
/cluster-Witness/components:
Name ID Admin State Operational State Mgmt Connectivity
---------- -- ----------- ------------------- -----------------
cluster-1 1 enabled in-contact ok
cluster-2 2 enabled in-contact ok
server - enabled clusters-in-contact ok
VPlexcli:/cluster-Witness> ll components/*
/cluster-Witness/components/cluster-1:
Name Value
---------------------- ------------------------------------------------------
admin-state enabled
diagnostic INFO: Current state of cluster-1 is in-contact (last
state change: 0 days, 13056 secs ago; last message
from server: 0 days, 0 secs ago.)
id 1
management-connectivity ok
operational-state in-contact
/cluster-witness/components/cluster-2:
Name Value
---------------------- ------------------------------------------------------
admin-state enabled
diagnostic INFO: Current state of cluster-2 is in-contact (last
state change: 0 days, 13056 secs ago; last message
from server: 0 days, 0 secs ago.)
```

id 2
management-connectivity ok
operational-state in-contact
/cluster-Witness/components/server:
Name Value
---------- ----------
admin-state enabled
diagnostic INFO: Current state is clusters-in-contact (last state
change: 0 days, 13056 secs ago.) (last time of
communication with cluster-2: 0 days, 0 secs ago.)
(last time of communication with cluster-1: 0 days, 0
secs ago.)
id -
management-connectivity ok
operational-state clusters-in-contact
Eefer to the VPLEX CLI guide found on Powerlink for more details
around VPLEX Witness CLI.