White Paper

Oracle Extended RAC with EMC® VPLEX™ Metro

Best Practices Planning

Abstract

This white paper describes EMC® VPLEX™ features and functionality relevant to Oracle Real Application Cluster (RAC) and Database. The best practices for configuring an Oracle Extended RAC environment to optimally leverage EMC VPLEX are also presented.

September 2011

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

The information in this publication is provided “as is”. EMC Corporation makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.

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

VMware is a registered trademark of VMware, Inc. All other trademarks used herein are the property of their respective owners.

Part Number h8930

Table of Contents

Executive summary

EMC® VPLEX™ is an enterprise-class storage federation technology that aggregates and manages pools of Fibre Channel (FC) attached storage within and across data centers. VPLEX resides between the servers and FC storage and presents local and distributed volumes to hosts. VPLEX storage aggregation allows online storage migrations and upgrades without any changes to host LUNs. VPLEX AccessAnywhere clustering technology allows read/write access to distributed volumes across distance where the volumes have the exact same SCSI LUN identity. This technology allows hypervisors to migrate virtual machines (VMs) across distance and clusterwares like Oracle Real Application Cluster (RAC) to provide high availability across distance. The VPLEX product line includes VPLEX Local (single site SAN federation), VPLEX Metro supporting synchronous distributed volumes with round-trip latency up to 5 ms, and VPLEX Geo support asynchronous distributed volumes with round-trip time up to 50 ms. This paper highlights VPLEX Metro working in conjunction with Oracle Extended RAC to provide improved application availability, resiliency to failures, and concurrent utilization of hardware resources at both data centers.

Oracle RACs support the transparent deployment of a single database across a cluster of servers, providing fault tolerance, high availability and scalability. Oracle Extended RAC is an architecture where servers in the cluster reside in locations that are physically separate. Oracle Extended RAC provides a way to scale out performance, utilize storage and server resources at multiple sites, and provide an increased resiliency to failure scenarios or maintenance operations without application downtime.

Main benefits of VPLEX Metro and Oracle Extended RAC:

- Continuous database availability through network, server, storage and site failures
- Database sessions automatic failover using Oracle Transparent Application Failover (TAF)
- Scale out architecture and full read-write access to the same database at both sites (no idle hardware at standby site)
- Simplification of deployment by using VPLEX:
    - Hosts need only to connect to local VPLEX cluster
    - No need to deploy Oracle voting disk and clusterware on a third site
    - No host CPU cycles consumed on mirroring. IO sent only once from host
    - Ability to create consistency groups that protect multiple databases or application files as a unit

- ○ VPLEX volumes do not require an application downtime or LUN ID changes during storage hardware refresh and migrations
- ○ Easy Disaster Recovery testing and validation by the fact that both sites can actively participate in the workload

The EMC Symmetrix® VMAX™ was used in this paper as the storage array behind VPLEX clusters. Symmetrix VMAX series provides an extensive offering of industry-leading features and functionality for the next era of high-availability virtual data centers and mission-critical applications. With advanced levels of data protection and replication, the Symmetrix VMAX system is at the forefront of enterprise storage area network (SAN) technology. Additionally, the Symmetrix VMAX array with FAST VP technology transparently optimizes storage tiers, improving performance and saving cost without disruption to host applications.

IT organizations are looking for ways to extend their database environments across data centers to reduce or avoid planned and unplanned downtime associated with hardware failures, disasters or even normal data center operations such as hardware refresh and migrations. With EMC VPLEX, such organizations are able to achieve new levels of flexibility in changes to their storage infrastructure, disaster resilience, and improved collaboration and scale out architecture. Thus EMC VPLEX system is a natural fit for this next generation of environments together with Oracle RAC technology. The capabilities of EMC VPLEX to federate storage systems and present globally unique devices not bound by physical data centers boundaries work hand in hand with the inherent capabilities of Oracle RAC to provide high availability and scalable database access. Oracle RAC together with EMC VPLEX and Symmetrix VMAX is an ideal choice for deploying a highly reliable cloud computing environment, reducing IT costs while also increasing infrastructure efficiency.

Audience

This white paper is intended for Oracle database administrators, storage administrators, and IT architects responsible for architecting, creating, managing, and using IT environments that focus on high availability with Oracle databases, VPLEX technologies, and Symmetrix VMAX storage. The white paper assumes readers are somewhat familiar with Oracle RAC and Oracle database technology, and EMC VPLEX and the Symmetrix storage array.

Introduction

Oracle Extended RAC provides a way to scale out performance, utilize storage and server resources at multiple sites, and provide an increased resiliency to failure scenarios or maintenance operations without application downtime - allowing organizations to eliminate database downtime, and continue business processing uninterrupted, even in the case of complete site failures.

It should be noted that while Oracle Stretched RAC provides high availability across a single database, it is still a good practice to deploy a Disaster Recovery (DR) solution over longer distances with technologies such as RecoverPoint SRDF®, Oracle Data Guard, and so on. Such remote replicas can help in case of database failure (such as erroneous file or LUN deletion, block corruptions, and so on). In addition, it is a good practice to deploy backup strategy to tape or VTL, potentially using clone/snapshot technology to offload the backup process from production.

This white paper introduces readers to the EMC VPLEX family, VPLEX Metro cluster architecture, and features and functionality that are relevant to Oracle Extended RAC deployment. The paper discusses the improved resiliency of the solutions against different failure conditions. It also provides the steps of provisioning VPLEX and Symmetrix VMAX storage for the Oracle Extended RAC platform, and running OLTP workload in a 4-node Oracle Extended RAC with VPLEX Metro and Symmetrix VMAX.

Products and features overview

VPLEX

EMC VPLEX is a storage virtualization solution for both EMC and non-EMC storage, as shown in Figure 1. The storage behind VPLEX can be heterogeneous, supporting both EMC storage and common arrays from other storage vendors, such as NetApp, HDS, HP, and IBM.

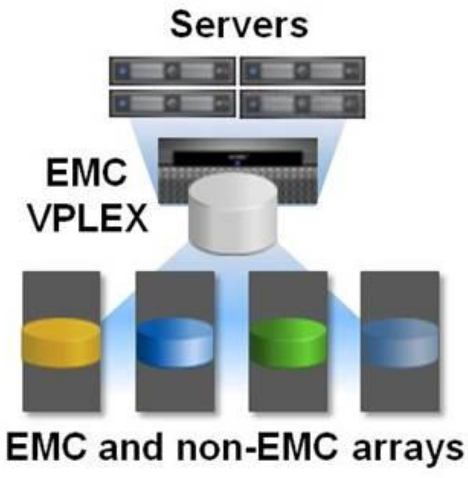


Figure 1. EMC VPLEX Local to federate heterogeneous storage

VPLEX can be extended across geographically dispersed data centers to provide simultaneous access to storage devices through the creation of VPLEX Distributed Virtual Volumes. VPLEX technology provides non-disruptive, heterogeneous data movement and volume management functionality.

Because of these capabilities, VPLEX delivers unique and differentiated value to address three distinct requirements:

- The ability to dynamically move applications and data across different compute and storage infrastructures, be they within or across data centers.
- The ability to create high-availability storage and compute infrastructure geographically dispersed with unmatched resiliency.
- The ability to provide efficient real-time data collaboration over distance.

VPLEX product offerings

EMC offers VPLEX in three configurations to address customer needs for high-availability and data mobility as seen in Figure 2:

- VPLEX Local
- VPLEX Metro
- VPLEX Geo

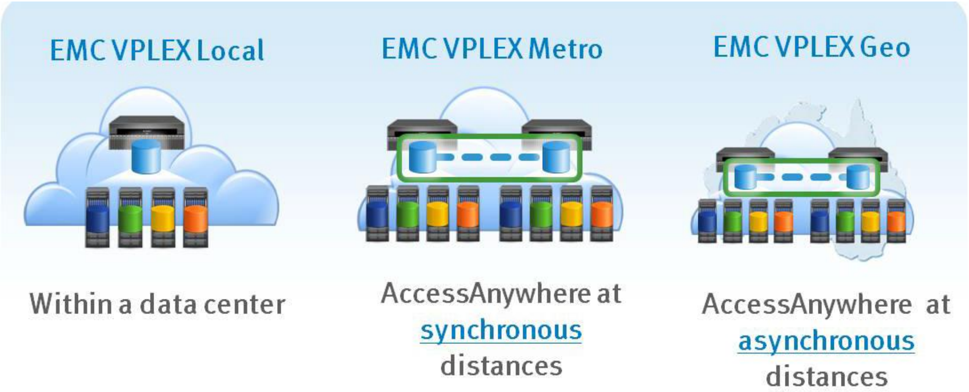

EMC VPLEX Local
EMC VPLEX Metro
EMC VPLEX Geo
Within a data center
AccessAnywhere at synchronous distances
AccessAnywhere at asynchronous distances

Figure 2. VPLEX topologies

VPLEX Local

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

VPLEX Local allows increased availability, simplified management, and improved utilization across multiple arrays.

VPLEX Metro with AccessAnywhere

VPLEX Metro with AccessAnywhere enables active-active, block level access to data between two sites within synchronous distances up to 5 ms round-trip time (RTT).

Following are two examples of using VPLEX Metro and Oracle for data mobility and high-availability.

- **Application and Data Mobility** —By itself, the hypervisor has the ability to move VMs without application downtime between physical servers. When combined with server virtualization, VPLEX distributed volumes allow users to transparently move and relocate VMs and their corresponding applications and data over distance. This provides a *unique capability* allowing users to relocate, share, and balance infrastructure resources between sites. An Oracle VM Live Migration mobility example with VPLEX Metro is covered in the white paper: Oracle VM Server for x86 Live Migration with EMC VPLEX and Symmetrix VMAX, as can be seen in Figure 3.

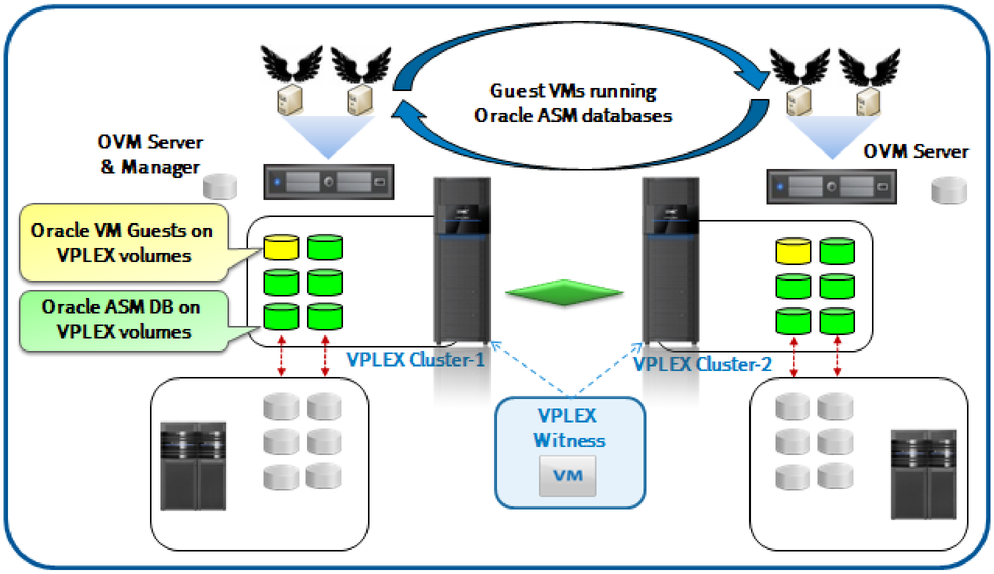


Figure 3. Oracle VM Live Migration with VPLEX Metro

- **High Availability Infrastructure** — Reduces recovery time objective (RTO). High Availability provides near continuous uptime for critical applications, and automates the restart of an application once a failure has occurred, with as little human intervention as possible. An example of Oracle Extended RAC with VPLEX Metro is shown in Figure 4. This solution is the focus of this white paper.

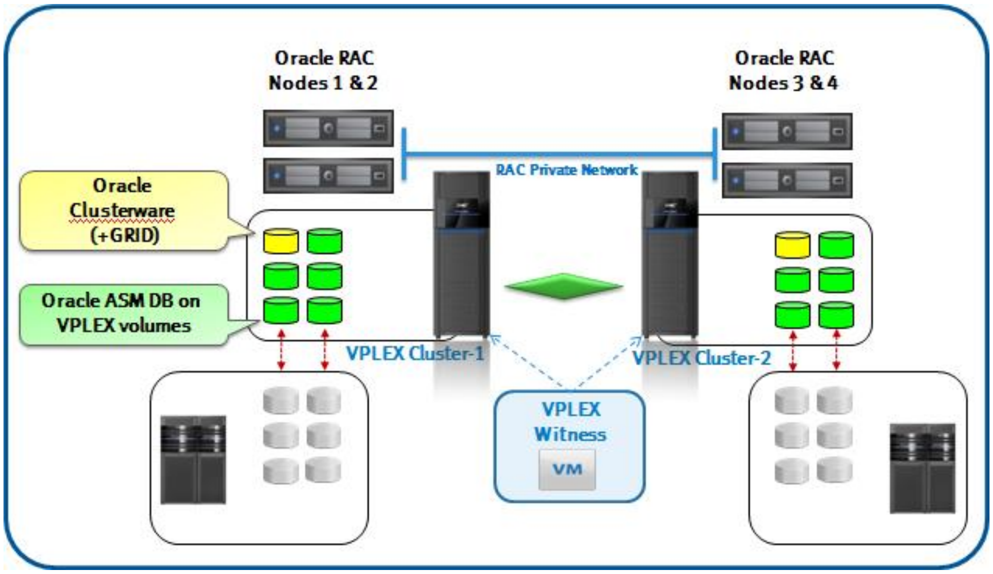


Figure 4. Oracle Extended RAC with VPLEX Metro

VPLEX Geo with AccessAnywhere

VPLEX Geo with AccessAnywhere enables active-active, block level access to data between two sites within asynchronous distances. VPLEX Geo enables better cost-effective use of resources and power. VPLEX Geo provides the same distributed device flexibility as Metro but extends the distance up to and within 50 ms RTT. As with any asynchronous transport media, bandwidth is important to consider for optimal behavior as well as application access method to distributed devices with long distances. An example of VPLEX Geo use case is when application 1 addressing VPLEX distributed volumes in consistency group 1 and performs read/write only at site 1. Application 2 addressing VPLEX distributed volumes in consistency group 2 and performs read/write only at site 2. VPLEX Geo maintains write IO fidelity for each application at the remote site, although slightly behind in time. In this example both applications can fail over and back independently, over long distances.

VPLEX architecture highlights[1]

The VPLEX family uses a unique clustering architecture to help customers break the boundaries of the data center and allow servers at multiple data centers to have read/write access to shared block storage devices. VPLEX Local includes a single cluster while VPLEX Metro and Geo each include two. A VPLEX cluster consists of one, two, or four engines as seen in Table 1. Each VPLEX engine provides SAN/WAN connectivity, cache and processing power with two redundant directors.

Table 1. VPLEX hardware components

Feature	Description
VPLEX Cluster	Contains one, two, or four engines
VPLEX Engine	Contains two directors, management modules, power supplies, battery power and fans
VPLEX Director	Contains some of the I/O modules, SSD, CPU and RAM

VPLEX Local uses write-through caching and lets writes pass directly and get acknowledged first at the storage behind the VPLEX volumes before acknowledging them back to the host. With EMC storage such as Symmetrix and VNX™, where writes only need to register with the storage persistent cache, application write response time is optimal. VPLEX Metro also uses write-through cache but will acknowledge the writes to the application only once they were registered with both local and remote storage. On the other hand, writes to VPLEX Geo are cached at the VPLEX layer and sent to the remote cluster in deltas that preserve write-order fidelity. In all three VPLEX deployments – Local, Metro, and Geo, reads can benefit from the VPLEX cache, and in VPLEX Metro and Geo, read-hits are served from the local VPLEX cluster cache.

[1] The details in this section are based on VPLEX release 5.0.1 and may be different in other releases. The VPLEX product guide provides exact version details.

VPLEX logical storage structure

VPLEX encapsulates traditional physical storage array devices and applies three layers of logical abstraction to the storage volumes as seen in Figure 5. Extents are the mechanism VPLEX uses to divide storage volumes. Extents may be all or part of the underlying storage volume. EMC VPLEX aggregates extents and it can apply RAID protection in the device layer. Devices are constructed using one or more extents and can be combined into more complex RAID schemes and device structures as desired. At the top layer of the VPLEX storage structures are virtual volumes. Virtual volumes are created from devices and inherit the size of the underlying device. Virtual volumes are the elements VPLEX exposes to hosts using its front-end (FE) ports. Access to virtual volumes is controlled using storage views, which are comparable to Auto-provisioning Groups on EMC Symmetrix or to storage groups on EMC CLARiiON®. They act as logical containers determining host initiator access to VPLEX FE ports and virtual volumes.

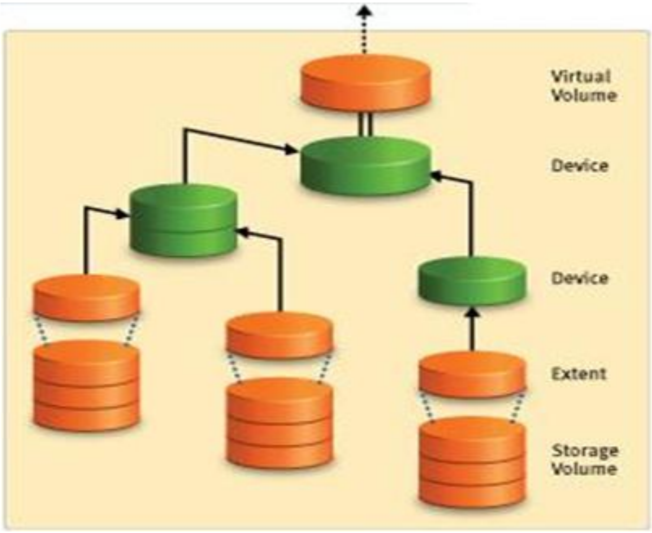


Figure 5. EMC VPLEX logical storage structures

VPLEX encapsulates storage area network (SAN) volumes by identifying storage devices with their WWNs, packaging them into sets of VPLEX virtual volumes with user-defined configuration and protection levels, and then presenting the virtual volumes to hosts. VPLEX has the ability to encapsulate/de-encapsulate existing storage devices (preserving their data), as well as the ability to divide and aggregate existing storage volumes for use in virtual volumes. Furthermore, virtual storage presented to hosts through VPLEX encapsulation can be non-disruptively moved within and between back-end storage arrays. It is recommended that when VPLEX encapsulates devices of EMC storage arrays such as VMAX and VNX the RAID protection is done at the storage array. This will keep the mapping between storage and VPLEX simple and will allow the use of storage features such as creating snaps, clones, and additional DR.

VPLEX consistency groups, detach rules and Witness

Consistency groups

Starting with GeoSynchrony 5.0 for VPLEX Metro and GeoSynchrony 5.0.1 for VPLEX Geo, consistency groups are used to organize virtual volumes. Consistency groups aggregate volumes together to provide a common set of properties to the entire group. In addition, you can move a consistency group from one cluster to another if required. Consistency groups are particularly important for databases and applications. All database LUNs (for example, Oracle data, control and log files) require preserving write-order fidelity to maintain data integrity, and therefore should always be placed in a single consistency group together. Often multiple databases have transaction dependencies, such as when database links are used to connect databases, or when the application issues transactions to multiple databases and expect them to be consistent with each other. In these cases the consistency group should include all LUNs that require preserving IO dependency (write-order fidelity).

There are two types of consistency groups:

Synchronous consistency groups — Provide a convenient way to apply the same detach rules 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. Synchronous consistency groups use write-through caching (known as synchronous cache mode in the VPLEX user interface) and with VPLEX Metro are supported on clusters separated by up to 5 ms of latency. This means that VPLEX Metro sends writes to the back-end storage volumes, and acknowledges the write to the application as soon as the back-end storage volumes in both clusters acknowledge the write.

Asynchronous consistency groups — Used for distributed volumes in a VPLEX Geo, separated by up to 50 ms of latency. All volumes in an asynchronous consistency group share the same detach rules and cache mode, and behave the same way in the event of an inter-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 VPLEX 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 compared and reconciled 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 (to preserve write order fidelity). Asynchronous cache mode can give better performance, but there is a higher risk that some data will be lost during a VPLEX cluster failure or inter-cluster link failure due to the Async nature of the replication, or if writes are performed at both local and remote locations.

Detach rules

Detach rules are predefined rules that determine consistency group I/O processing semantics whenever connectivity with remote cluster is lost (for example network partitioning or remote cluster failure). In these situations, until communication is restored, most workloads require specific sets of virtual volumes to resume I/Os on one cluster and suspend I/Os on the other.

In a VPLEX Metro configuration detach rules can depict a static preferred cluster by setting[2]: *winner:cluster-1*, *winner:cluster-2* or *No Automatic Winner* (where the last one specifies no preferred cluster). When the system is deployed without VPLEX Witness (discussed in the next section), I/Os to consistency group devices proceeds on the preferred cluster and are suspended on the non-preferred cluster.

Asynchronous consistency groups detach rules can be *Active Cluster Wins* or *No Automatic Winner*. When using the active cluster wins rule, I/Os to the consistency group continue at the cluster where the application was actively writing last (provided there was only one such cluster. If dirty buffers exist in both clusters for the same asynchronous consistency group, I/Os to the consistency group will suspend on both clusters). In this way, the active cluster is the preferred cluster.

VPLEX Witness

VPLEX Witness, introduced with GeoSynchrony 5.0, is an optional external server that is installed as a virtual machine. 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 cluster(s) to distinguish between inter-cluster network partition failures and cluster failures and automatically resume I/O in these situations at the appropriate site. With GeoSynchrony 5.0 VPLEX Witness affects only synchronous consistency groups in a VPLEX Metro configuration, and only when the detach rules depict either cluster-1 or cluster-2 as preferred for the consistency group (that is, VPLEX Witness does not affect consistency groups where No Automatic Winner rule is in effect).

Without VPLEX Witness, if two VPLEX clusters lose contact, the consistency group detach rules in effect define which cluster continues operation and which suspends I/O as explained earlier. The use of detach rules alone to control which site is a winner may add an unnecessary complexity in case of a site failure since it may be necessary to manually intervene to resume I/O to the surviving site. VPLEX Witness handles such an event dynamically and automatically. It provides the following features:

- Automatic load balancing between data centers
- Active/active use of both data centers
- Fully automatic failure handling

[2] Based on management GUI options. CLI uses slightly different terms to specify the same rules.

In order for VPLEX Witness to be able to distinguish correctly between failure conditions, it must be installed in an independent failure domain from either cluster using distinct network interfaces to each. This will eliminate the possibility of a single fault affecting both the cluster and the VPLEX Witness. 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.

Cache vault

To avoid metadata loss or data loss (when write back caching is used) under emergency conditions, VPLEX uses a mechanism called cache vaulting to safeguard cache information to a persistent local storage.

Oracle Real Application Clusters

Oracle Real Application Clusters) are an optional feature of Oracle Database Enterprise Edition. Oracle RAC supports the transparent deployment of a single database across a cluster of servers, providing fault tolerance from hardware failures or planned outages. If a node in the cluster fails, the Oracle database continues running on the remaining nodes. If more processing power is needed, new nodes can be added to the cluster by providing horizontal scaling.

Oracle RAC supports mainstream business applications of all kinds. This includes Online Transaction Processing (OLTP) and Decision Support System (DSS).

Oracle Automatic Storage Management

Oracle Automatic Storage Management (ASM) is an integrated database filesystem and disk manager. ASM filesystem and volume management capabilities are integrated with the Oracle database kernel. This reduces the complexity of managing the storage for the database. In addition to providing performance and reliability benefits, ASM can also increase database availability because disks can be added or removed without shutting down the database. ASM automatically rebalances the files across the disk group after disks have been added or removed.

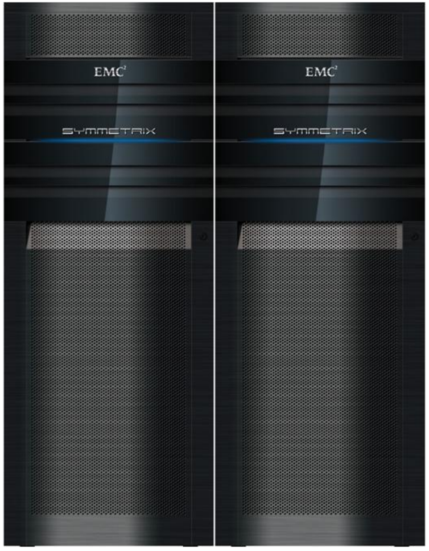
In Oracle Database 11gR2, Oracle ASM and Oracle Clusterware have been integrated into a single set of business and named Oracle Grid Infrastructure. This now provides all the cluster and storage services required to run an Oracle RAC database. Oracle ASM has also been extended to include support for Oracle Cluster Registry (OCR) and voting files to be placed within ASM disk groups.

Oracle RAC on Extended Distance Clusters

Oracle RAC on Extended Distance Clusters is an architecture where servers in the cluster reside in locations that are physically separate. Oracle RAC on Extended Distance Clusters provides greater availability than local Oracle RAC. Oracle RAC on Extended Distance Clusters provides extremely fast recovery from a site failure and allows for all servers, in all sites, to actively process transactions as part of a single database cluster. While this architecture creates great interest and has been successfully implemented, it is critical to understand where this architecture best fits especially in regards to distance, latency, and degree of protection it provides. The high impact of latency, and therefore distance, creates some practical limitations as to where this architecture can be deployed. This architecture fits best where the two datacenters are located relatively close and where the costs of setting up direct dedicated channels between the sites have already been taken.

Symmetrix VMAX

Symmetrix VMAX is built on the strategy of simple, intelligent, modular storage, and incorporates a new Virtual Matrix™ interconnect that connects and shares resources across all nodes, allowing the storage array to seamlessly grow from an entry-level configuration into the world's largest storage system. It provides the highest levels of performance and availability featuring new hardware capabilities as seen in Figure 6.



- 2 – 16 director boards
- Up to 2.1 PB usable capacity
- Up to 128 FC FE ports
- Up to 64 FICON FE ports
- Up to 64 GigE / iSCSI FE ports
- Up to 1 TB global memory (512 GB usable)

Figure 6. The Symmetrix VMAX platform

Symmetrix VMAX provides the ultimate scale-out platform. It includes the ability to incrementally scale front-end and back-end performance by adding processing modules (nodes) and storage bays. Each processing module provides additional front-end, memory, and back-end connectivity.

Symmetrix VMAX also increases the maximum hyper size to 240 GB (64 GB on Symmetrix DMX™). This allows ease of storage planning and device allocation, especially when using Virtual Provisioning™ where the thin storage pool is already striped and large hypers can be easily used.

Symmetrix VMAX TimeFinder product family

The EMC TimeFinder® family of local replication technology allows for creating multiple, nondisruptive, read/writeable storage-based replicas of database and application data. It satisfies a broad range of customers' data replication needs with speed, scalability, efficient storage utilization, and minimal to no impact on the applications – regardless of the database size. TimeFinder provides a solution for backup, restart, and recovery of production databases and applications, even when they span Symmetrix arrays. TimeFinder is well integrated with other EMC products such as SRDF and allows the creation of replicas on a remote target without interrupting the synchronous or asynchronous replication. If a restore from a remote replica is needed, TimeFinder and SRDF will restore data incrementally and in parallel, to achieve a maximum level of availability and protection. The TimeFinder product family supports the creation of dependent write-consistent replicas by using EMC consistency technology, and replicas that are valid for Oracle backup/recovery operations, as described later in the use cases.

Virtual Provisioning

Symmetrix thin devices are logical devices that you can use in many of the same ways that Symmetrix devices have traditionally been used. Unlike traditional Symmetrix devices, thin devices do not need to have physical storage preallocated at the time the device is created and presented to a host (although in some cases customers interested only in the thin pool wide striping and ease of management choose to fully preallocate the thin devices). You cannot use a thin device until it has been bound to a thin pool. Multiple thin devices may be bound to any given thin pool. The thin pool is comprised of devices called data devices that provide the actual physical storage to support the thin device allocations. Table 2 describes the basic Virtual Provisioning definitions.

Table 2. Definitions of Virtual Provisioning devices

Device	Description
Thin device	A host-accessible device that has no storage directly associated with it.
Data devices	Internal devices that when placed in a thin pool provide storage capacity to be used by thin devices.
Thin pool	A collection of data devices that provide storage capacity for thin devices.

Implementing VPLEX Metro with Oracle Extended RAC

Oracle Extended RAC deployment considerations in a VPLEX Metro environment

EMC VPLEX breaks the physical barriers of data centers and allows users to access data at different geographical locations concurrently. Oracle Extended RAC with VPLEX Metro allows for transparent workload sharing between multiple sites accessing a single database, while providing the flexibility of migrating workloads between sites in anticipation of planned events such as hardware maintenance. Furthermore, in case of an unplanned event that causes disruption of services at one of the data centers, the failed client connections can be automatically restarted at the surviving site.

Oracle Extended RAC and VPLEX Metro deployment guidelines

The following points describe some of the main deployment guidelines.

Oracle Clusterware and VPLEX Witness deployment

Often deployments of Oracle Extended RAC emphasize deploying a third site for one of the Oracle clusterware voting disks (possible based on NFS). However in the deployment of Oracle Extended RAC with VPLEX Metro, it is the VPLEX Witness alone that is deployed in an additional fault domain (third site in the case of multi-site deployment) as described in the VPLEX Witness section earlier:

- In the case of Oracle interconnect partitioning alone (not a true site failure, and no affect on VPLEX interconnect) then Oracle clusterware will reconfigure based on node majority and access to the voting disk. No need for a third site for Oracle clusterware.
- In the case of VPLEX interconnect partitioning (or a true site failure) VPLEX immediately suspends IOs at both clusters while VPLEX Witness resolves which of them should continue to service IOs (based on detach rules and site availability). The Oracle cluster nodes will therefore have access to voting disks only where VPLEX resumes IO, and Oracle clusterware will therefore reconfigure the cluster in accordance. Again, no need for Oracle clusterware to use a third site.

Oracle Clusterware deployment

Oracle Cluster Ready Service (CRS) is deployed on VPLEX distributed volumes alone (no third sites) as described in the previous section. In Oracle database, release 11gR2 CRS was merged with ASM to create the grid infrastructure. Therefore, the first ASM disk group is created at the time of CRS installation. In prior database releases, the Oracle clusterware uses either raw devices or a clustered filesystem:

- When ASM is used for Oracle clusterware (starting with Oracle 11gR2) it is recommended to create a unique disk group for CRS alone, for example: +GRID (that is, no database content such as log or data files will be placed there). The +GRID disk group will benefit from using Normal or High redundancy. This way Oracle will create multiple voting disks (instead of only one such as the case if CRS ASM disk group was using External redundancy[3]). Because no database content is included, the size of the distributed VPLEX devices that are used for this disk group can be relatively very small.
- The recommended approach when EMC storage is used behind VPLEX, such as VMAX or VNX, or when VPLEX RAID protection is used, is to set all other ASM disk groups with external redundancy. This will provide adequate protection for the ASM members, based on VPLEX or EMC storage array RAID.

[3] Starting with Oracle 11gR2, the number of CRS voting disks is determined automatically by the ASM redundancy level, such as External redundancy implies 1 voting disk, Normal redundancy implies 3 voting disks, and High redundancy implies 5 voting disks.

- Since all Oracle cluster nodes require access to all CRS and database devices, both CRS and Oracle database should use only the VPLEX distributed volumes, whether ASM is used, raw devices, or a clustered filesystem.
- To align with timeout values associated with VPLEX cluster and Witness resolution of failure states, it is recommended to modify the CRS CSS Miscount from the default of 30 seconds to 45 seconds. This can be done by using the following commands:

```
# crsctl get css miscount

# crsctl set css miscount 45
```

- No other timeout value changes are necessary (that is, such as crs disk timeout).

Additional notes

On x86-based server platform, ensure that partitions are aligned. VPLEX requires alignment at 4 KB offset; however, if Symmetrix is used align at 64 KB (128 blocks) offset (which is natively aligned at 4 KB boundary as well):

- On Windows diskpar or diskpart can be used. On Linux fdisk or parted can be used.
- An example of aligning partition to 64 KB offset using fdisk is shown later in the section: Create partitions on PowerPath devices.

Oracle Extended RAC and VPLEX Metro protection from unplanned downtime

The combination of Oracle Extended RAC and VPLEX Metro provides improved availability and resiliency to many failure conditions and therefore increased the availability of mission-critical database and applications.

Table 3 summarizes the list of failure scenarios and the best practices that will make the database able to continue operations in each of them. Note that the list does not cover failover to a standby system (such as Oracle Data Guard, RecoverPoint, SRDF, and so on). *EMC VPLEX with GeoSynchrony 5.0 Product Guide* provides more information regarding VPLEX connectivity best practices.

Table 3. Summary of VPLEX Metro and Oracle Extended RAC improved resiliency

Resiliency to host and site failure scenarios			
Failure	**Oracle database single server (non-RAC)**	**Oracle RAC (not extended)**	**Oracle Extended RAC with VPLEX Metro**
Host HBA port failure	✓ • Each host should have more than a single path to the storage. Use multiple HBA ports (initiators). • Use multipathing software (such as EMC PowerPath®) for automatic path failover and load balancing. • Ideal SAN connectivity will use redundant switches where the HBA ports (initiators) will spread across them.	✓ Same as single server	✓ Same as single server
Host hardware failure or crash	✗ • Downtime implied until host and application can resume operations.	✓ • Oracle RAC provides database resiliency for a single server failure by performing automatic instance recovery and having other cluster nodes ready for user connections. • Configure Oracle Transparent Application Failover to allow sessions to automatically failover to a surviving cluster node.	✓ Same as Oracle RAC
Lab/building/site failure	✗ • Downtime implied until host and application can resume operations.	✗ • Downtime implied until host and application can resume operations.	✓ • By installing VPLEX clusters and Witness in independent failure domains (such as another building or site) it becomes resilient to a lab, building, or site failures. • The VPLEX cluster in the failure-domain not affected by the disaster will continue to serve I/Os to the application.

			• Use Oracle Transparent Application Failover to allow automatic user connection failover to the surviving cluster nodes.
Resiliency to database/network related failure scenarios			
Failure	**Oracle database single server (non-RAC)**	**Oracle RAC (not extended)**	**Oracle Extended RAC with VPLEX Metro**
Database instance crash or public network disconnect	✗ • Downtime implied until instance can be brought up again or public network reconnected.	✓ • Oracle RAC provides database resiliency for a single server failure by performing automatic instance recovery and having other cluster nodes ready for user connections. • Configure Oracle Transparent Application Failover to allow sessions to automatically failover to a surviving cluster node.	✓ Same as Oracle RAC
RAC Cluster interconnect partitioning	• N/A	✓ • RAC natively and automatically handles by cluster reconfiguration. Cluster nodes will compete for voting disk access.	✓ Same as Oracle RAC
Resiliency to Storage failure scenarios			
Failure	**Oracle database single server (non-RAC)**	**Oracle RAC (extended)**	**Oracle Extended RAC with VPLEX Metro**
Front-end port failure	✓ • SAN connectivity should include multiple storage front end ports, ideally across Symmetrix directors. If using Symmetrix with multiple engines connect to ports on different engines as well to gain even higher protection..	✓ Same as single server	✓ Same as single server
Physical drive failure	✓ • Use storage RAID protection. Symmetrix storage uses RAID protection where RAID1 and RAID5 protect from a single disk failure within a RAID	✓ Same as single server	✓ Same as single server

	group and RAID6 protects from two disks failures within a RAID group. In either case the application will continue to run undisturbed. • If a drive starts failing Symmetrix hot-spare drive will copy its data and EMC Enginuity™ will initiate Call-Home to inform EMC support immediately.		
Storage array components, including director board (cache, IO)	✓ • Symmetrix components are fully redundant, including mirrored cache that is also persistent (uses vaulting in the case of elongated power failure), redundant directors and power supplies. • Symmetrix data is T10 DIF protected from the time it enters the storage until it leaves it.	✓ Same as single server	✓ Same as single server
Loss of connectivity to storage array	✗ • Downtime implied until storage array connectivity can be resumed.	✗ • Downtime implied until storage array connectivity can be resumed.	✓ • VPLEX Metro synchronous consistency group continues to serve I/O's at both sites, even if one of the storage arrays is not available. • Oracle clusterware would not know about the storage unavailability as VPLEX cluster continues to service all I/Os.
Resiliency to VPLEX failure scenarios			
Failure	**Oracle Database Single Server (non-RAC)**	**Oracle RAC (not extended)**	**Oracle Extended RAC with VPLEX Metro**
Front-end port	✓ • SAN connectivity should include multiple storage front-end ports, ideally across Symmetrix directors. If using VPLEX with multiple engines connect to ports on different engines as well to gain even higher protection.	✓ Same as single server	✓ Same as single server

	• Use multipathing software (such as PowerPath) for automatic path failover and load balancing. • Ideal SAN connectivity will use redundant switches with connectivity to multiple VPLEX front-end ports. ✓	✓	✓
Back-end port	• In a similar fashion to VPLEX front-end ports, also use redundant switch connectivity with VPLEX back-end ports as they connect to the storage array. ✓	Same as single server ✓	Same as single server ✓
VPLEX hardware component	• VPLEX components are fully redundant, including persistent cache (uses vaulting in the case of elongated power failure), redundant directors and power supplies.	Same as single server	Same as single server ✓
VPLEX Interconnect partition	N/A ✗	N/A ✗	If both sites are still available the VPLEX Preferred Cluster detach rules will determine which cluster resumes I/Os and which suspends, without downtime for host connected to the surviving Cluster. ✓
VPLEX cluster unavailable	• Downtime implied until VPLEX cluster connectivity can be resumed.	• Downtime implied until VPLEX cluster connectivity can be resumed.	• VPLEX Witness will allow I/O to resume at the surviving VPLEX cluster. RAC cluster nodes connected to that VPLEX cluster will resume operations without downtime. • Use Oracle Transparent Application Failover (TAF) to allow automatic user connection failover to the surviving cluster nodes.

Although the architecture of VPLEX is designed to support concurrent access at multiple locations, the current version of the product supports a two-site configuration separated by synchronous distance with a maximum round trip latency of 5 ms between the two sites. In addition, the VPLEX Metro solution requires extension of a VLAN to different physical data centers. Technologies such as Cisco's Overlay Transport Virtualization (OTV) can be leveraged to provide the service. *EMC VPLEX Architecture and Deployment: Enabling the Journey to the Private Cloud* TechBook, available on EMC Powerlink®, provides further information on EMC VPLEX Metro configuration.

VPLEX Metro with Oracle Extended RAC lab configurations and tests

Lab configuration and setup

The following section describes the technologies and components used in the test cases documented in this paper.

Physical environment

Figure 7 illustrates the overall physical architecture of the Oracle Extended RAC configuration deployment that was used for the tests shown in this paper. The Oracle Extended RAC consisted of four Oracle RAC cluster nodes, two at each simulated data center (Site A and Site B).

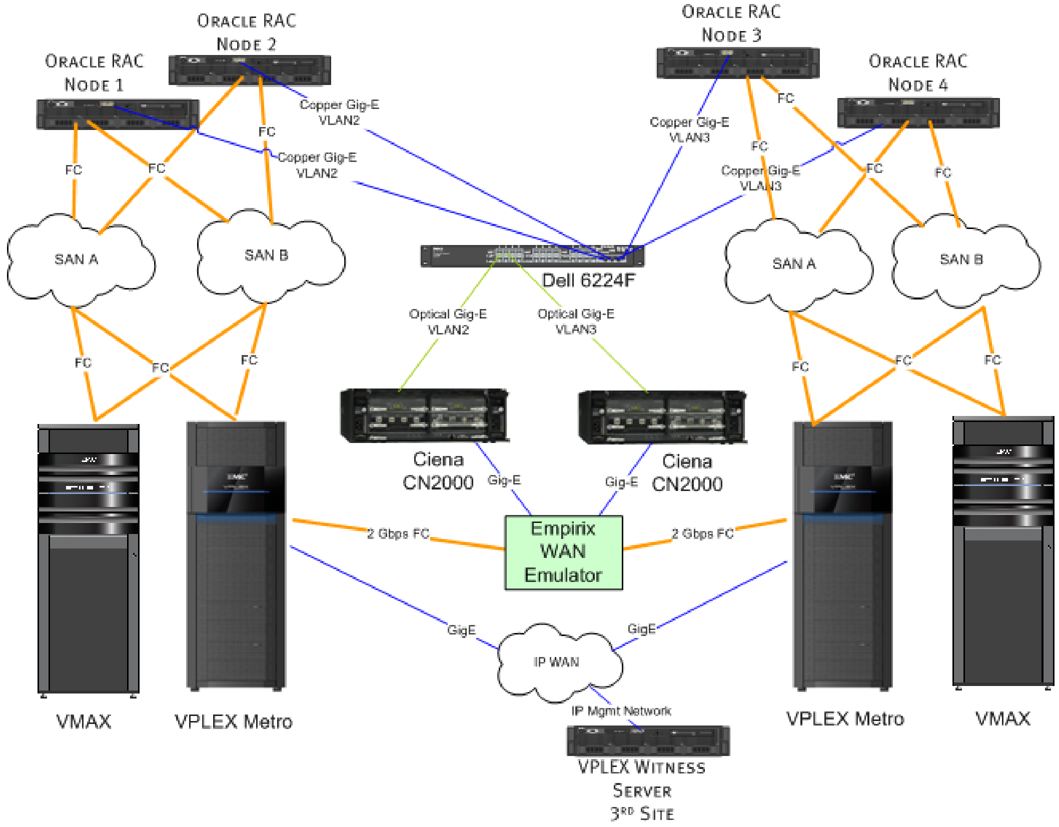


Figure 7. VPLEX Metro configuration for Oracle

The hardware setup included setting up Fibre Channel connectivity among the hosts, VPLEX clusters, and Symmetrix VMAX storage arrays to the redundant switches. In each site, the hosts are zoned to their local VPLEX cluster front-end ports, and the storage to the VPLEX cluster back-end ports. The VPLEX clusters are zoned to each other (through the distance simulator in this case). The Oracle RAC interconnect between sites used the same distance simulator to experience the same latency overhead as the VPLEX interconnect. The distance simulator used 2 x 1 GigE links with WAN compression and Ciena switches for protocol change.

Table 4 shows the VPLEX, hosts, and Symmetrix storage array hardware details. They included 2 x 2-engine VPLEX clusters setup in Metro configuration, and 2 x 1-engine Symmetrix VMAX arrays providing local storage to each VPLEX cluster. The 4 x RAC nodes consisted of Dell 2950 dual core servers with 16 GB of cache. It should be noted that the focus of the tests was on generating high IO workload rather than making use of lots of RAM to achieve high transaction rate numbers. VPLEX Witness was set and connected to each VPLEX cluster.

Table 5 shows the distance simulation hardware. It included Dell PowerConnect 6224F Fiber Ethernet Switch that converts Gigabit IP network into Fibre Channel network. Ciena CN 2000 network manager provides virtual optical networking with inputs of Fibre Channel, FICON, or GbE. It supports both data compression and Dynamic Bandwidth Allocation (DBA) to provide port level quality of service to client connections. Empirix PacketSphere Network Emulator provided a controlled network emulation for IP network impairments, such as network latency, for full-duplex, Gigabit wire speed network traffic.

Table 4. Extended RAC hardware environment

Hardware	Quantity	Release and configuration
EMC VPLEX Metro	2	VPLEX Metro with GeoSynchrony 5.0.1 Two engine and four directors on each cluster
VPLEX Witness	1	Dell R900 running VPLEX Witness virtual machine (VM)
Symmetrix VMAX	2	Single engine VMAX with Enginuity 5875, 112 x 450 GB/15k FC drives using Virtual Provisioning
Dell 2950 server (RAC nodes)	4	2 x dual core, 16 GB RAM
Emulex Express (2 HBA ports used per server)	4	

The main guidelines during the connectivity step are to maximize hardware redundancy such as using two switches, more than a single HBA, multipath for dynamic path failover, and load balancing.

Table 5. Distance simulation hardware environment

Hardware	Quantity	Release and configuration
Dell PowerConnect 6224F Fiber Ethernet Switch	1	Support up to 4 x 10 Gigabit fiber and 2 x 10GBase-T copper Ethernet uplinks
Ciena CN 2000 Network Manager	1	Ciena ON-Center CN 2000 Network Manager 5.0.1
Empirix PacketSphere Network Emulator	1	Empirix Network Emulator NE XG 1.0B8, 2 1-GbE network emulators
Dell 2950 server	4	2 x dual core, 16 GB RAM

Table 6 describes the host software used.

Table 6. Host software

Software	Release
Server OS	Oracle Enterprise Linux Release 5 Update 4 x86_64
EMC PowerPath	Version 5.5 for Linux x86_64
Oracle	Oracle Clusterware 11g R2 (11.2.02) and Oracle Database 11g R2 (11.2.0.2) for Linux x86-64

Storage setup and device allocation planning

Symmetrix VMAX Virtual Provisioning and storage device configuration

Table 7 describes Symmetrix VMAX device configuration with storage Virtual Provisioning and volume layout for VPLEX Metro with Oracle Extended RAC testing environment.

This configuration placed the Oracle data and log files in separate thin pools which allowed each to use distinct RAID protection. In this configuration, data files were placed in the RAID5 protected thin pool and redo logs in the RAID1 protected thin pool:

- Symmetrix RAID5 protected thin pool offers a good combination of protection, performance, and capacity utilization benefits for the data files with optimized writes and rotating parity, therefore being a good choice for data files. RAID1 protected thin pool in certain cases may offer a slight improvement in availability and performance over RAID5 and therefore was used for the log files. Note that the same group of physical disks was shared by both thin pools to allow full sharing of physical resources.
- It should be noted that a different configuration, focused on simplicity rather than pure storage performance/availability optimization, could be deployed where both data and log files share the same thin pool and RAID protection (be it RAID1, RAID5, or RAID6).

Multiple ASM disk groups were used in the configuration:

- +GRID: As discussed earlier in Oracle Clusterware deployment section, when ASM is used for Oracle clusterware (starting with Oracle 11gR2) it is recommended to create a unique disk group for CRS alone, for example: +GRID. This approach of separating the clusterware LUNs from the database is useful when there are plans to use storage technologies such as clones or snapshots to create additional database copies for repurposing, backup, and so on. These copies will not include the clusterware LUNs. It is also helpful for DR solutions such as RecoverPoint and SRDF since the replicated ASM disk groups exclude the clusterware LUNs, and will be mounted to an already configured CRS stack at the DR target site.
- Separating +DATA, +LOG, and +FRA ASM disk groups allow the storage technology to be used for offloading backups from production. During hot backup process the +DATA and +FRA diskgroups will be cloned at different times. In addition, remote replications such as RecoverPoint and SRDF natively create a database restartable replica. Restartable replicas do not access archive logs during crash or instance recovery and therefore the archive logs (+FRA disk group) do not need to be part of the replication.
- Temp files often can be included in the +DATA disk group. The only reason they used a separate ASM disk group in the tests was just for monitoring purposes and it is not a specific deployment recommendation.

Table 7. Storage and database devices configuration and assignment

	Thin devices (LUNs)			Data devices
	ASM disk group and LUNs assignment	Thin pool binding	Thin devices	
Oracle RAC Grid/ASM **Instance**	+Grid ASM disk group	Redo_Pool	5 x 20 GB thin LUNs (15F:163)	56 x 30GB RAID1
Database: Name: ERPFINDB Size: 1 TB Num. LUNs: 38	+REDO ASM disk group	Redo_Pool	5 x 20 GB thin LUNs (164:168)	
	+DATA: ASM disk group	Data_Pool	25 x 60 GB thin LUNs (1A5:1B4)	56 x 230 GB RAID5 (3+1) (C5:FC)
	+TEMP: ASM disk group	Temp_Pool	6 x 50 GB thin LUNs (17D:182)	56 x 60 GB RAID5 (3+1) (8D:C4)
	+FRA: ASM disk group		2 x 50 GB thin LUNs	
VPLEX	VPLEX Meta devices		2 x 2 x 80 GB thin LUNs (2E5:2E8)	
	VPLEX Log		2 x 50 GB thin LUNs (2E9:2EA)	

Symmetrix VMAX storage provisioning for VPLEX Metro system

The following steps were followed to provision storage from the Symmetrix VMAX system to a VPLEX virtual storage environment, which is basically the same as to provision storage to physical or virtual servers. The process described assumes this is the first time storage is being provisioned from the Symmetrix VMAX array to the VPLEX Metro system. It also assumes that the VPLEX have been zoned to front-end ports of the Symmetrix VMAX storage array. The Symmetrix actions should be executed from a management host connected through gatekeeper to the Symmetrix, or using a Symmetrix Management Console client. The following list shows the setup activities with focus on CLI although they can be easily executed using Symmetrix Management Console as well.

Step	Action
1	Create Symmetrix devices on both local and remote arrays using either Symmetrix Management Console or the Solutions Enabler command line interface.

2	Create a Symmetrix storage group by executing the following command:
	`symaccess -sid <symm_id> -name <group name> -type storage devs create`
	For example, the command:
	`symaccess -sid 191 -name VPLEX1_Storage_Group -type storage devs 15F:168 create`
	creates the storage group, Storage_Group_Test.

3	Create a Symmetrix port group. The command:
	`symaccess -sid 191 -name VPLEX1_Port_Group -type port -dirport <Dir>:<Port> create`
	For example, the command:
	`symaccess -sid 191 -name VPLEX1_Port_Group -type port -dirport 7E:1 create`
	creates the port group, Port_Group_Test.

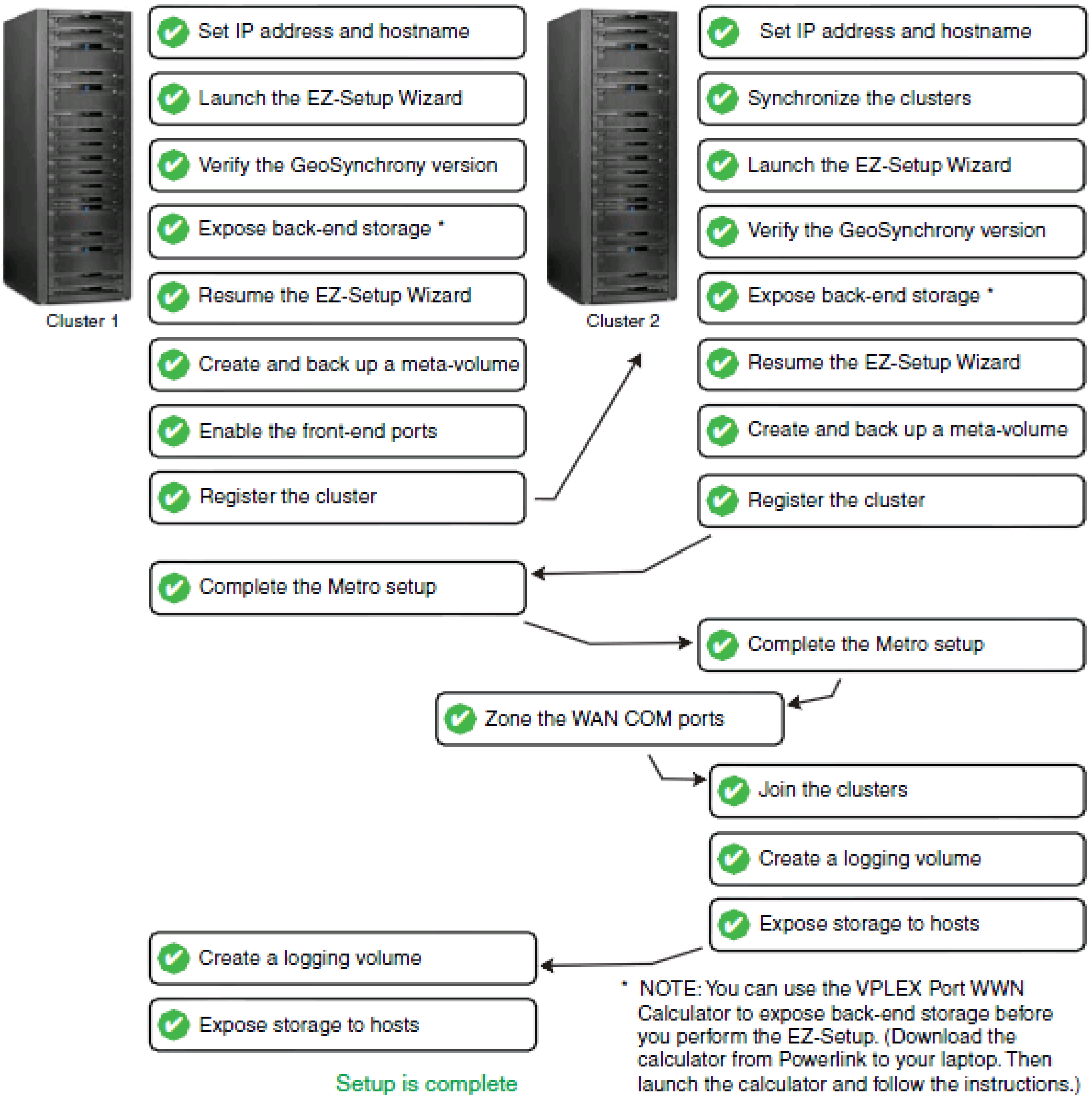
4	Create a Symmetrix initiator group where the VPLEX back-end ports' WWNs are the "host" initiators to the Symmetrix initiator group. The creation of the initiator group, Initiator_Group_Test, can be achieved by running the following command:
	`symaccess -sid 191 -name VPLEX1_Initiator_Group -type init -wwn <WWN> create`
	For example, the command:
	`symaccess -sid 191 -name VPLEX1_Initiator_Group -type init -wwn 500014426011ee10 create`

5	Create the Symmetrix masking view to group the storage, port, and initiator groups: `symaccess -sid 191 create view -name VPLEX1_View -storgrp VPLEX1_Storage_Group -portgrp VPLEX1_Port_Group -initgrp VPLEX1_Initiator_Group`
6	Repeat steps 1–5 for storage provisioning the second VPLEX system (VPLEX2) to the second VMAX array (sid 219)

VPLEX Metro setup

VPLEX Metro cluster setup steps

Figure 8 lists the main tasks that are required for VPLEX Metro setup.


Cluster 1
Set IP address and hostname
Launch the EZ-Setup Wizard
Verify the GeoSynchrony version
Expose back-end storage *
Resume the EZ-Setup Wizard
Create and back up a meta-volume
Enable the front-end ports
Register the cluster
Cluster 2
Set IP address and hostname
Synchronize the clusters
Launch the EZ-Setup Wizard
Verify the GeoSynchrony version
Expose back-end storage *
Resume the EZ-Setup Wizard
Create and back up a meta-volume
Register the cluster
Complete the Metro setup
Complete the Metro setup
Zone the WAN COM ports
Join the clusters
Create a logging volume
Expose storage to hosts
Create a logging volume
Expose storage to hosts
Setup is complete

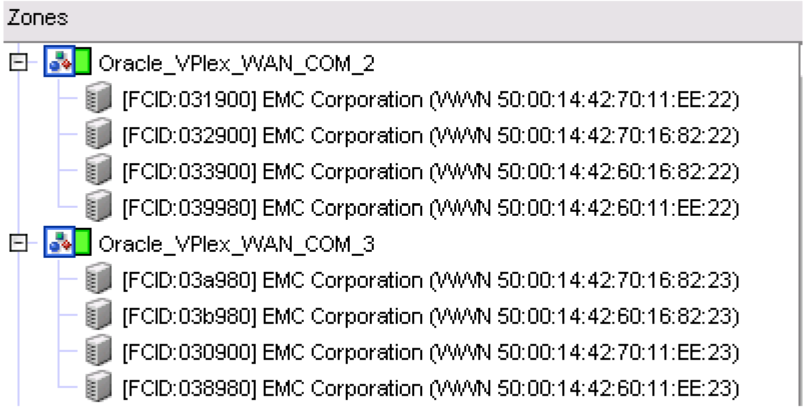
* NOTE: You can use the VPLEX Port WWN Calculator to expose back-end storage before you perform the EZ-Setup. (Download the calculator from Powerlink to your laptop. Then launch the calculator and follow the instructions.)

Figure 8. Overview of VPLEX Metro setup tasks

Note: You must set up both VPLEX Metro clusters as described. You cannot set each cluster up individually and then join them later.

Set up Metro-Plex cluster connectivity

Metro-Plex connectivity is the communication between clusters in a Metro-Plex. The two key components of Metro-Plex communication are FC (FCIP, DWDM) and VPN WAN. FC MAN is the Fibre Channel connectivity between directors of each cluster and the VPN is connectivity between management servers for management purposes. A Metro-Plex should be set up with redundant (dual fabrics) and completely independent Fibre Channel connectivity between clusters for inter-cluster communication. This provides maximum performance, fault isolation, fault tolerance, and availability. Figure 9 is an example of zoning inter-cluster WAN connectivity.



```
Zones
Oracle_VPlex_WAN_COM_2
    [FCID:031900] EMC Corporation (WWN 50:00:14:42:70:11:EE:22)
    [FCID:032900] EMC Corporation (WWN 50:00:14:42:70:16:82:22)
    [FCID:033900] EMC Corporation (WWN 50:00:14:42:60:16:82:22)
    [FCID:039980] EMC Corporation (WWN 50:00:14:42:60:11:EE:22)
Oracle_VPlex_WAN_COM_3
    [FCID:03a980] EMC Corporation (WWN 50:00:14:42:70:16:82:23)
    [FCID:03b980] EMC Corporation (WWN 50:00:14:42:60:16:82:23)
    [FCID:030900] EMC Corporation (WWN 50:00:14:42:70:11:EE:23)
    [FCID:038980] EMC Corporation (WWN 50:00:14:42:60:11:EE:23)
```

Figure 9. Zoning example of VPLEX inter-cluster WAN connectivity

Checking cluster connectivity

To check for FC MAN connectivity, log in to the VPLEX CLI and run the following command:

```
ll **/hardware/ports/
```

An example:

```
VPlexcli:/> ll **/hardware/ports/

/engines/engine-1-1/directors/director-1-1-A/hardware/ports:

Name     Address             Role       Port Status

-------  ------------------  ---------  -----------

A2-FC00  0x500014426011ee20  wan-com    up

A2-FC01  0x500014426011ee21  wan-com    up

A2-FC02  0x500014426011ee22  wan-com    up

A2-FC03  0x500014426011ee23  wan-com    up


/engines/engine-1-1/directors/director-1-1-B/hardware/ports:

Name     Address             Role       Port Status

-------  ------------------  ---------  -----------


B2-FC00  0x500014427011ee20  wan-com    up

B2-FC01  0x500014427011ee21  wan-com    up
```

```
B2-FC02  0x500014427011ee22  wan-com    up
B2-FC03  0x500014427011ee23  wan-com    up

/engines/engine-2-1/directors/director-2-1-A/hardware/ports:
Name     Address             Role       Port Status
-------  ------------------  ---------  -----------
A2-FC00  0x5000144260168220  wan-com    up
A2-FC01  0x5000144260168221  wan-com    up
A2-FC02  0x5000144260168222  wan-com    up
A2-FC03  0x5000144260168223  wan-com    up

/engines/engine-2-1/directors/director-2-1-B/hardware/ports:
Name     Address             Role       Port Status
-------  ------------------  ---------  -----------
B2-FC00  0x5000144270168220  wan-com    up
B2-FC01  0x5000144270168221  wan-com    up
B2-FC02  0x5000144270168222  wan-com    up
B2-FC03  0x5000144270168223  wan-com    up
```

To check FC MAN link status, run **cluster summary** command.

An example:

```
VPlexcli:/> cluster summary
Clusters:
Name       Cluster ID  Connected  Expelled  Operational Status Health State
---------  ----------  ---------  --------  ------------------ ------------
cluster-1  1           true       false     ok                  ok
cluster-2  2           true       false     ok                  ok
Islands:
  Island ID  Clusters
  ---------  --------------------
  1          cluster-1, cluster-2
```

VPLEX Metro host connectivity

To ensure the highest level of connectivity and availability to Oracle Extended RAC even during abnormal operations for connecting Oracle RAC Servers to EMC VPLEX, each Oracle RAC Server in the Oracle Extended RAC Infrastructure environment should have at least two physical HBAs, and each HBA should be connected to front-end ports on different directors on EMC VPLEX. This configuration ensures continued availability of the Oracle RAC Server even if one of the front-end ports of the EMC VPLEX goes offline for either planned maintenance events or unplanned disruptions.

When a single VPLEX engine configuration is connected to an Oracle Extended RAC environment, each HBA should be connected to the front-end ports provided on both the A and B directors within the VPLEX engine. Connectivity to the VPLEX front-end ports should consist of first connecting unique hosts to port 0 of each I/O module emulating the front-end directors before connecting additional hosts to the remaining ports on the I/O module. If multiple VPLEX engines are available, the HBAs from the Oracle RAC Servers should be connected to different engines.

The connectivity from the VPLEX engines to the storage arrays should follow the best practices recommendation for the array. A detailed discussion of the best practices for connecting the back-end storage is beyond the scope of this paper. The *EMC VPLEX Architecture and Deployment: Enabling the Journey to the Private Cloud* TechBook provides more information.

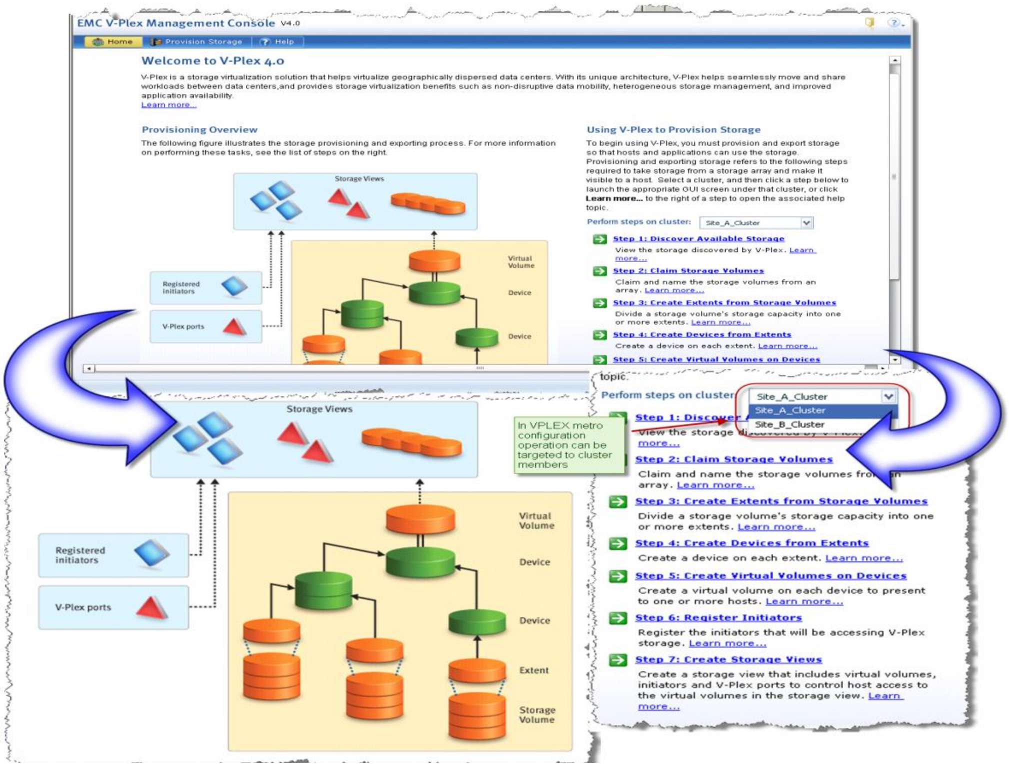
VPLEX Metro administration

Administration of VPLEX Metro can be done primarily through the Management Console, although the same functionality exists with VPLEX CLI. On authenticating to the secure web-based GUI, the user is presented with a set of on-screen configuration options, listed in the order of completion. The EMC VPLEX Management Console online help provides more information about each step in the workflow. The following table summarizes the steps to be taken from the discovery of the arrays up to the storage being visible to the host.

Step	Action
1	**Discover available storage** VPLEX Metro automatically discovers storage arrays that are connected to the back-end ports. All arrays connected to each director in the cluster are listed in the Storage Arrays view.
2	**Claim storage volumes** Storage volumes must be claimed before they can be used in the cluster (with the exception of the metadata volume, which is created from an unclaimed storage volume). Only after storage volume is claimed, it can be used to create extents, devices, and then virtual volumes.
3	**Create extents** Create extents for the selected storage volumes and specify the capacity.
4	**Create devices from extents** A simple device is created from one extent and uses storage in one cluster only.
5	**Create a virtual volume** Create a virtual volume using the device created in the previous step.
6	**Register initiators** When initiators (hosts accessing the storage) are connected directly or through a Fibre Channel fabric, VPLEX Metro automatically discovers them and populates the **Initiators view**. Once discovered, you must register the initiators with VPLEX Metro before they can be added to a storage view and access storage. Registering an initiator gives a

	meaningful name to the port’s WWN, which is typically the server’s DNS name, to allow you to easily identify the host.
7	**Create a storage view** For storage to be visible to a host, first create a storage view and then add VPLEX Metro front-end ports and virtual volumes to the view. Virtual volumes are not visible to the hosts until they are in a storage view with associated ports and initiators.
8	**Create consistency group** Create a consistency group on both VPLEX systems, and add all virtual volumes assigned for Oracle Extended RAC ASM devices, including Grid and ASM devices for Oracle RAC database, that require write-order consistency to the consistency group.

Figure 10 shows the GUI interface for logical layout and provisioning storage from EMC VPLEX.



Welcome to V-Plex 4.0

V-Plex is a storage virtualization solution that helps virtualize geographically dispersed data centers. With its unique architecture, V-Plex helps seamlessly move and share workloads between data centers,and provides storage virtualization benefits such as non-disruptive data mobility, heterogeneous storage management, and improved application availability.
Learn more...

Provisioning Overview

The following figure illustrates the storage provisioning and exporting process. For more information on performing these tasks, see the list of steps on the right.

Using V-Plex to Provision Storage

To begin using V-Plex, you must provision and export storage so that hosts and applications can use the storage. Provisioning and exporting storage refers to the following steps required to take storage from a storage array and make it visible to a host. Select a cluster, and then click a step below to launch the appropriate GUI screen under that cluster, or click **Learn more...** to the right of a step to open the associated help topic.

Perform steps on cluster: Site_A_Cluster

Step 1: Discover Available Storage
View the storage discovered by V-Plex. Learn more...

Step 2: Claim Storage Volumes
Claim and name the storage volumes from an array. Learn more...

Step 3: Create Extents from Storage Volumes
Divide a storage volume's storage capacity into one or more extents. Learn more...

Step 4: Create Devices from Extents
Create a device on each extent. Learn more...

Step 5: Create Virtual Volumes on Devices

In VPLEX metro configuration operation can be targeted to cluster members

topic.

Perform steps on cluster Site_A_Cluster
Site_A_Cluster
Site_B_Cluster

Step 1: Discover
View the storage
more...

Step 2: Claim Storage Volumes
Claim and name the storage volumes fro array. Learn more...

Step 3: Create Extents from Storage Volumes
Divide a storage volume's storage capacity into one or more extents. Learn more...

Step 4: Create Devices from Extents
Create a device on each extent. Learn more...

Step 5: Create Virtual Volumes on Devices
Create a virtual volume on each device to present to one or more hosts. Learn more...

Step 6: Register Initiators
Register the initiators that will be accessing V-Plex storage. Learn more...

Step 7: Create Storage Views
Create a storage view that includes virtual volumes, initiators and V-Plex ports to control host access to the virtual volumes in the storage view. Learn more...

Figure 10. EMC VPLEX GUI management interface

The browser-based management interface, as seen in Figure 10, schematically shows the various components involved in the process. Storage from EMC VPLEX is exposed using a logical construct called “Storage View” that is a union of the objects, “Registered initiators,”, “VPLEX ports,” and “Virtual Volume”. The “Registered initiators” object lists the WWPN of the initiators that need access to the storage. In the case of an Oracle VM Server environment, the “Registered initiators” entity contains the WWPN of the HBAs in the Oracle VM Servers connected to the EMC VPLEX. The object “VPLEX ports” contains the front-end ports of the VPLEX array through which the “Registered initiators” access the virtual volumes. The “Virtual Volume” object is a collection of volumes that are constructed from the storage volumes that are provided to the EMC VPLEX from the back-end storage arrays. It can be seen from the inset in the bottom left corner of Figure 10 that a virtual volume is constructed from a “Device” that in turn can be a combination of different devices built on top of an abstract entity called an “Extent”. The figure also shows that an “Extent” is created from the “Storage Volume” exposed to the EMC VPLEX. However, to leverage array-based replication technologies, we map each storage device from VMAX in its entire (one-to-one mapping) pass-through configuration (device capacity = extent capacity = storage volume capacity) to VPLEX and has a RAID 0 (single extent only) VPLEX device geometry. By doing so, the underlying storage devices are left untouched by VPLEX and the back-end array LUN replication technology, such as TimeFinder/Clone and TimeFinder/Snap, continues to function normally.

Also shown in Figure 10 in the bottom right corner inset are the seven steps that are required to provision storage from EMC VPLEX. The wizard supports a centralized mechanism for provisioning storage to different cluster members in case of EMC VPLEX Metro. The first step in the process of provisioning storage from EMC VPLEX is the discovery of the storage arrays connected to it. This step needs to be rarely executed since the EMC VPLEX proactively monitors for changes to the storage environment. The second step in the process is the “claiming” of storage that has been exposed to EMC VPLEX. The process of claiming the storage creates the object’s “Storage Volume” that is shown in Figure 10. The Create Storage View wizard enables you to create a storage view and add initiators, ports, and virtual volumes to the view. Once all the components are added to the view, it automatically becomes active. When a storage view is active, hosts can see the storage and begin I/O to the virtual volumes. After creating a storage view, you can only add or remove virtual volumes through the GUI. To add or remove ports and initiators, use the CLI. The EMC VPLEX CLI Guide provides comprehensive information above VPLEX Metro commands.

VPLEX Metro with VPLEX Witness

VPLEX Witness is installed as a closed virtual machine 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). 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 cluster(s) to distinguish between inter-cluster network partition failures and cluster failures and automatically resume I/O in these situations.

Host and Oracle setup

Multipathing software setup

As part of the best practice for the connectivity of Oracle RAC Servers to VPLEX, each Oracle RAC Server should have two HBA ports with each port connected to a separate

FC switch for increased availability. In such a configuration, the host must use a multipathing solution to handle multiple paths to the same storage device for the purpose of providing high availability, load balancing, and live migration. You can either install EMC PowerPath for Oracle RAC Server as the multipathing solution, or use the Linux native multipath solution, device mapper, to configure Oracle RAC Servers. For Oracle Extended RAC in an EMC VPLEX Metro configuration, as described in this paper, we installed EMC PowerPath 5.5 on four physical servers.

Install PowerPath rpm on each host

```
 [root@ RAC NODE 1: licoc039 ]rpm -i EMCpower.LINUX-5.5.0.00.00-
275.RHEL5.x86_64.rpm
```

A reboot may be required the first time after PowerPath is installed for the host to register the /dev/emcpower pseudo devices.

Install a PowerPath license on each host

```
  [root@ RAC NODE 1: licoc039 ]emcpreg -add <key>
```

Configure PowerPath on each host

```
[root@ RAC NODE 1: licoc039 ] powermt config

[root@ RAC NODE 1: licoe039 ] powermt display

…

Pseudo name=emcpowerk

Invista ID=FNM00100600231

Logical device ID=6000144000000010A002636D3C679C6A

state=alive; policy=ADaptive; priority=0; queued-IOs=0

==============================================================================

---------------- Host ---------------   - Stor -   -- I/O Path -  -- Stats ---

###  HW Path                I/O Paths    Interf.   Mode    State  Q-IOs Errors

==============================================================================

   1 lpfc                      sdaq      08        active  alive      0      0

   2 lpfc                      sdbu      00        active  alive      0      0

   2 lpfc                      sdcy      08        active  alive      0      0

   1 lpfc                      sdm       00        active  alive      0      0
```

Match PowerPath pseudo device names across Oracle RAC Server nodes

To match the PowerPath pseudo device names across Oracle RAC Server nodes, EMC recommends to use the PowerPath utility **emcpadm**. The utility allows to export the mapping from one host and to import it to another. It also allows the renaming of pseudo devices one at a time if necessary.

```
<source host> emcpadm export_mapping -f <mapping_file_name>
```

Then copy that file to the other hosts. Shut down any application using storage devices, unmount any filesystem, or export any LVM volumes and run:

```
<target host> emcpadm check_mapping [-v] -f <mapping_file_name>
<target host> emcpadm import_mapping -f <mapping_file_name>
```

Create partitions on PowerPath devices

EMC strongly recommends to align x86-based server platform partitions offset to a 64 KB when Symmetrix is used. If another storage array is used it may have different requirements although the size should always be aligned on 4 KB offset to match VPLEX block size. Figure 11 shows alignment at 64 KB to create the partitions on a PowerPath device, simply start fdisk as shown below, and once the partitions are created, type “x” to enter Expert mode. Type “p” to show (print) the current partition table, including the offset in block units. Type “b” to change any partition offset. For example, move partition 1 from its default offset of 32 blocks to 128. Since each block is 512 bytes, then 128 x 512 bytes = 64 KB offset. If more than one partition is created on the LUN, verify that the rest of the partitions are aligned, or follow a similar step to change their offset to a number that is 128 blocks (64 KB) aligned.

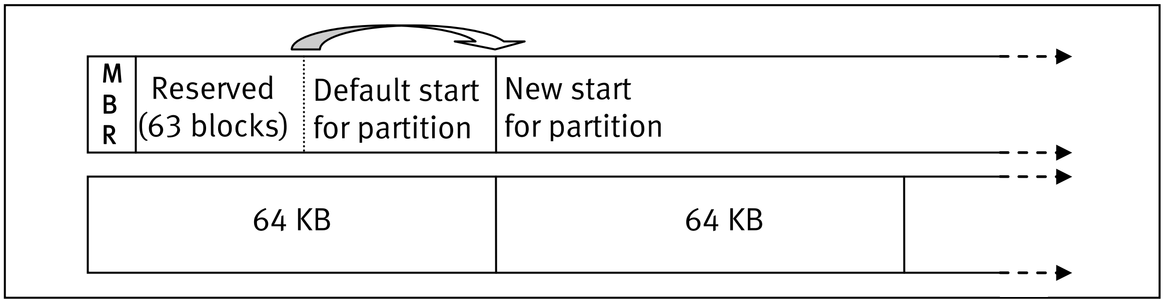


Figure 11. Partition alignment to Symmetrix track size boundary (64 KB)

In this example, one partition is created for a PowerPath device that is going to be used as an Oracle ASM device.

```
[root@licoc091 ~]# fdisk /dev/emcpowerd
Command (m for help): n
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
First cylinder (1-52218, default 1): [ENTER]
Using default value 1
Last cylinder or +size or +sizeM or +sizeK (1-52218, default 52218):
[ENTER]

Command (m for help): p

Disk /dev/emcpowerd: 54.7 GB, 54755328000 bytes
64 heads, 32 sectors/track, 52218 cylinders
Units = cylinders of 2048 * 512 = 1048576 bytes
```

```
   Device           Boot   Start       End       Blocks   Id  System
/dev/emcpowerd1              1       52218     53471168   83  Linux
```

Align partitions on a 64 KB boundary (128 blocks).

```
Command (m for help): x (going into export mode)

Expert command (m for help): p (print partition table)
Note that partition 1 starts at 32 blocks

Disk /dev/emcpowerk: 64 heads, 32 sectors, 52218 cylinders

Nr AF  Hd Sec  Cyl  Hd Sec  Cyl     Start      Size ID
 1 00   1   1    0  63  32 1023         32    106942336 83
 2 00   0   0    0   0   0    0          0          0 00
 3 00   0   0    0   0   0    0          0          0 00
 4 00   0   0    0   0   0    0          0          0 00
Expert command (m for help): b (move partition start)
Partition number (1-4): 1
New beginning of data (32-2002943, default 32): 128

Expert command (m for help): p

Disk /dev/emcpowerk: 64 heads, 32 sectors, 52218 cylinders

Nr AF  Hd Sec  Cyl  Hd Sec  Cyl     Start      Size ID
 1 00   1   1    0  63  32 1023        128  106942336 83
 2 00   0   0    0   0   0    0          0          0 00
 3 00   0   0    0   0   0    0          0          0 00
 4 00   0   0    0   0   0    0          0          0 00

Expert command (m for help): w
The partition table has been altered!

Calling ioctl() to re-read partition table.
Syncing disks.
[root@licoc091 ~]#
```

After the partitions were created ensure the other nodes recognize them. It may be necessary to run the fdisk command on each other node and write (“w”) the partition table. Alternatively, a rescan of the SCSI bus or reboot of the other nodes will refresh the information as well.

Install Oracle and set up RAC database

The following table summarizes the steps to be taken to configure Oracle server nodes for Oracle Grid Infrastructure and ASM database installation, according to *Oracle 11gR2.0.2 Grid Infrastructure Installation Guide for Linux* and *11g Release 2.0.2 Database Installation Guide for Linux*. The installation guide provides further details.

Step	Action
1	Configure server nodes private network, OS /etc/hosts file, OS kernel parameters and edit the /etc/sysctl.conf file.

2	Create Oracle user groups and accounts for installing and maintaining Oracle 11gR2 on each RAC nodes.
3	Update Boot Script (`/etc/rc.d/rc.local`) to set Oracle permissions for devices designated for Oracle CRS and ASM.
4	Set up ssh for Oracle user on each RAC nodes.
5	Set the shell limits for Oracle user in the /etc/security/limits.conf file.
6	Modify the /etc/pam.d/login file and the /etc/profile file accordingly.
7	Install additional required OS packages for Oracle installation.
8	Install Oracle 11g R2.0.2 clusterware for grid infrastructure and create ASM instance for Oracle CRS cluster on all the nodes.
9	Install Oracle 11g R2.0.2 database software.
10	Configure Oracle Cluster Synchronization Services (CSS) for ASM.
11	Create additional ASM disk groups for Oracle database.
12	Create an Oracle database with size and init parameters to meet the desired database workload and performance requirements.
13	Register Database with CRS.

OLTP database workload tests

We used a standard Oracle OLTP workload (70/30 random-read/write ratio, respectively) to illustrate that VPLEX Metro cluster provides Oracle Extended RAC with high performance and workload resiliency over Metro distance of 100 km. As mentioned in the implementation section of the paper, our Oracle Extended RAC with VPLEX Metro test environment consists of two local RAC nodes and two remote RAC nodes, the interconnect between VPLEX clusters and RAC nodes is affected by Metro distance with Empirix PacketSphere Network Emulator for network latency. As shown in Figure 12 and Figure 13, a 16-drivers workload was carried out from each one of Oracle RAC nodes, OLTP workload transactions rates (transactions per minute) were recorded for single node, two nodes, three nodes, or four nodes workload. The workloads were increased with the addition of each node to show the scalability benefits of RAC. The transaction rate increased proportionally as workload increase with addition of more RAC nodes (16, 32, 48, and 64 workload drivers with 1–4 nodes, respectively). Similar transaction rate increases were shown for VPLEX Metro with Oracle RAC at 0 km distance, as well as at 100 km Metro distance. Furthermore, VPLEX Metro is capable of delivering high performance with approximately 85–90percent of transaction rate for Oracle Extended RAC at Metro distance of 100 km, in comparison with baseline, achieved at 0 km distance. In addition, VPLEX Metro is capable of achieving over 90percent of workload performance for Oracle Extended RAC at Metro distance of 50 km or less (data not shown). Therefore, VPLEX Metro provides Oracle Extended RAC with high I/O performance and resiliency for OLTP workload where the two datacenters are located at a Metro distance of 100 km. Tests were performed also with 500 km distance (5 ms RTT) showing similar scalability although with relative lower transaction rate due to increase latency. Overall the solution proved both VPLEX and RAC ability to increase application performance and provide higher availability at the same time. Note that since the OLTP benchmark was indeed completely random, no block contention was encountered. In deployments with real customer workload, the DBA should pay attention (as they always do) to potential block contention across cluster nodes that, especially if occurs between remote nodes, may reduce the overall transaction rate.

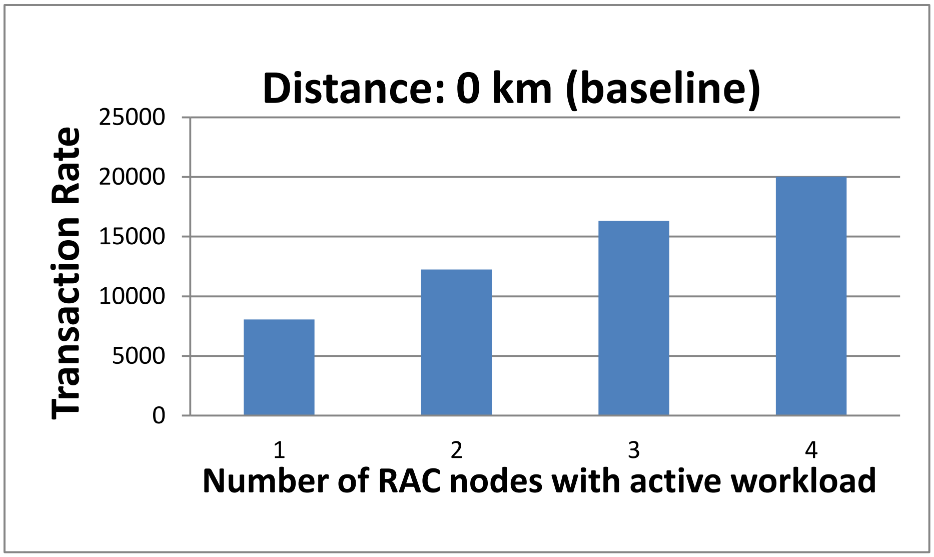


Figure 12. Transaction rate at 0 km distance

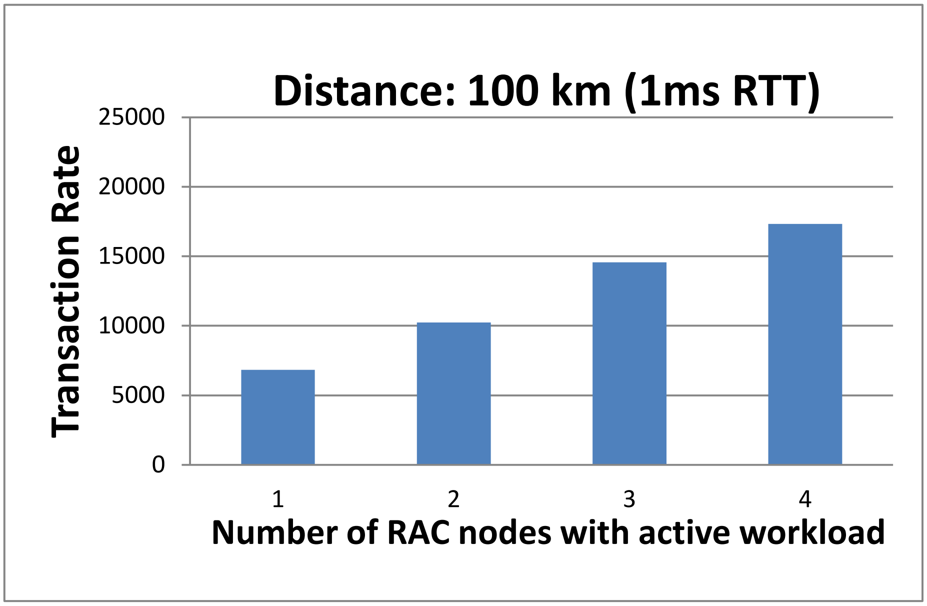


Figure 13. Transaction rate at 100 km distance

Failure conditions tests

The environment was tested successfully for many failure scenarios driven by the project team, E-Lab^{TM} (EMC quality and qualification organization) and based on Oracle RAC engineering provided test plan. All the tests completed successfully with the expected result. All the tests were performed while transaction workload was running (when applicable). A very partial list of the tests performed includes:

- RAC interconnect partitioning (‘split-brain’) while VPLEX Metro infrastructure is intact.

- RAC interconnect and VPLEX Metro interconnect partitioning.
- A single storage system disconnect without application downtime.
- Site failure simulation where the surviving site continues running workload.
- Host connectivity loss for different RAC nodes without affecting the rest of the cluster.
- ASM rebalance, storage, and VPLEX config changes and software updates.

Conclusion

EMC VPLEX running the EMC Metro or GeoSynchrony operating system is an enterprise-class SAN-based federation technology that aggregates and manages pools of Fibre Channel attached storage arrays that can be either collocated in a single data center or multiple data centers that are geographically separated by metro distances. Furthermore, with a unique scale-up and scale-out architecture, EMC VPLEX's advanced data caching and distributed cache coherency provide workload resiliency, automatic sharing, and balancing and failover of storage domains, and enable both local and remote data access with predictable service levels. Oracle Extended RAC dispersed in two datacenters within metro distance backed by the capabilities of EMC VPLEX provides improved performance, scalability, and flexibility. In addition, the capability of EMC VPLEX to provide nondisruptive, heterogeneous data movement, and volume management functionality within synchronous distances enables customers to offer nimble and cost-effective cloud services spanning multiple physical locations.

References

The following documents include more information on VPLEX and can be found on EMC.com and Powerlink:

- *Implementation and Planning Best Practices for EMC VPLEX Technical Notes*
- *EMC VPLEX Metro Witness – Technology and High Availability* TechBook
- *Conditions for stretched hosts cluster support on EMC VPLEX Metro*
- *EMC VPLEX with GeoSynchrony 5.0 Product Guide*