

EMC Solutions for Oracle Database 10g/11g for Midsize Enterprises EMC Celerra Unified Storage Platform

Best Practices Planning



Abstract

The EMC[®] Celerra[®] Unified Storage Platform is a remarkably versatile device. It provides both a world-class NAS device providing NFS access as well as a world-class midrange SAN device, through the front-end ports on the CLARiiON[®] CX4-240 back-end storage array. EMC has provided a unique solution that combines the high performance of FCP with the manageability of NFS. This white paper presents the best practices for configuration, backup, recovery, and protection of this solution in a customer environment.

February 2009

Copyright © 2007, 2008, 2009 EMC Corporation. All rights reserved.

Published February 2009

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

Benchmark results are highly dependent upon workload, specific application requirements, and system design and implementation. Relative system performance will vary as a result of these and other factors. Therefore, this workload should not be used as a substitute for a specific customer application benchmark when critical capacity planning and/or product evaluation decisions are contemplated.

All performance data contained in this report was obtained in a rigorously controlled environment. Results obtained in other operating environments may vary significantly. EMC Corporation does not warrant or represent that a user can or will achieve similar performance expressed in transactions per minute.

No warranty of system performance or price/performance is expressed or implied in this document. 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.

All other trademarks used herein are the property of their respective owners.

Part Number H4160.4

Table of Contents

Executive summary	6
Introduction	6
<i>Audience</i>	7
Configuration	7
<i>Tested configuration</i>	7
Physically booted RAC solutions	8
Virtualized single-instance solutions	9
Solution diagrams.....	10
<i>Storage setup and configuration.....</i>	15
Disk drive recommendations	15
RAID groups.....	15
Number of shelves	15
<i>LUN/RAID group layout: Pure NFS solution.....</i>	17
<i>LUN/RAID group layout: Blended FCP/NFS solution.....</i>	19
<i>Storage setup and configuration specific to the blended FCP/NFS solution.....</i>	21
ASM diskgroup guidelines.....	21
PowerPath/ASM workaround	21
SnapView clone.....	22
Reserved LUN pool.....	22
<i>Storage setup and configuration specific to the pure NFS solution.....</i>	23
Stripe size.....	23
Load distribution	23
High availability	23
Control Station security	24
Data Mover parameters setup.....	24
NFS mount point parameters	24
Protocol overhead	26
sunrpc.tcp_slot_table_entries.....	26
<i>Network setup and configuration</i>	26
Gigabit connection.....	26
Virtual local area networks	26
Jumbo frames for the RAC interconnect.....	26
<i>Database servers setup and configuration</i>	27
BIOS.....	27
Hyper-Threading	27
Memory: Best practices	27
<i>Oracle 10g.....</i>	27
Shared memory	27
SHMMAX setting	27
SHMMNI setting	28
<i>Oracle 11g.....</i>	28
Linux setup and configuration.....	28
Database setup and configuration	29
<i>Initialization parameters.....</i>	29
<i>Control files and log files.....</i>	30
Control files.....	30
Online and archived redo log files	30

Oracle Direct NFS client	30
<i>Implementing DNFS: Mandatory fixes and patches</i>	30
Oracle 11.1.0.7 patch	30
Celerra Data Mover: Enabling transChecksum	30
<i>DNFS network setup</i>	30
<i>Database tuning for Oracle 11g DNFS</i>	31
Reserved port configuration	31
Mounting DNFS	31
Mounting multiple servers	31
Degree of Parallelism (DOP)	31
Backup and recovery: Pure NFS solution	31
<i>Comparison: Logical storage backup and Flashback Database</i>	32
<i>Data protection</i>	33
<i>Database cloning</i>	33
Backup and recovery: Blended FCP/NFS solution	34
<i>Logical storage backup</i>	34
<i>Physical storage backup</i>	35
<i>Comparison: Logical storage backup and Flashback Database</i>	35
<i>CX4 cache configuration for SnapView snapshot</i>	36
Advanced backup and recovery using de-duplication (pure NFS and blended FCP/NFS solutions)	36
Managing and monitoring Celerra	36
<i>Celerra Manager</i>	36
<i>Enterprise Grid Control storage monitoring plug-in</i>	36
Test/dev using Celerra SnapSure writeable snapshots	37
<i>CSS disktimeout for Test/dev solution</i>	37
<i>RAC-to-RAC cloning</i>	37
Advanced protect using RecoverPoint	37
<i>Journal volumes</i>	37
<i>Repository volumes</i>	38
<i>WAN compression guidelines</i>	38
<i>Clusters</i>	38
<i>Zoning</i>	38
Virtualization using VMware ESX	38
<i>LUN discovery</i>	38
<i>VMotion storage requirements</i>	38
<i>NFS connectivity requirements</i>	39
<i>NFS volume recommendations</i>	39
<i>Performance comparison: Pure NFS versus blended FCP/NFS</i>	40
<i>Storage configuration using virtualized solutions</i>	40
<i>VMware HA cluster</i>	41
Compared with Oracle RAC	41
Name resolution	41
Virtualization using Oracle VM	42
Replication Manager	42

<i>Oracle home location</i>	42
<i>Dedicated server process</i>	42
Conclusion	43
References	43
Appendix A: Sample ks.cfg	44

Executive summary

The EMC[®] Celerra[®] Unified Storage Platform is a remarkably versatile device. It provides both a world-class NAS device providing NFS access as well as a world-class midrange SAN device, using the front-end ports on the EMC CLARiON[®] CX4-240 back-end storage array. EMC has provided a unique solution that combines the high performance of FCP with the manageability of NFS in the context of Oracle RAC 10g/11g on Linux.

Oracle over FCP and NFS on a Celerra Unified Storage Platform provides the benefits described in [Table 1](#).

Table 1. Benefits of Oracle over FCP on a Celerra

Benefit	Details
Lower total cost of ownership (TCO)	Reduces acquisition, administration, and maintenance costs more than equivalent to direct-attached storage (DAS).
Reduced cooling, space, and power costs	Virtualization and de-duplication increase density and reduce physical and power costs.
Greater manageability	Eases implementation, provisioning, and LUN management; virtualization provides flexibility in terms of server provisioning and migration.
Simplified Real Application Cluster (RAC) implementation	Provides storage for shared file systems.
High availability	Clustering architecture provides very high levels of data availability.
Improved protection	Integrates both availability and backup.
Benefits of EMC Information Lifecycle Management (ILM)	Implements tiered storage.
Increased flexibility	<ul style="list-style-type: none">• Easily makes databases, or copies of databases, available (through clones) to other servers using storage-based cloning.• Useful for testing and development environments.• Cloned databases are mounted on virtualized servers, further improving flexibility and manageability.

Introduction

This white paper describes the best practices for running Oracle RAC 10g/11g on Red Hat Enterprise Linux and Oracle Enterprise Linux servers with Celerra unified storage systems using both the NFS protocol and the FCP protocol. Oracle performance tuning is beyond the scope of this paper.

The topics covered include the setup and configuration of the following:

- Storage
- Network
- Database server hardware and BIOS
- Linux operating system install
- Oracle software install
- Database parameters and settings
- RAC resiliency
- Backup and recovery, including de-duplication

-
- Virtualization
 - Advanced protect using EMC RecoverPoint

Oracle performance tuning is beyond the scope of this white paper. The [Oracle Database Performance Tuning Guide](#) provides more information on this topic.

Information in this white paper can be used as the basis for a solution build, white paper, best practices document, or training. Information in this paper can also be used by other EMC organizations (for example, the technical services or sales organization) as the basis for producing documentation for a technical services or sales kit.

Audience

The primary target audience for this white paper is database administrators, system administrators, storage administrators, and architects who analyze, design, implement, and maintain robust database and storage systems. Readers should be familiar with Oracle RAC 10g/11g software, basic Linux system administration, basic networking, and EMC products. As such, readers should already be familiar with the installation and administration of their server operating environment and Oracle RAC 10g/11g software.

Configuration

Tested configuration

Four separate configurations were tested, two storage configurations and two database server configurations. The storage configurations included blended (FCP and NFS in combination) and pure (NFS only). The database server configurations included physical four-node RAC and virtualized four database server VMs running on a VMware ESX server. In addition, both Oracle RAC 10g and 11g were tested. Unless otherwise indicated, the comments in this white paper apply to both versions of Oracle RAC software.

Physically booted RAC solutions

The following components were tested:

Table 2. Solution components

Component	Description
Scale-Up OLTP	<ul style="list-style-type: none">• Real-world performance and capacity testing.• Utilizes an industry-standard OLTP database performance benchmark, while providing only real-world tunings on a reasonably priced and configured platform.• All of the scalability is provided on a single database instance. This assumes a monolithic application where all users must have access to all of the data in the database.
Basic Backup and Recovery	<ul style="list-style-type: none">• Uses only the functionality provided by the database server and the operating system software to perform backup and recovery.• Uses the database server's CPU, memory, and I/O channels for all backup, restore, and recovery operations.
Advanced Backup and Recovery (snapshot)	<ul style="list-style-type: none">• Uses additional software components at the storage layer to free up the database server's CPU, memory, and I/O channels from the effects of operations relating to backup, restore, and recovery.• Provides high-performance backup and restore operations, improved space efficiency, or other benefits in comparison to basic backup and recovery.
Advanced Backup and Recovery (de-duplication)	<ul style="list-style-type: none">• Saves acquisition, power, space and cooling costs by increasing the density of storage of Oracle database backups by using a specialized hardware de-duplication array.
Basic Protect	<ul style="list-style-type: none">• Uses tools provided by the operating system and database server software (in the same sense as basic backup) to provide disaster recovery.• Uses the database server's CPU, memory, and I/O channels for all operations relating to the disaster recovery configuration.
Advanced Protect	<ul style="list-style-type: none">• Uses additional software components at the storage layer to enable disaster recovery, thereby freeing up the database server's CPU, memory, and I/O channels from the effects of these operations.• Enables the creation of a writeable copy of the production database on the disaster recovery target, allowing this database to be used for operations such as backup, test/dev, and data warehouse staging.
Resiliency	<ul style="list-style-type: none">• Every significant layer of the solution is tested by introducing faults in an effort to cause the solution to fail. In the process, the entire solution is shown to be resilient to faults at every layer, including database clustering, networking, and storage.
Test/dev	<ul style="list-style-type: none">• A running production OLTP database is cloned with minimal, if any, performance impact on the production server, as well as no downtime. The resulting dataset is provisioned on another server for use for testing and development. This is a critical capability for many midsize enterprise customers.

Virtualized single-instance solutions

The following components were tested:

Table 3. Solution components

Component	Description
Scale-Out OLTP	<ul style="list-style-type: none">• Real-world performance and capacity testing.• Utilizes an industry-standard OLTP database performance benchmark, while providing only real-world tunings on a reasonably priced and configured platform.• Scalability is provided by adding additional database instances that are not clustered and that access their own physical database. This assumes that the database application can be broken down into many small, independent databases, and that no single user needs to see the data of any other user outside of the database associated with that user. A typical example would be Software as a Service (SaaS).
Basic Backup and Recovery	<ul style="list-style-type: none">• Uses only the functionality provided by the database server and the operating system software to perform backup and recovery.• Uses the database server's CPU, memory, and I/O channels for all backup, restore, and recovery operations.
Advanced Backup and Recovery (snapshot)	<ul style="list-style-type: none">• Uses additional software components at the storage layer to free up the database server's CPU, memory, and I/O channels from the effects of operations relating to backup, restore, and recovery.• Provides high-performance backup and restore operations, improved space efficiency, or other benefits in comparison to basic backup and recovery.
Basic Protect	<ul style="list-style-type: none">• Uses tools provided by the operating system and database server software (in the same sense as basic backup) to provide disaster recovery.• Uses the database server's CPU, memory, and I/O channels for all operations relating to the disaster recovery configuration.
Advanced Protect	<ul style="list-style-type: none">• Uses additional software components at the storage layer to enable disaster recovery, thereby freeing up the database server's CPU, memory, and I/O channels from the effects of these operations.• Enables the creation of a writeable copy of the production database on the disaster recovery target, allowing this database to be used for operations such as backup, test/dev, and data warehouse staging.
Resiliency	<ul style="list-style-type: none">• Every significant layer of the solution is tested by introducing faults in an effort to cause the solution to fail. In the process, the entire solution is shown to be resilient to faults at every layer, including database clustering, networking, and storage.
Test/dev	<ul style="list-style-type: none">• A running production OLTP database is cloned with minimal, if any, performance impact on the production server, as well as no downtime. The resulting dataset is provisioned on another server for use for testing and development. This is a critical capability for many midsize enterprise customers.

Solution diagrams

The next five pages provide figures of the five solution configurations.

Figure 1 provides an overview of the pure NFS physically booted solution for Oracle RAC 10g/11g Celerra.

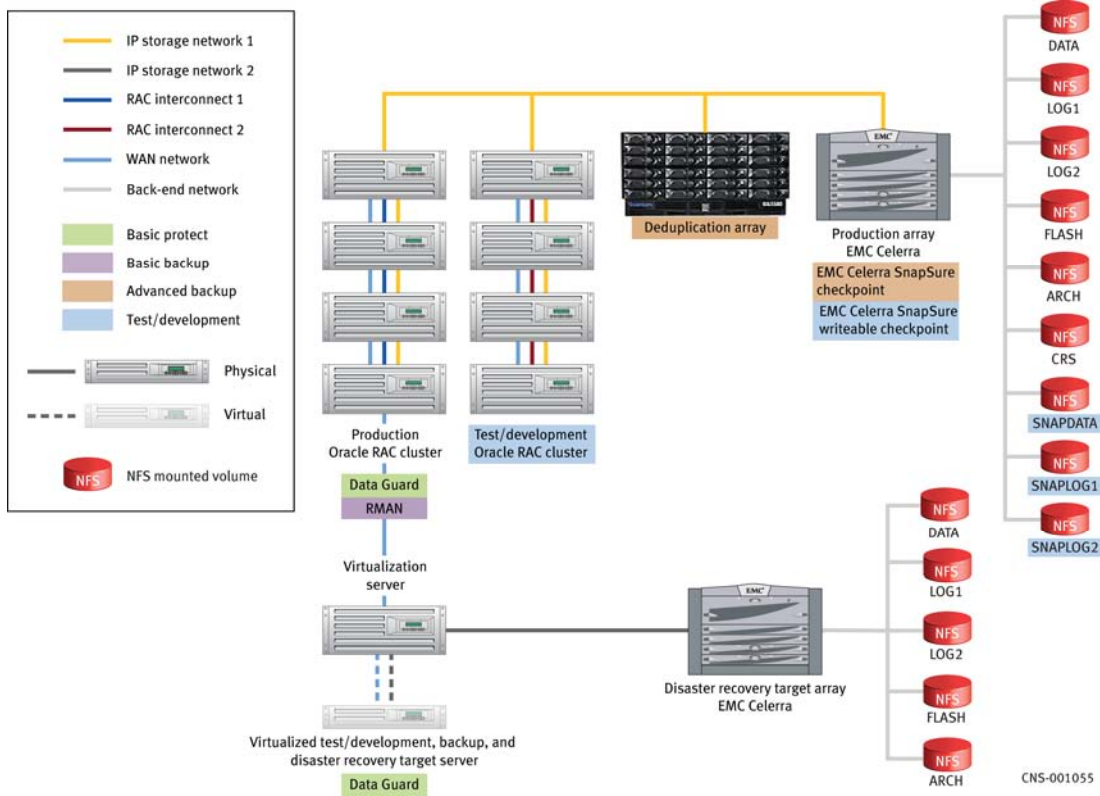


Figure 1. Pure NFS physically booted solution: Oracle RAC and Celerra

Figure 2 provides an overview of the blended FCP/NFS physically booted solution for Oracle RAC 10g/11g and Celerra.

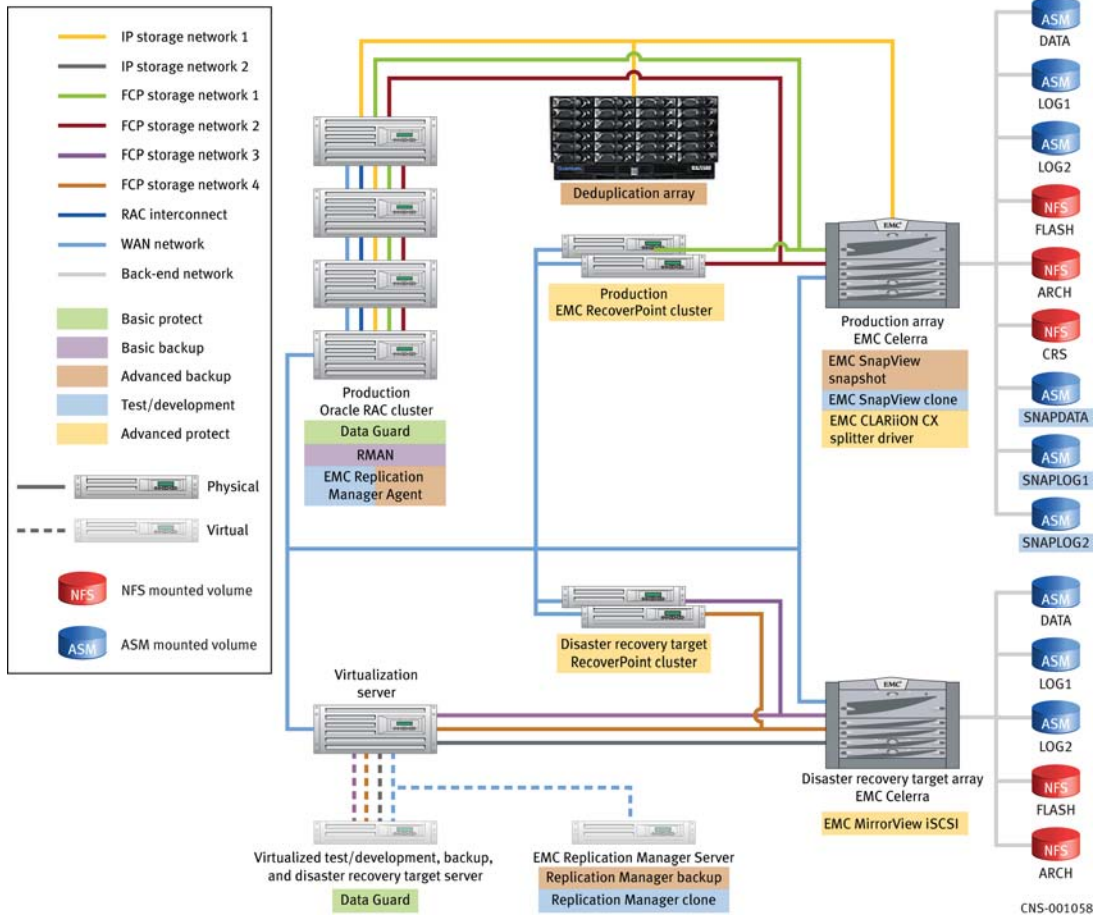


Figure 2. Blended FCP/NFS physically booted solution: Oracle RAC and Celerra

Figure 3 provides an overview of the pure NFS virtualized solution for Oracle RAC 10g/11g and Celerra.

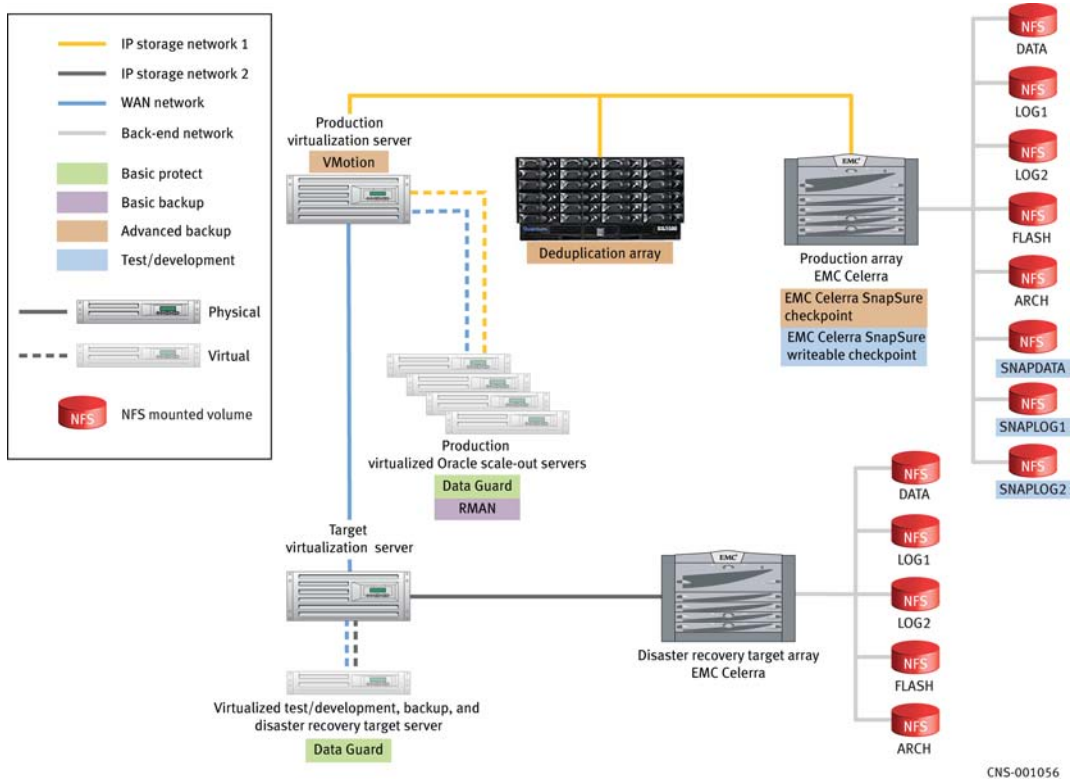


Figure 3. Pure NFS virtualized solution: Oracle RAC and Celerra

Figure 4 provides an overview of the pure NFS VMware high-availability (HA) cluster solution for Oracle RAC 10g/11g and Celerra.

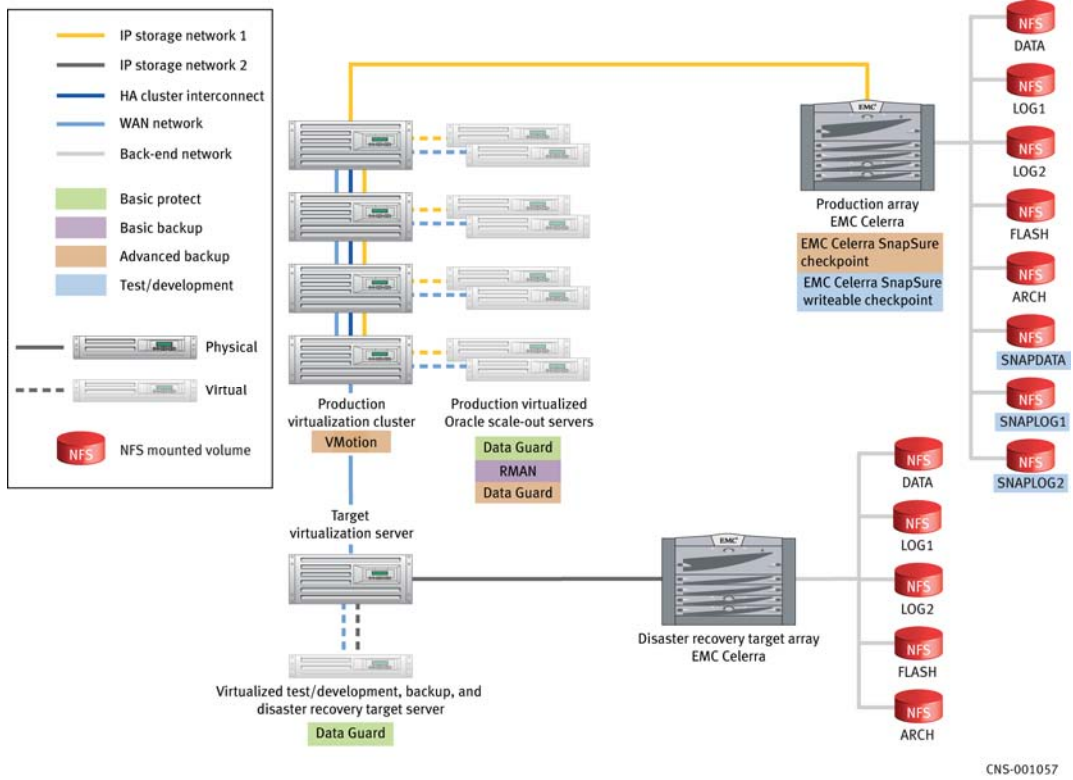


Figure 4. Pure NFS VMware HA cluster solution: Oracle RAC and Celerra

Figure 5 provides an overview of the blended FCP/NFS virtualized solution for Oracle RAC 10g/11g and Celerra.

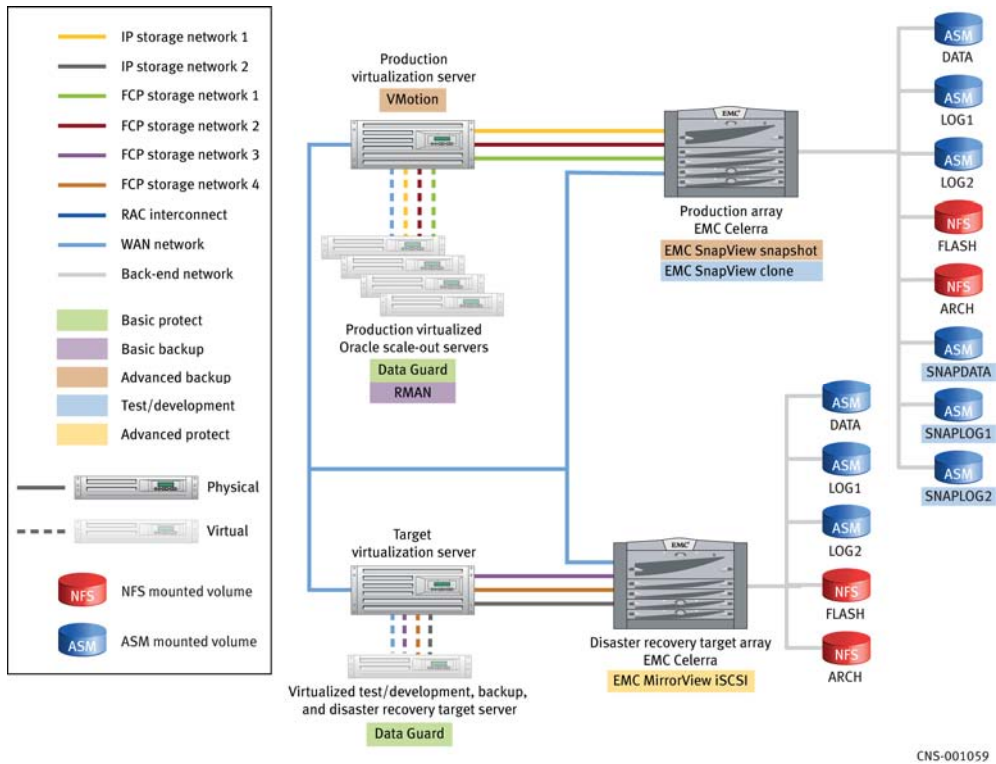


Figure 5. Blended FCP/NFS virtualized solution: Oracle RAC and Celerra

Storage setup and configuration

Disk drive recommendations

The following are the general recommendations for disk drive settings:

- Drives with higher revolutions per minute (rpm) provide higher overall random-access throughput and shorter response times than drives with slower rpm. For optimum performance, higher-rpm drives are recommended.
- Because of significantly better performance, Fibre Channel drives are always recommended for storing datafiles and online redo log files.
- Serial Advanced Technology-Attached (SATA II) drives have slower response, rotational speed, and moderate performance with random I/O. However, they are less expensive than the Fibre Channel drives for the same or similar capacity. SATA drives are frequently the best option for storing archived redo logs and the flashback recovery area. In the event of high performance requirements for backup and recovery, Fibre Channel drives can also be used for this purpose.

RAID groups

Table 4 summarizes the general recommendations for the RAID types corresponding to different Oracle file types.

Table 4. Recommendations for RAID types corresponding to Oracle file types

Description	RAID 10/FC	RAID 5/FC	RAID 5/SATA II
Datafiles	Recommended ¹	Recommended	Avoid
Control files	Recommended	Recommended	Avoid
Online redo logs	Recommended ²	Avoid	Avoid
Archived logs	OK	OK	Recommended
Flashback recovery area	OK	OK	Recommended
OCR file/voting disk	OK ³	OK	Avoid

The tables in the “LUN/RAID group layout: Pure NFS solution” and “LUN/RAID group layout: Blended FCP/NFS solution” sections contain the storage templates that can be used for Oracle RAC 10g/11g databases on a Celerra. That section can help you determine the best configuration to meet your performance needs.

Number of shelves

For high performance, EMC recommends that you use a minimum of two Fibre Channel shelves and one SATA shelf to store Oracle databases on a Celerra. The most common error when planning storage is designing for capacity rather than for performance. The most important single storage parameter for

¹ In some cases, if an application creates a large amount of temp activity, placing your tempfiles on RAID 10 devices may be faster due to RAID 10's superior sequential I/O performance. This is also true for undo. Further, an application that performs many full table scans or index scans may benefit from these datafiles being placed on a RAID 10 device.

² Online redo log files should be put on RAID 1 devices. You should not use RAID 5 because sequential write performance of distributed parity (RAID 5) is not as high as that of mirroring (RAID 1). Further, RAID 1 provides the best data protection, and protection of online redo log files is critical for Oracle recoverability.

³ You should use FC disks for these files as unavailability of these files for any significant period of time (due to disk I/O performance issues) may cause one or more of the RAC nodes to reboot and fence itself from the cluster.

performance is disk latency. High disk latency is synonymous with slower performance; low disk counts lead to increased disk latency.

The recommendation is a configuration that produces average database I/O latency (the Oracle measurement db file sequential read) of less than or equal to 20 ms. In today's disk technology, the increase in storage capacity of a disk drive has outpaced the increase in performance. Therefore, the performance capacity must be the standard to use when planning an Oracle database's storage configuration, not disk capacity.

LUN/RAID group layout: Pure NFS solution

Two sets of RAID and disk configurations were tested on the NFS protocol, as shown in Figure 6 and Figure 7:

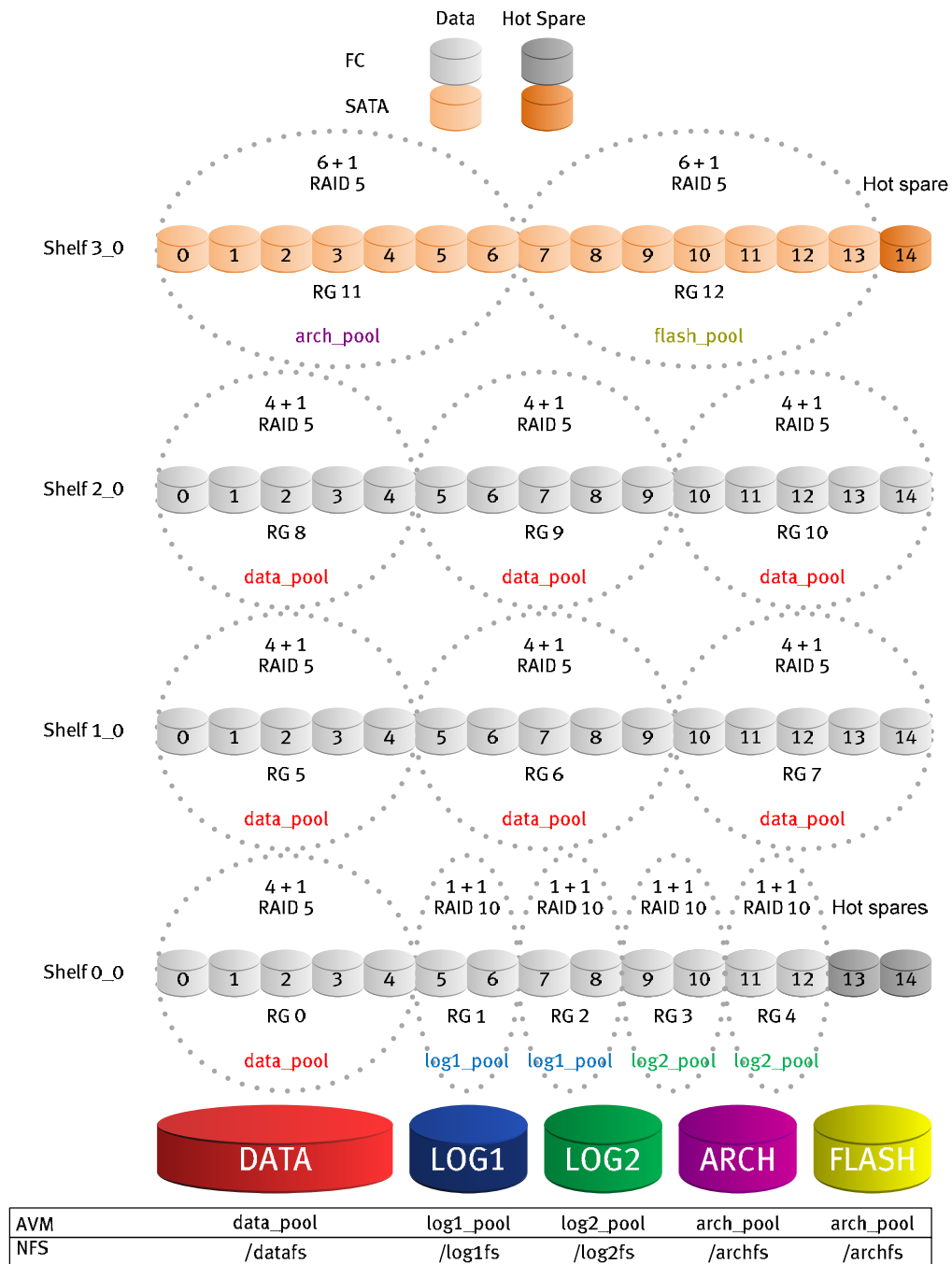


Figure 6. Pure NFS: AVM with user-defined storage pools: 3-FC shelf configuration

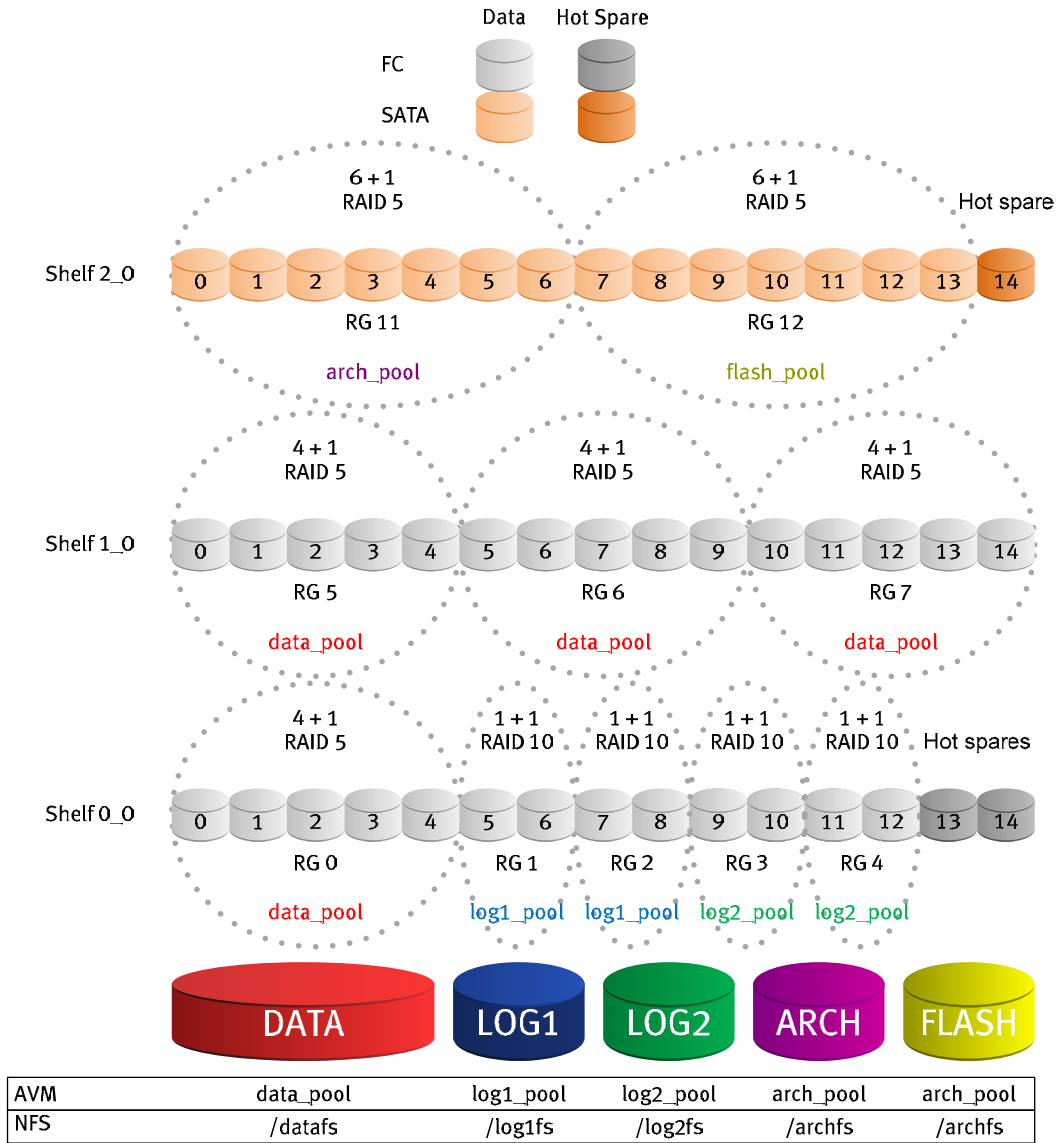


Figure 7. Pure NFS: AVM with user-defined storage pools: 2-FC shelf configuration

LUN/RAID group layout: Blended FCP/NFS solution

A LUN/RAID group configuration consisting of three Fibre Channel shelves with RAID 10 and RAID 1 was tested and found to provide good performance for Oracle RAC 11g databases on Celerra.

Two sets of RAID and disk configurations were tested over the FCP protocol, as shown in [Table 5](#):

Table 5. RAID and disk configuration

Figure	Configuration	Description
Figure 8	Configuration 1	3 FC shelf RAID 10/RAID 1
Figure 9	Configuration 2	3 FC shelf RAID 5/RAID 1

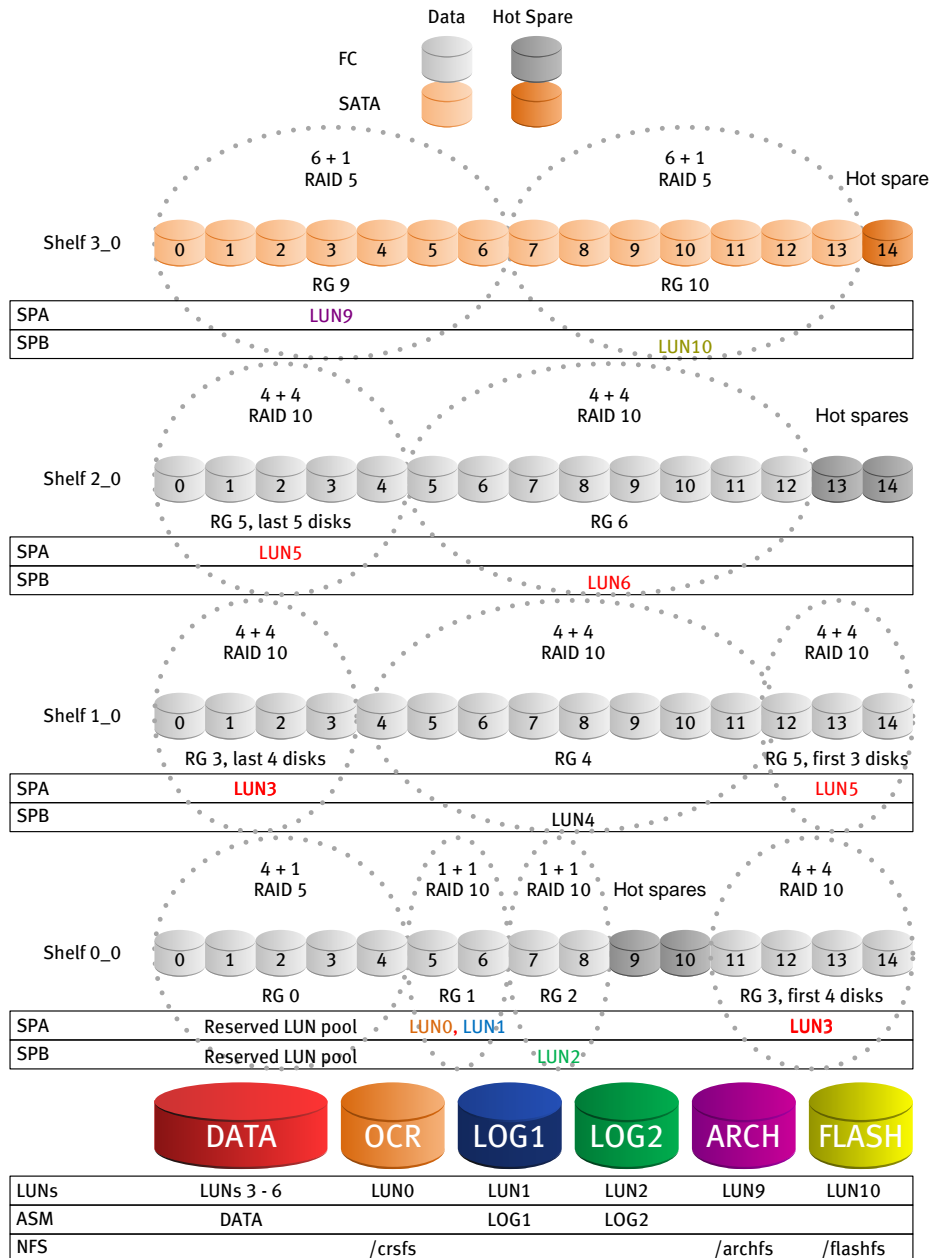


Figure 8. Blended configuration 1: 3 FC shelf RAID 10

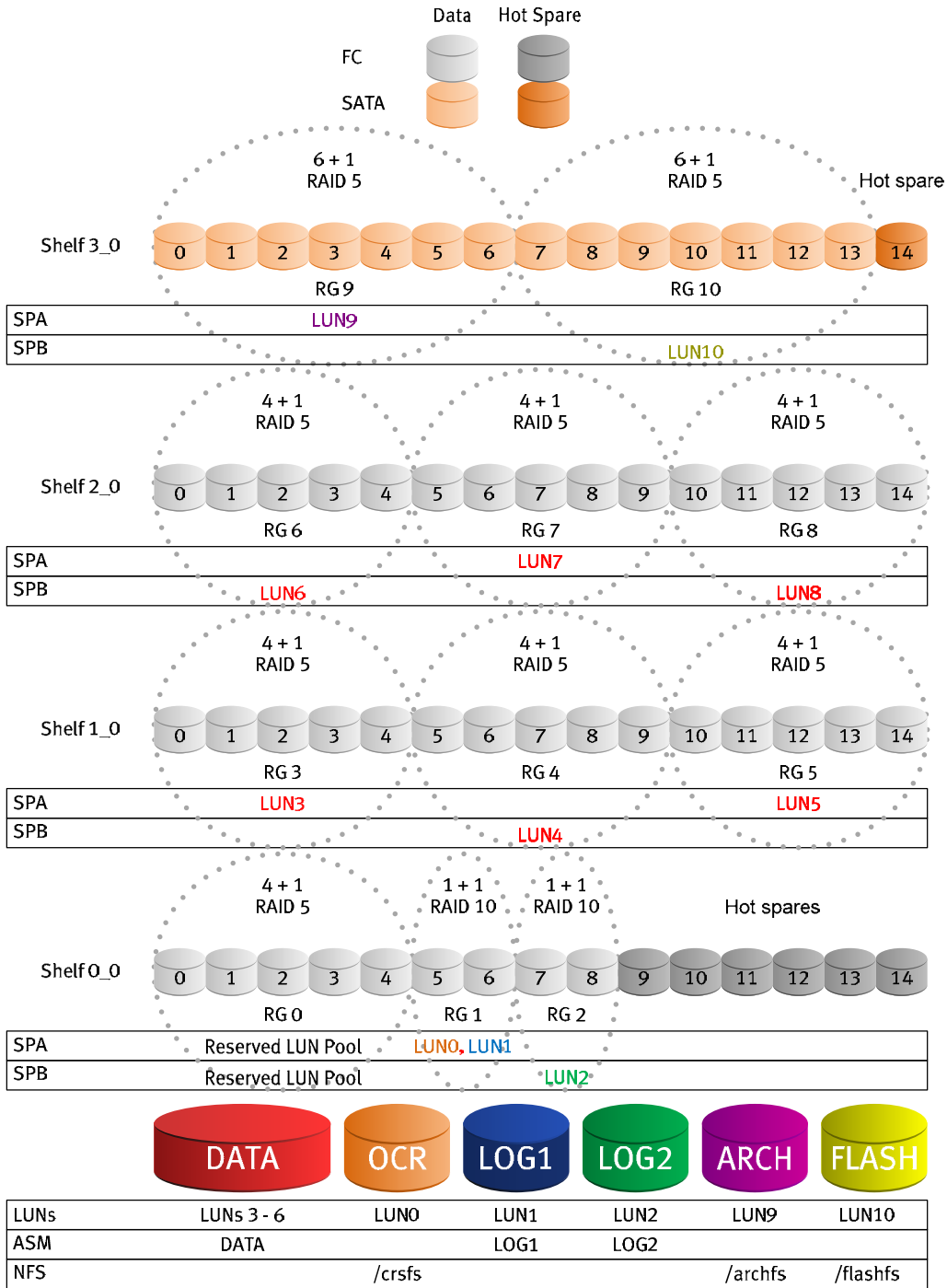


Figure 9. Blended configuration 2: 3 FC shelf RAID 5/RAID 10

Storage setup and configuration specific to the blended FCP/NFS solution

ASM diskgroup guidelines

Automatic Storage Management (ASM) is used to store the database objects requiring high performance. This does not work for the Oracle Cluster Registry file and the voting disk. These two files must be stored on shared storage that is not on the ASM file system. Therefore, NFS is used to store these files.

In addition, files not requiring high performance are stored on NFS. These NFS file systems are in turn stored on low-cost SATA II drives. This drives down the cost in terms of storage, while improving manageability.

Table 6 contains a detailed description of all the database objects and where they are stored.

Table 6. File system layout

File system/ mount point	File system type	LUNs stored on	Contents
/u02	NFS	LUN 0	Oracle Cluster Registry file and voting disk
+DATA	ASM	LUNs 5 through 10	Oracle datafiles
+LOG1 and +LOG2	ASM	LUNs 1 through 4	Online redo logs and control file (mirrored copies)
/u03	NFS	LUN 9	Flashback recovery area (all backups stored here)
/u04	NFS	LUN 10	Archived log dump destination

Best practices for ASM diskgroup design dictate that a diskgroup should consist entirely of LUNs that are all of the same RAID type and that consist of the same number and type of component spindles. Thus, EMC does not recommend mixing any of the following within a single ASM diskgroup:

- RAID levels
- Disk types
- Disk rotational speeds

PowerPath/ASM workaround

The validation was performed using Oracle Enterprise Linux 5.1 on both pure NFS and blended FCP/NFS solutions. PowerPath® 5.1 was required for this version of Linux for the blended FCP/NFS solution but this version was not yet GA at the time the validation was started. Therefore a GA release candidate was used for these tests.

On this version of PowerPath, we observed the following issue while creating ASM disks.

```
[root@mteoradb55 ~]# service oracleasm createdisk LUN1 /dev/emcpower1
Marking disk "/dev/emcpower1" as an ASM disk: asmtool: Device "/dev/emcpower1"
is not a partition
[ FAILED ]
```

We used the following workaround to create the ASM disks.

```
[root@mteoradb55 ~]# /usr/sbin/asmtool -C -l /dev/oracleasm -n LUN1 \
```

```
> -s /dev/emcpower1 -a force=yes
asmtool: Device "/dev/emcpower1" is not a partition
asmtool: Continuing anyway
```

```
[root@mteoradb55 ~]# service oracleasm scandisks
Scanning system for ASM disks: [ OK ]
[root@mteoradb55 ~]# service oracleasm listdisks
LUN1
[root@mteoradb55 ~]#
```

SnapView clone

SnapView™ clone can be used to clone a running Oracle RAC 10g/11g production database for rapidly creating testing and development copies of this database. The steps to accomplish this are provided in the *EMC Solutions for Oracle Database 10g/11g for Midsize Enterprises EMC Celerra Unified Storage Platform - Applied Technology Guide*. A number of best practices should be followed when doing so.

Flashback recovery area and archived log LUNs

While cloning the LUNs, it is better to avoid cloning the archived log and flashback recovery area LUNs. These LUNs are usually large. Further, they are typically stored on LCFC or SATA II disks. The combination of these settings means that cloning these LUNs will take much longer than cloning the datafile LUNs. Instead, you can configure a separate set of LUNs for the archived logs and flashback recovery area for the testing or development database, if desired. Since the test/dev database can be easily refreshed, you may choose to simply forgo backing up these databases. In this case, a flashback recovery area is not required. The archived logs from the production database should be accessible to the test/dev database, in order for it to recover successfully in many cases. The one exposure is that significant transactional I/O on the test/dev database could create an archived log. If you do not change the archive dump destination of the test/dev database, and it is using the same LUNs for storing the archived logs as the production database, the test/dev database could overwrite the archived logs of the production database. This could destroy the recoverability of the production database. Thus, you must always move the archived log dump destination of the test/dev database to a new set of LUNs before opening this database for transactional use. This setting is contained in the pfile or spfile of the test/dev database in the parameters LOG_ARCHIVE_DUMP_DEST1 to LOG_ARCHIVE_DUMP_DEST10.

The time taken to clone these LUNs is very high due to the huge size of these LUNs and also due to the fact that these are SATA II disk-based LUNs.

SyncRate

While initializing the clone, the most important setting is the SyncRate. If you need the test/dev database to be created rapidly, you should specify the option -SyncRate as high. This will speed up the synchronization process, at the cost of a greater performance impact on the production database. If performance on the production database is your primary concern, you should specify the option -SyncRate as low.

An example of the naviseccli command using the SyncRate option is shown below.

```
naviseccli -address 10.6.24.205 snapview -addclone -name lun0CloneGrp -luns 50
-SyncRate <high|medium|low|value>
```

If you do not specify this option, then medium is set as the default.

Reserved LUN pool

It is better to configure the reserved LUN pool with a higher number of LUNs with less capacity than with a lower number of LUNs with higher capacity. There is no benefit in assigning LUNs with a high capacity, such as 25 GB to 30 GB, to the reserved LUN pool. Usually a very small capacity like 2 GB to 3 GB is only used out of these reserved LUNs. It is better to have more LUNs assigned to the reserved LUN pool with capacities of 5 GB to 8 GB. Approximately 20 to 25 small LUNs are sufficient for most purposes.

Storage setup and configuration specific to the pure NFS solution

Stripe size

EMC recommends a stripe size of 32 KB for all types of database workloads.

The default stripe size for all the file systems on FC shelves (redo logs and data) should be 32 KB. Similarly, the recommended stripe size for the file systems on SATA II shelves (archive and flash) should be 256 KB.

Currently, the default stripe size for AVM is 8 KB. If you decide to use AVM, you should change this setting to 32 KB for optimal performance. To change the default AVM stripe size for volumes created using the `clar_r5_performance` profile, execute the following command on the Celerra Control Station:

```
nas_cmd @nas_profile -modify data_stripe -stripe_size 32768
```

Load distribution

For tablespaces with heavy I/O workloads consisting of concurrent reads and writes, EMC recommends spreading the I/O across multiple datafiles.

High availability

The Data Mover failover capability is a key feature unique to the Celerra. This feature offers redundancy at the file-server level, allowing continuous data access. It also helps to build a fault-resilient RAC architecture.

EMC recommends that you set up an auto-policy for the Data Mover, so if a Data Mover fails, either due to hardware or software failure, the Control Station immediately fails the Data Mover over to its partner. The standby Data Mover assumes the faulted Data Mover's:

- Network identity: The IP and MAC addresses of all its NICs
- Storage identity: The file systems that the faulted Data Mover controlled
- Service identity: The shares and exports that the faulted Data Mover controlled

This ensures continuous file sharing transparently for the database without requiring users to unmount and remount the file system. The NFS applications and NFS clients do not see any significant interruption in I/O.

Data Mover failover occurs if any of these conditions exists:

- Failure (operation below the configured threshold) of both internal network interfaces by the lack of a heartbeat (Data Mover timeout)
- Power failure within the Data Mover (unlikely as the Data Mover is typically wired into the same power supply as the entire array)
- Software panic due to exception or memory error
- Data Mover hang

Data Mover failover does not occur under these conditions:

- Removing a Data Mover from its slot
- Manually rebooting a Data Mover

Since manual rebooting of Data Mover does not initiate a failover, EMC recommends that you initiate a manual failover before taking down a Data Mover for maintenance.

The synchronization services component (CSS) of Oracle Clusterware maintains two heartbeat mechanisms:

- The disk heartbeat to the voting disk
- The network heartbeat across the RAC interconnects that establishes and confirms valid node membership in the cluster

Both of these heartbeat mechanisms have an associated time-out value. For more information on Oracle Clusterware `MissCount` and `DiskTimeout` parameters see [Metalink Note 2994430.1](#).

EMC recommends setting the disk heartbeat parameter `disktimeout` to 160 seconds. You should leave the network heartbeat parameter `misscount` at the default of 60 seconds. These settings will ensure that the RAC nodes do not evict when the active Data Mover fails over to its partner. The command to configure this option is:

```
$ORA_CRS_HOME/bin/crsctl set css disktimeout 160
```

Control Station security

The Control Station is based on a variant of Red Hat Linux. Therefore it is possible to install any publicly available system tools that your organization may require.

Data Mover parameters setup

Noprefetch

EMC recommends that you turn off file-system read prefetching for an online transaction processing (OLTP) workload. Leave it on for Decision Support System (DSS) workload. Prefetch will waste I/Os in an OLTP environment, since few, if any, sequential I/Os are performed. In a DSS, setting the opposite is true.

To turn off the read prefetch mechanism for a file system, type:

```
$ server_mount <movername> -option <options>,noprefetch <fs_name> <mount_point>
```

For example:

```
$ server_mount server_3 -option rw,noprefetch ufs1 /ufs1
```

Network File System threads

EMC recommends that you use the default Network File System (NFS) thread count of 256 for optimal performance. Please do not set this to a value lower than 32 or higher than 512. The *Celerra Network Server Parameters Guide* has more information.

file.asyncthreshold

EMC recommends that you use the default value of 32 for the parameter `file.asyncthreshold`. This provides optimum performance for databases. The *Celerra Network Server Parameters Guide* has more information.

NFS mount point parameters

For optimal reliability and performance, the NFS client options in [Table 7](#) are recommended. The mount options are listed in the `/etc/fstab` file.

Table 7. NFS client options

Option	Syntax	Recommended	Description
Hard mount	hard	Always	With this option, the NFS file handles are kept intact when the NFS server does not respond. When the NFS server responds, all the open file handles resume, and do not need to be closed and reopened by restarting the application. This option is required for Data Mover failover to occur transparently without having to restart the Oracle instance.
NFS protocol version	vers= 3	Always	This option sets the NFS version to be used. Version 3 is recommended.
TCP	proto=tcp	Always	With this option, all the NFS and RPC requests will be transferred over a connection-oriented protocol. This is required for reliable network transport.
Background	bg	Always	This setting enables client attempts to connect in the background if the connection fails.
No interrupt	nointr	Always	This toggle allows or disallows client keyboard interruptions to kill a hung or failed process on a failed hard-mounted file system.
Read size and write size	rsize=32768,wsiz=32768	Always	This option sets the number of bytes NFS uses when reading or writing files from an NFS server. The default value is dependent on the kernel. However, throughput can be improved greatly by setting rsize/wsize= 32768
No auto	noauto	Only for backup/utility file systems	This setting disables automatic mounting of the file system on boot-up. This is useful for file systems that are infrequently used (for example, stage file systems).
ac timeo	actimeo=0	RAC only	This sets the minimum and maximum time for regular files and directories to 0 seconds.
Timeout	timeo=600	Always	This sets the time (in tenths of a second) the NFS client waits for the request to complete.

The NFS parameter `sunrpc.tcp_slot_table_entries` was set to the maximum allowable setting of 128 (this was done to increase the concurrent I/Os to be submitted to the storage system from the default value of 16):

```
[root@mteoradb51 mterac5]# sysctl sunrpc.tcp_slot_table_entries
sunrpc.tcp_slot_table_entries = 128
```

Protocol overhead

Typically, in comparison to the host file system implementations, NFS implementations increase database server CPU utilization by 1 percent to 5 percent. However, most online environments are tuned to run with significant excess CPU capacity. EMC testing has confirmed that in such environments protocol CPU consumption does not affect the transaction response times.

sunrpc.tcp_slot_table_entries

There is a NFS module called “sunrpc.tcp_slot_table_entries”. This parameter controls the concurrent I/Os to the storage system and it should be set to the maximum value for enhanced I/O performance.

The command to configure this option is:

```
[root@mteoraesx2-vm3 ~]# sysctl -w sunrpc.tcp_slot_table_entries=128
sunrpc.tcp_slot_table_entries = 128
```

Network setup and configuration

Gigabit connection

EMC recommends that you use Gigabit Ethernet for the RAC interconnects if RAC is used.

Virtual local area networks

Virtual local area networks (VLANs) are logical groupings of network devices.

EMC recommends that you use VLANs to segment different types of traffic to specific subnets. This provides better throughput, manageability, application separation, high availability, and security.

[Table 8](#) describes the database server network port setup.

Table 8. Database server network port setup

VLAN ID	Description	CRS setting
1	Client network	Public
2	RAC interconnect	Private

Our tested configuration also used dedicated redundant network switches for the RAC interconnect, as shown in [Figure 3](#).

Jumbo frames for the RAC interconnect

Maximum Transfer Unit (MTU) sizes of greater than 1,500 bytes are referred to as jumbo frames. Jumbo frames require Gigabit Ethernet across the entire network infrastructure – server, switches, and database servers. Whenever possible, EMC recommends the use of jumbo frames on all legs of the RAC interconnect networks. For Oracle RAC 10g/11g installations, jumbo frames are recommended for the private RAC interconnect to boost the throughput as well as to possibly lower the CPU utilization due to the software overhead of the bonding devices. Jumbo frames increase the device MTU size to a larger value (typically 9,000 bytes).

Typical Oracle database environments transfer data in 8 KB and 32 KB block sizes, which require multiple 1,500 frames per database I/O, while using an MTU size of 1,500. Using jumbo frames, the number of frames needed for every large I/O request can be reduced and thus the host CPU needed to generate a large number of interrupts for each application I/O is reduced. The benefit of jumbo frames is primarily a complex function of the workload I/O sizes, network utilization, and Oracle database server CPU utilization, so it is not easy to predict.

Detailed instructions on configuring jumbo frames are contained in the *EMC Solutions for Oracle Database 10g/11g for Midsize Enterprises EMC Celerra Unified Storage Platform - Applied Technology Guide*. For information on using jumbo frames with the RAC Interconnect, see [Metalink Note 300388.1](#).

Database servers setup and configuration

BIOS

Dell PowerEdge 2900 servers were used in our testing. These servers were preconfigured with the A06 BIOS. Upgrading the BIOS to the latest version (2.2.6 as of the time of this publication) resolved a range of issues, including hanging reboot problems and networking issues.

Regardless of the server vendor and architecture, you should monitor the BIOS version shipped with the system and determine if it is the latest production version supported by the vendor. If it is not the latest production version supported by the vendor, then flashing the BIOS is recommended.

Hyper-Threading

Intel Hyper-Threading Technology allows multi-threaded operating systems to view a single physical processor as if it were two logical processors. A processor that incorporates this technology shares CPU resources among multiple threads. In theory, this enables faster enterprise-server response times and provides additional CPU processing power to handle larger workloads. As a result, server performance will supposedly improve. In EMC's testing, however, performance with Hyper-Threading was poorer than performance without it. For this reason, EMC recommends disabling Hyper-Threading. There are two ways to disable Hyper-Threading: in the kernel or through the BIOS. Intel recommends disabling Hyper-Threading in the BIOS because it is cleaner than doing so in the kernel. Please refer to your server vendor's documentation for instructions.

Memory: Best practices

Oracle 10g

EMC's Oracle RAC 10g testing was done with servers containing 24 GB of RAM. This was the maximum memory capacity of the server platform (Dell PowerEdge 2900) at the time it was purchased.

Please refer to your database server documentation to determine the total number of memory slots your database server has, and the number and density of memory modules that you can install. EMC recommends that you configure the system with the maximum amount of memory feasible to meet the scalability and performance needs. Compared to the cost of the remaining components in an Oracle database server configuration, the cost of memory is minor. Configuring an Oracle database server with the maximum amount of memory is entirely appropriate.

Shared memory

Oracle uses shared memory segments for the Shared Global Area (SGA), which is an area of memory that is shared by Oracle processes. The size of the SGA has a significant impact on the database performance.

EMC's Oracle RAC 10g testing was done with servers using 20 GB of SGA.

SHMMAX setting

This parameter defines the maximum size in bytes of a single shared memory segment that a Linux process can allocate in its virtual address space. Since the SGA is comprised of shared memory, SHMMAX can potentially limit the size of the SGA. SHMMAX should be slightly larger than the SGA size.

As the SGA size was set to 20 GB, SHMMAX was set to 24 GB as shown:

```
kernel.shmmax = 25769803776
```

SHMMNI setting

This parameter sets the system-wide maximum number of shared memory segments. Oracle recommends SHMMNI to be at least 4096 for Oracle 10g as shown:

```
kernel.shmmni = 4096
```

Oracle 11g

EMC's Oracle RAC 11g testing was carried out with Dell PowerEdge 2900 servers containing 24 GB of memory. Configuring this amount of memory in an x86-32 environment was possibly the most challenging aspect of the entire configuration. However, failure to configure the memory properly results in extremely poor performance. Therefore, this is an area where the DBA and system administrator should focus their attention. As stated in the *EMC Solutions for Oracle Database 10g/11g for Midsize Enterprises EMC Celerra Unified Storage Platform - Applied Technology Guide*, you must configure a shared memory file system to contain the SGA shared memory structures, in order to exceed the 32-bit memory mapping limitation, which is effectively 4,294,967,295 bytes (4 GB minus 1). The vast majority of Oracle database servers could benefit greatly from an SGA that is larger than that allowed by a maximum addressable memory space of 4 GB. See the Applied Technology Guide for detailed instructions on how to accomplish this.

Linux setup and configuration

Kickstart provides a way for users to automate a Red Hat Enterprise Linux installation. This is particularly critical in RAC environments where the OS configuration should be identical, and the required packages are more specific. Using kickstart, a single file can be created containing the answers to all the questions that would normally be asked during a Linux installation. These files can be kept on a single server system and read by individual database servers during the installation, thus creating a consistent, repeatable Linux install.

Kickstart provides a way for users to automate a Red Hat Enterprise Linux installation. This is particularly critical in RAC environments where the OS configuration should be identical, and the required packages are more specific. Refer to the Oracle documentation [Installing Oracle Database 11g Release 2 on Linux x86](#) for detailed instructions and recommendations on this.

The steps for kickstart installation are as follows:

1. Create a kickstart file.
2. Create a boot media with the kickstart file or make the kickstart file available on the network.
3. Make the installation tree available.
4. Start the kickstart installation.

“[Appendix A: Sample ks.cfg](#)” provides a sample ks.cfg file that you can use. This file was used in EMC's testing. For a clean, trouble-free Oracle Clusterware install, use these packages, and no others, for an Oracle RAC 10g installation.

The only other major issue we encountered concerned the package libaio. Our platform was EM64T. On this platform, both the 32-bit and 64-bit versions of libaio are required. In order to install this rpm successfully on this platform, the following procedure is required (assuming the current working directory contains both the 32-bit and 64-bit versions of this rpm):

```
rpm -e --nodeps libaio
rpm -Uvh libaio*rpm
```

Database setup and configuration

Initialization parameters

To configure the Oracle instance for optimal performance on the CLARiiON CX4 Series, we recommend the initialization options in [Table 9](#). These options are in the spfile or init.ora file for the Oracle instance.

Table 9. Initialization options

Parameter	Syntax
Description	
Database block size	DB_BLOCK_SIZE=n
For best database performance, DB_BLOCK_SIZE should be a multiple of the OS block size. For example, if the Linux page size is 4096, DB_BLOCK_SIZE =4096 *n.	
Direct I/O	FILESYSTEM_IO_OPTIONS=setall
This setting enables direct I/O and async I/O. Direct I/O is a feature available in modern file systems that delivers data directly to the application without caching in the file system buffer cache. Direct I/O preserves file system semantics and reduces the CPU overhead by decreasing the kernel code path execution. I/O requests are directly passed to network stack, bypassing some code layers. Direct I/O is a very beneficial feature to Oracle's log writer, both in terms of throughput and latency. Async I/O is beneficial for datafile I/O.	
Multiple database writer processes	DB_WRITER_PROCESSES=2*n
The recommended value for db_writer_processes is to at least match the number of CPUs. But during this test, we observed very good performance by just setting db_writer_processes to 1.	
Multi Block Read Count	DB_FILE_MULTIBLOCK_READ_COUNT= n
DB_FILE_MULTIBLOCK_READ_COUNT determines the maximum number of database blocks read in one I/O during a full table scan. The number of database bytes read is calculated by multiplying the DB_BLOCK_SIZE and DB_FILE_MULTIBLOCK_READ_COUNT. The setting of this parameter can reduce the number of I/O calls required for a full table scan, thus improving performance. Increasing this value may improve performance for databases that perform many full table scans, but degrade performance for OLTP databases where full table scans are seldom (if ever) performed. Setting this value to a multiple of the NFS READ/WRITE size specified in the mount limits the amount of fragmentation that occurs in the I/O subsystem. This parameter is specified in DB Blocks and NFS settings are in bytes, so adjust as required. EMC recommends that DB_FILE_MULTIBLOCK_READ_COUNT be set between 1 and 4 for an OLTP database and between 16 and 32 for DSS.	
Disk Async I/O	DISK_ASYNC_IO=true
RHEL 4 update 3 and later support async I/O with direct I/O on NFS. Async I/O is now recommended on all the storage protocols.	
Use Indirect Memory Buffers	USE_INDIRECT_DATA_BUFFERS=true
Required to support the use of the /dev/shm in-memory file system for storing the SGA shared memory structures.	
Database Block Buffers	DB_BLOCK_BUFFERS=<a number appropriate to your server>
Using the 32-bit version of Oracle RAC 11g (which is the only version available as of the publication of this white paper), you must manually set the size of your database buffer cache. The automatic memory management features of Oracle did not produce acceptable performance in our testing.	
Shared Pool Size	SHARED_POOL_SIZE=2800M
Using the 32-bit version of Oracle RAC 11g (which is the only version available as of the publication of this white paper), you must manually set the size of your shared pool. The automatic memory management features of Oracle did not produce acceptable performance in our testing. In addition, in our testing, setting the shared pool size to greater than 2.8 GB resulted in an out-of-memory error. This appears to be a limitation of the 32-bit version of Oracle RAC 11g.	

Control files and log files

Control files

EMC recommends that when you create the control file, allow for growth by setting MAXINSTANCES, MAXDATAFILES, MAXLOGFILES, and MAXLOGMEMBERS to high values.

EMC recommends that your database has a minimum of two control files located on separate physical ASM diskgroups. One way to multiplex your control files is to store a control file copy on every diskgroup that stores members of the redo log groups.

Online and archived redo log files

EMC recommends that you run a mission-critical, production database in ARCHIVELOG mode. EMC also recommends that you multiplex your redo log files for these databases. Loss of online redo log files could result in a database recovery failure. The best practice to multiplex your online redo log files is to place members of a redo log group on different ASM diskgroups. To understand how redo log and archive log files can be placed, refer to [Figure 3](#).

Oracle Direct NFS client

Direct NFS (DNFS), a new feature introduced in Oracle RAC 11g, was validated for the pure NFS solution. DNFS integrates the NFS client directly inside the database kernel instead of the operating system kernel. As part of the pure-NFS solution, the storage elements for Oracle RAC 11g were accessed using the DNFS protocol.

Implementing DNFS: Mandatory fixes and patches

Oracle 11.1.0.7 patch

Oracle 11g R1 has a known bug with regard to DNFS resiliency. The bug is resolved in the 11.1.0.7 patch. Do not implement Direct NFS unless the appropriate Oracle 11.1.0.7 patch has been installed and configured.

WARNING

If the appropriate 11.1.07 patch is not applied, it can have serious implications for the stability and continuity of a running database when configured to use DNFS.

See Oracle Metalink for more information on downloading and installing the Oracle 11.1.0.7 patch.

Celerra Data Mover: Enabling transChecksum

EMC recommends that you enable transChecksum on the Data Mover that serves the Oracle DNFS clients. This avoids the likelihood of tcp Port and XID (transaction identifier) reuse by two or more databases running on the same physical server, which could possibly cause data corruption.

To enable the transChecksum, type:

```
#server_param <movername> -facility nfs -modify transChecksum -value 1
```

Note: This applies to NFS version 3 only. Refer to the *NAS Support Matrix* available on Powerlink to understand the Celerra version that supports this parameter.

DNFS network setup

Port bonding and load balancing are managed by the Oracle DNFS client in the database; therefore there are no additional network setup steps.

If OS NIC/connection bonding is already configured, you should reconfigure the OS to release the connections so that they operate as independent ports. DNFS will then manage the bonding, high availability, and load balancing for the connections.

Dontroute specifies that outgoing messages should not be routed using the operating system, but sent using the IP address that they are bound to. If dontroute is not specified, it is mandatory that all paths to the Celerra are configured in separate network subnets.

The network setup can now be managed by an Oracle DBA, through the oranfstab file. This frees up the database sysdba from specific bonding tasks previously necessary for OS LACP-type bonding, for example, the creation separate subnets.

Database tuning for Oracle 11g DNFS

Oracle 11g DNFS requires little additional tuning, other than the tuning considerations necessary in any IP Storage environment with Oracle. In an unchanging environment, once tuned, DNFS requires no ongoing maintenance.

Examples of Celerra/Database tuning for Oracle 11g DNFS are described below.

Reserved port configuration

Some NFS file servers require NFS clients to connect using reserved ports. If your file server is running with reserved port checking, then you must disable it for DNFS to operate.

Mounting DNFS

If you use DNFS, then you must create a new configuration file, oranfstab, to specify the options/attributes/parameters that enable Oracle Database to use DNFS. These include:

- Add oranfstab to the ORACLE_BASE\ORACLE_HOME\db directory
- Oracle RAC: replicate the oranfstab file on all nodes and keep synchronized

Mounting multiple servers

If you use DNFS, then you must create a new configuration file. This specifies the options, attributes and parameters that enable Oracle Database to use DNFS. “Mounting DNFS” above provides additional details.

When oranfstab is placed in the ORACLE_BASE\ORACLE_HOME\db directory, the entries in this file are specific to a single database. The DNFS Client searches for the mount point entries as they appear in oranfstab. DNFS uses the first matched entry as the mount point.

Degree of Parallelism (DOP)

BI/DSS and data warehousing workloads with complex query generation involving outer table joins or full database table scans can be optimized on DNFS by configuring the degree of parallelism used by the database in the int.ora file. DOP is set to eight by default for a vanilla database install. Validation testing for DSS workloads with DNFS concluded that DOP set to 32 was optimum for the TPC-H like workloads applied to the servers during this testing.

Backup and recovery: Pure NFS solution

The best practice for the backup of Oracle Database 10g/11g is to perform approximately six logical storage backups per day, at four-hour intervals, using Celerra SnapSure™ checkpoints. The Celerra checkpoint command (fs_ckpt) allows a database administrator to capture an image of the entire file system as of a point in time. This image takes up very little space and can be created very rapidly. It is thus referred to as a logical image. Creating an Oracle backup using a logical image is referred to as a logical storage backup. This is to distinguish this operation from creating a backup using a copy to a different physical media, which is referred to as a physical backup.

To facilitate the ability to recover smaller granularities than the datafile (a single block for example), you should catalog all the SnapSure checkpoint backups within the RMAN catalog. In addition, as logical backups do not protect you from hardware failures (such as double-disk failures), you should also perform one physical backup per day, typically during a period of low user activity. For this purpose, EMC recommends RMAN using an incremental strategy, if the database is larger than 500 GB, and using a full strategy otherwise. Refer to the *EMC Solutions for Oracle Database 10g/11g for Midsize Enterprises EMC Celerra Unified Storage Platform - Applied Technology Guide* for more information. Further, EMC recommends that the RMAN backup be to a SATA II disk configuration rather than to tape.

Mean time to recovery is optimized by this approach. In the event of a fault that is not related to the hardware, you can restore instantly from a SnapSure checkpoint (according to Oracle, approximately 90 percent of all restore/recovery events are not related to hardware failures, but rather to user errors such as deleting a datafile or truncating a table). Further, the improved frequency of backups over what can be achieved with a pure physical backup strategy means that you have fewer logs to apply, thereby improving mean time to recovery. Even in the case where you need to restore from physical backup, the use of SATA II disk will improve restore time.

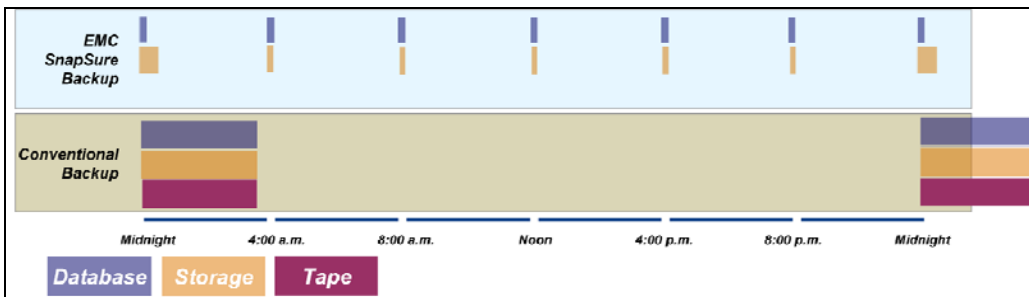


Figure 10. Multiple restore points using EMC SnapSure

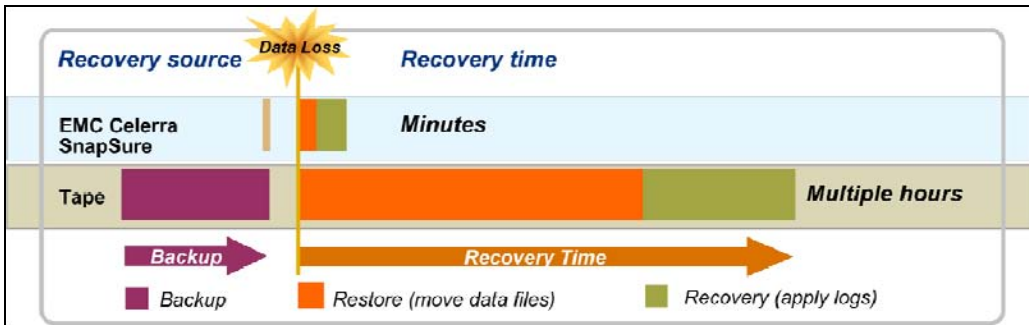


Figure 11. Rapid restore and recovery

Comparison: Logical storage backup and Flashback Database

With Oracle Database 10g/11g, Oracle introduced the FLASHBACK DATABASE command. In some respects it is similar to a logical backup. Both features provide you with the ability to revert the database to a point in time. Thus, both features allow you to undo certain user errors that affect the database. However, Flashback Database has certain limitations:

- A separate set of logs is required, increasing I/O at the database layer. The SnapSure checkpoints require some I/O as well, but this is at the storage layer, significantly lower in the stack than the database. In general, SnapSure checkpoints are lighter in weight than the flashback logs.
- The amount of time required to restore a database to a point in time using Flashback Database will be longer than that using Celerra SnapSure checkpoint restore. However, SnapSure checkpoints require you to apply archive logs, and Flashback Database does not. Thus, the mean time to recovery may vary between the two features. For Flashback Database, the mean time to recovery will be strictly

proportional to the amount of time you are discarding. In the case of Celerra SnapSure, the number of archived redo logs that must be applied is the major factor. Because of this, the frequency of logical backup largely determines the mean time to recovery.

- Flashback Database does not protect you from all logical errors. For example, deleting a file or directory in the file system cannot be recovered by Flashback Database but can be recovered using Celerra SnapSure checkpoints. Only errors or corruptions created within the database can be corrected using Flashback Database.

Evaluate both technologies carefully. Many customers choose to use both.

Data protection

As shown in Figure 12, the best practice for disaster recovery of an Oracle Database 10g/11g over NFS is to use the Celerra fs_copy for seeding the disaster recovery copy of the production database, and then to use the Oracle Data Guard log transport and log apply services. The source of the database used for seeding the disaster recovery site can be a hot backup of the production database within a Celerra SnapSure checkpoint. This avoids any downtime on the production server relative to seeding the disaster recovery database. The steps for creating this configuration are contained in the *EMC Solutions for Oracle Database 10g/11g for Midsize Enterprises EMC Celerra Unified Storage Platform - Applied Technology Guide*.

For best practices on Oracle Data Guard configuration, refer to the Oracle documentation on this subject.

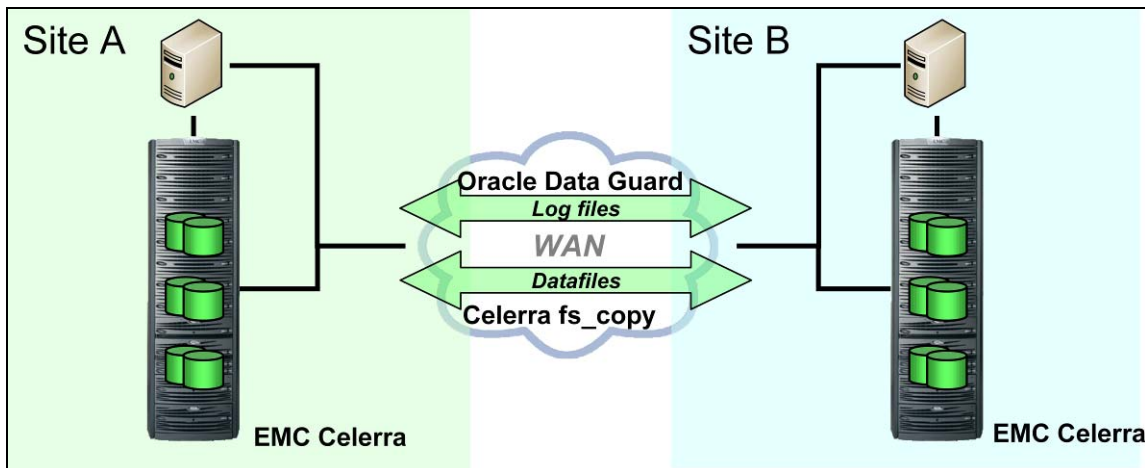


Figure 12. Remote disaster recovery for business protection

Database cloning

The ability to clone a running production Oracle database is a key requirement for many customers. The creation of test and dev databases, enabling of datamart and data warehouse staging, and Oracle and OS version migration are just a few applications of this important functionality.

EMC provides online, zero-downtime cloning of Oracle databases using the Celerra fs_copy feature. The best practice for creating a writeable clone of a production Oracle Database 10g/11g over NFS is to take a hot backup of the database using the SnapSure checkpoint, and then copy that hot backup to another location (possibly within the same Celerra array) using Celerra fs_copy. At that point, you can run a recovery against the hot backup copy to bring it to a consistent state.

Two methods can be used for database cloning. The full clone, involving a full copy of the entire database, is recommended for small databases or for a one-time cloning process. The alternative is incremental cloning. Incremental cloning is more complex but allows you to create a clone, making a full copy on the first iteration, and thereafter making an incremental clone for all other iterations, by copying only the changed data in order to update the clone. This is recommended for larger databases as well as for an ongoing or continuous need to clone the production database. Refer to the *EMC Solutions for Oracle*

Database 10g/11g for Midsize Enterprises EMC Celerra Unified Storage Platform - Applied Technology Guide for detailed steps on both of these methods.

Backup and recovery: Blended FCP/NFS solution

EMC offers you two choices for backup and recovery of an Oracle RAC 10g/11g stored on an EMC CLARiiON CX4 series. Refer to the sections “[Logical storage backup](#)” and “[Physical storage backup](#)” for more information.

Logical storage backup

The best practice for backup of Oracle RAC 10g/11g stored on a Celerra is to perform approximately six logical storage backups per day, at four-hour intervals, using CLARiiON SnapView. Navisphere® SnapView allows a database administrator to capture an image of an entire LUN, or using consistency technology, a group of LUNs, as of a point in time. This image takes up very little space and can be created very rapidly. It is thus referred to as a logical image. Creating an Oracle backup using a logical image is referred to as a logical storage backup. This is to distinguish this operation from creating a backup using a copy to a different physical media, which is referred to as a physical backup.

Mean time to recovery is optimized by this approach. You have the ability to restore instantly from a SnapView snapshot in the event that the event causing the fault has nothing to do with the hardware (according to Oracle, approximately 90 percent of all restore/recovery events are not related to hardware failures, but rather to user errors such as deleting a datafile or truncating a table). Further, the improved frequency of backups over what can be achieved with a pure physical backup strategy means that you have fewer logs to apply, thereby improving mean time to recovery. Even in the case where you need to restore from physical backup, the use of SATA II disk will improve restore time.

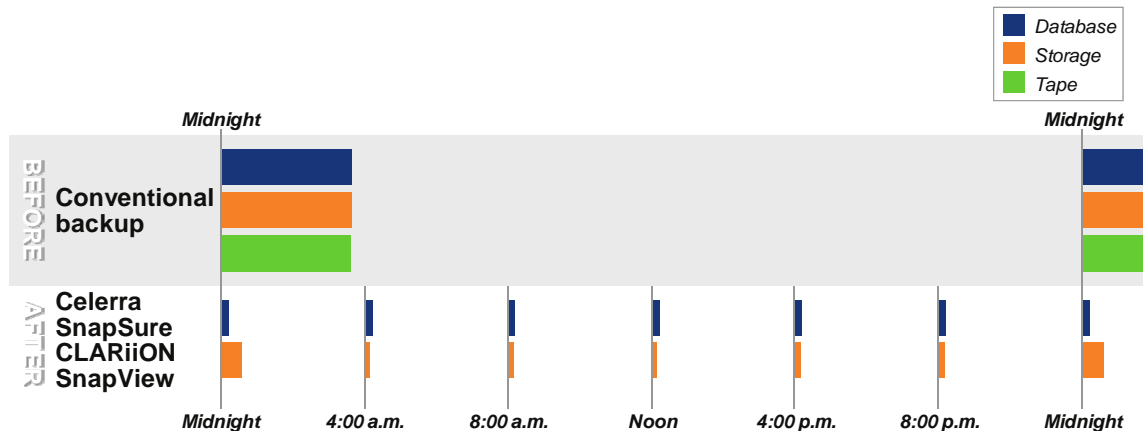


Figure 13. Multiple restore points using Celerra and SnapView

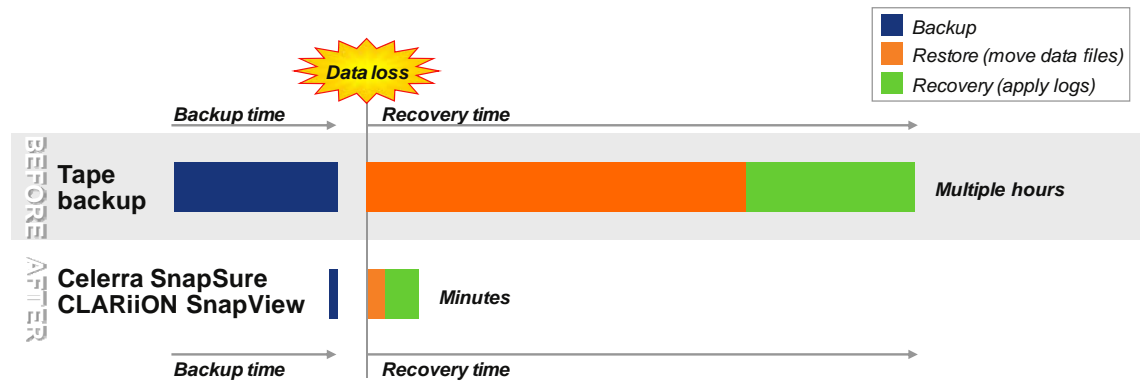


Figure 14. Rapid restore and recovery

Physical storage backup

The best practice for physical backup of Oracle Database 10g/11g databases is to use RMAN to back up to SATA or LCFS disk using ASM diskgroups. For more information refer to the *EMC Solutions for Oracle Database 10g/11g for Midsized Enterprises EMC Celerra Unified Storage Platform - Applied Technology Guide*.

Comparison: Logical storage backup and Flashback Database

With Oracle Database 10g/11g, Oracle introduced the FLASHBACK DATABASE command. In some respects it is similar to a logical backup. Both features provide you with the ability to revert the database to a point in time. Thus, both features allow you to undo certain user errors that affect the database. However, Flashback Database has certain limitations:

- A separate set of logs is required, increasing I/O at the database layer. SnapView snapshots require some I/O as well, but this is at the storage layer, significantly lower in the stack than the database. In general, SnapView snapshots are lighter in weight than the flashback logs.
- The amount of time required to restore a database to a point in time using Flashback Database will be longer than that using Celerra with SnapView rollback. However, using Celerra with SnapView rollbacks requires you to apply archive logs, and Flashback Database does not. Thus, the mean time to recovery may vary between the two features. For Flashback Database, the mean time to recovery will be strictly proportional to the amount of time you are discarding. In the case of using Celerra with SnapView rollback, the number of archived redo logs that must be applied is the major factor. Because of this, the frequency of logical backup largely determines the mean time to recovery.
- Flashback Database does not protect you from all logical errors. For example, deleting a file or directory in the file system cannot be recovered by Flashback Database but can be recovered using Celerra with SnapView rollback. Only errors or corruptions created within the database can be corrected using Flashback Database.

Evaluate both technologies carefully. Many customers choose to use both.

CX4 cache configuration for SnapView snapshot

Poor performance was observed using SnapView snapshot with the default settings. This was in the context of a scale-up OLTP workload using a TPC-C-like benchmark. If you have a similar workload and wish to use SnapView snapshot on a CX4 CLARiiON array, you will experience better performance by setting your cache settings as described in Table 10.

Table 10. CX4: Recommended cache settings

Cache setting	Value
Low watermark	10 percent
High watermark	30 percent
SP A read cache memory	200 MB
SP B read cache memory	200 MB
Write cache memory	1061 MB

Advanced backup and recovery using de-duplication (pure NFS and blended FCP/NFS solutions)

We validated the backup/restore of Oracle database using a Quantum DXi5500 disk-based backup appliance. Our results proved that Quantum offers huge benefits with respect to increasing the amount of backup data that can be retained on the same amount of disks and also with faster backup/restore operations.

We performed four full backups on both the NFS and the Quantum share to validate the benefits. The total size used on the NFS share for full backups was approximately 858 GB in the case of NFS and the same four full backups consumed only 435 GB on the Quantum appliance. After four full backups, we also performed a restore of the database from both the NFS and the Quantum share. The restore operation took 4 hours and 46 minutes on the NFS share whereas it took only 3 hours and 36 minutes on the Quantum share.

The other features of Quantum appliances are:

- Backup performance is comparable to a normal NFS server as a backup target.
- There is no impact on the production database while taking a backup.
- Saves a huge amount of disk space by using de-duplication technology. We saw about two to one compression on space during our testing.
- It is very easy to configure and manage the Quantum storage.

Managing and monitoring Celerra

Celerra Manager

The Celerra Manager is a web-based graphical user interface (GUI) for remote administration of a Celerra Unified Storage Platform. Various tools within Celerra Manager provide the ability to monitor the Celerra. These tools are available to highlight potential problems that have occurred or could occur in the future. Some of these tools are delivered with the basic version of Celerra Manager, while more detailed monitoring capabilities are delivered in the advanced version.

Enterprise Grid Control storage monitoring plug-in

EMC recommends use of the Oracle Enterprise Manager monitoring plug-in for the EMC Celerra Unified Storage Platform. Use of this system monitoring plug-in offers the following benefits:

- Realize immediate value through out-of-box availability and performance monitoring

-
- Realize lower costs through knowledge: know what you have and what has changed
 - Centralize all of the monitoring information in a single console
 - Enhance service modeling and perform comprehensive root cause analysis

The plug-in for an EMC Celerra server is available on the Oracle Technology Network.

Test/dev using Celerra SnapSure writeable snapshots

CSS disktimeout for Test/dev solution

The Cluster Synchronization Services (CSS) component of Oracle Clusterware maintains a heartbeat parameter called “disktimeout.” This parameter guarantees the amount of time that RAC nodes will not evict when there is no active I/O at the back-end storage.

The validation of the Test/dev solution using writeable checkpoints for pure NFS found that the disktimeout parameter needed to be set to a value of at least 900, so that the test/dev operation could be performed successfully without impacting the production database. Setting the disktimeout parameter to a higher value does not have any performance impact.

To configure a value for the disktimeout parameter, type the following command:

```
$ORA_CRS_HOME/bin/crsctl set css disktimeout 900
```

RAC-to-RAC cloning

The Oracle Clusterware and database software should be installed at the same location at both the production site and the clone target site. The following paths should be identical on both the source and the clone target site:

```
ORA_CRS_HOME=/u01/crs/oracle/product/10/crs  
ORACLE_HOME=/u01/app/oracle/oracle/product/10.2.0/db_1
```

The Oracle Cluster Registry (OCR) file and the voting disks for the source and clone target sites should be placed on separate, independent file systems.

The kernel parameters, memory settings, and the database directory structure should be identical on both the source and clone target sites.

Advanced protect using RecoverPoint

Journal volumes

If there are multiple consistency groups, then configure journal volumes for each consistency group on different RAID groups so that the journal of one group will not slow down the other groups.

Configure journal volumes on separate RAID groups from the user volumes.

Journal volumes can be corrupted if any host writes to it other than the RecoverPoint Appliance (RPA). So, ensure that the journal volumes are zoned only with RPAs.

RecoverPoint performs striping on journal volumes; using a large number from different RAID groups increases the performance.

The size of the journal volumes should be at least 20 percent that of the data being replicated.

Journal volumes are required on both local and remote sides for Continuous Remote Replication (CRR) to support failover.

All journal volume LUNs should be of the same size because RecoverPoint uses the smallest LUN size, and it stripes the snapshot across all LUNs. It will not be able to stripe evenly across different sized LUNs.

Repository volumes

Repository volumes should be at least 4 GB and an additional 2 GB per consistency group.

In a CRR configuration, there must be one repository volume for the RPA cluster on the local and remote site.

WAN compression guidelines

Setting a strong compression will cause CPU congestion and also when it is set to low, it will cause high loads.

EMC recommends a 5x-10x compression ratio for Oracle databases.

Clusters

The RecoverPoint clustering does not support one-to-many RPAs between sites. The configuration should have two RPAs on both sites for every site.

Zoning

When discovering the CX array splitters, make sure that all the RPA ports and all the CX SPA/SPB ports are included in the same zone. You must ensure that this zone is present on both sites. This can be accomplished by using a zone that spans multiple FCP switches if required.

Virtualization using VMware ESX

LUN discovery

LUNs can be discovered on an ESX server in two ways. The first common method is Virtual Machine File System (VMFS) and the other one is Raw Device Mapping (RDM). We recommend using RDM for discovering the LUNs on the ESX server as RDM provides better disk I/O performance and also supports VMotion.

VMotion storage requirements

You must have a common storage network configured on both source and target ESX servers to perform VMotion. Even the network configuration including the vSwitch names should be exactly the same and the connectivity to the LUNs on the back-end storage from the ESX servers also should be established in the same way. Figure 15 shows a typical configuration:

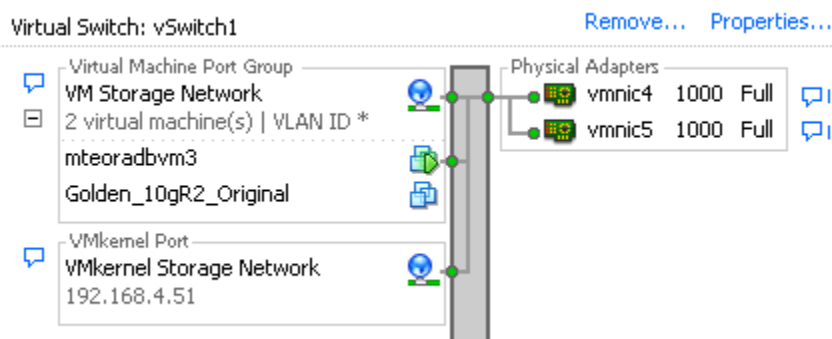


Figure 15. Storage network configuration





Storage					Refresh
Identification	Device	Capacity	Free	Type	
 results	10.6.115.168:/or...	12.39 TB	9.80 TB	NFS	
 ip-dart-qa (Reado...	128.222.1.24:/pd...	2.61 TB	26.89 GB	NFS	
 vm	192.168.4.101:/...	480.83 GB	151.35 GB	NFS	
 mteoradb51:stora...	vmhba0:0:0:3	128.50 GB	127.95 GB	vmfs3	

Figure 16. Storage configuration

In this case, the VMkernel Storage Network is being used to store the files for the VMs (through NFS). The storage pane in Figure 17 shows that the NFS-mounted volume “vm” is where these files are stored.

All ESX servers must have an identical configuration, other than the IP address for the VMkernel port, of course.

If RDM LUNs are mounted on a VM that must be migrated via NFS, then the VMDK pointer files must be stored on a shared VMFS file system using shared SAN (either FCP or iSCSI) storage. These files cannot be stored on an NFS mount point.

NFS connectivity requirements

When NFS connectivity is used, it is a best practice to have a dedicated private connection to the back-end storage from each of the VMs. We assigned four NICs (one NIC for each VM) on the ESX server, assigned private IPs to the same, and set up the connectivity from these four NICs to the Data Movers of the back-end storage using a Dell PowerConnect switch.

In addition, if you are running an Oracle database with NFS storage, the NFS mount points should be mounted directly onto the VM using `/etc/fstab` and normal Linux semantics. Do not mount the NFS mount points on the ESX server and then configure them as VMFS file systems within the VM. Vastly superior performance can be achieved through direct NFS mounts on the VM, as opposed to on the ESX server.

NFS volume recommendations

When configuring NFS storage volumes for VMs, avoid using the same NFS volume for multiple VMs. This is especially true for the datafile volume - the volume that is typically replicated when using snapshots and clones.

The use of the same NFS volume for multiple VMs will have the following negative impacts:

1. A snap or clone of that volume for the purposes of backing up one VM’s database will contain extraneous data in the form of data from the other VMs. This is wasteful of storage resources.
2. If you perform a rapid restore (using the snapshot functionality, for example) of a shared volume, you will wipe out more than one VM’s database data. This will limit flexibility greatly.

You can use the same user-defined pool (`data_pool` in the case of our standard configuration) for all of these volumes. Doing so allows all of the volumes to share the same disk resources, thereby leveraging these disks most effectively.

You should not create a very small volume on a small number of disks, as this would also be wasteful of resources, and will limit performance greatly.

Figure 17 shows the configuration that we tested with VMware HA cluster and which was found to work. The configuration shown in Figure 17 assumes the LUN/RAID group layout described in Figure 6.

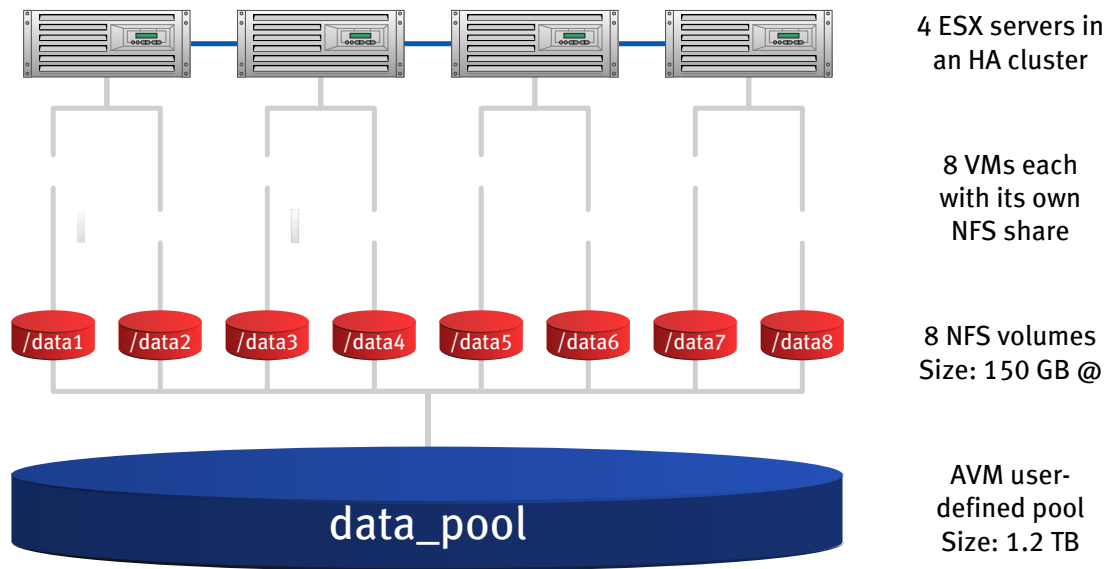


Figure 17. NFS volumes: Tested configuration

Performance comparison: Pure NFS versus blended FCP/NFS

We validated both the pure NFS solution and the blended FCP/NFS solution using VMware. We observed significantly better performance on the pure NFS solution compared to the blended solution. In the case of the pure NFS solution, there is no overhead introduced as file systems are directly mapped on VMs using NFS. The I/O is directly between the VMs and the storage in this case.

But in the case of the blended configuration, the LUNs are discovered on VMs using RDM and I/O from the storage must be routed via the ESX server. There is also a little delay in the I/O due to additional applications, PowerPath and ASM in the case of the blended solution.

As a result, for virtualization, we recommend the pure NFS solution for most purposes. However, please note the recommendation in the “NFS connectivity requirements” section on the manner in which the NFS file systems are mounted.

Storage configuration using virtualized solutions

A three-shelf RAID-10 disk layout is the recommended configuration for blended virtualized solutions. The performance on a RAID 10 configuration was better compared to RAID 5. Also, some of the solutions failed when using a RAID 5 configuration, for example “Advanced Backup using SnapView snapshot”.

VMware HA cluster

Compared with Oracle RAC

The use of VMware HA cluster for high availability has manageability and performance advantages over Oracle RAC. However, the VMware HA cluster solution does not work in every situation. There are two considerations to make when selecting the correct solution for your needs, these are discussed next.

Real-time high availability

VMware is an OS failover cluster. The VM that is running the Oracle database will restart on a surviving node of the VMware HA cluster in the event of a failure on the lost node; this will be seen by the Oracle database as downtime. The amount of time required for the failover is relatively brief — a few minutes. With Oracle RAC, on the other hand, a node failure will not result in any database downtime. There may be connectivity loss to client connections for the failed node, but even that can be protected against by use of Oracle products that virtualize the client connections to the cluster. No such protection is available for VMware HA cluster at this time.

Some Oracle databases require that there is absolutely no downtime. An excellent example would be a database that captures call records from a telephony switch. In that case, any downtime would result in the call records being lost, with a resulting loss of revenue to the customer. Another example would be an online securities trading application where loss of connectivity or downtime could cause legal issues for the customer due to a trade failing to complete in a timely manner. In either of these two scenarios, Oracle RAC is the correct solution as it provides absolute guaranteed uptime for the database.

The vast majority of databases do not fall into these categories, however. For databases where a brief period of downtime can be withstood in the event of hardware failure, the use of a VMware HA cluster should be preferred.

Scale-up compared to scale-out

Since a VMware VM cannot currently be a member of a RAC cluster, the number of CPUs and the amount of memory that a VM can use is limited. As a result, VMware is capable of handling a relatively small database, compared to the size of database that Oracle RAC can handle. This makes use of VMware HA clusters unsuitable in scale-up scenarios, but very appropriate in scale-out scenarios.

Scale-up scenario

In a scale-up scenario, every user of the database must be able to see every row in the database, and the database is typically quite large – in the range of terabytes or more. In this scenario, Oracle RAC is required (or a proprietary large CPU count SMP UNIX system). A typical example would be an online banking application of a large financial institution; decision makers need to be able to see the data of all customers of the enterprise. RAC is the correct solution in a scale-up scenario.

Scale-out scenario

A scale-out is a scenario similar to Software as a Service (SaaS). In this situation, a large number of similar databases are used to scale a piece of software that is being marketed to online customers. No single customer needs to be able to see the data of any other customer. As a result, the database workload can be split up and handled as a large number of small database instances. A VMware HA cluster is perfect for this scenario; not only does it handle the load very well, the cloning and replication capabilities of VMware can be used to make server provisioning very easy and convenient.

You should carefully map these considerations onto your customer's workload. Where VMware HA cluster can be implemented, very significant manageability and performance advantages can be achieved.

Name resolution

Each host in a VMware HA cluster must be able to resolve the name of any other host in the cluster. If name resolution fails, this will result in cluster instability.

The most reliable method of name resolution is a static hosts file. If DNS is used for name resolution, ensure that the DNS server configuration is reliable. For example, redundant DNS servers are recommended.

Virtualization using Oracle VM

By default, when a virtual machine is created using `virt-install` on an Oracle VM Server, the configuration file is created in the `/etc/xen` directory. When virtual machines are managed through Oracle VM Manager, it expects this configuration file to be located in `/OVS/running_pool/<vm_name>/vm.cfg`. It is recommended to move and rename the virtual machine configuration file from `/etc/xen/<vm_name>` to `/OVS/running_pool/<vm_name>/vm.cfg`.

Oracle VM documentation states that a virtual machine can also be created from Oracle VM Manager. We observed several issues such as NFS errors and bugs while creating OVMs from OVM Manager. Our recommendation is to create all OVM virtual machines through the command line interface using the `virt-install` command on the Oracle VM Server.

Hugepage settings have not been implemented in Oracle VM for either paravirtualized or fully virtualized OVMs. Hugepages have significant performance improvements in Oracle environments. VMware ESX provides hugepages support.

Replication Manager

The Test/dev and Advanced Backup solution components of the Oracle blended FCP/NFS solution are now integrated with EMC Replication Manager. This has significant advantages over the previous solutions in that Replication Manager provides a layered GUI application to manage these processes. This includes a scheduler so that the jobs can be run on a regular basis. Replication Manager, however, introduces a few issues that are covered in this section.

Oracle home location

Currently, Replication Manager does not support ASM and Oracle having separate Oracle homes. This may be confusing, because the Oracle installation guide presents an installation in which ASM is located in its own home directory. However, having the normal Oracle home and the ASM home in the same directory is supported. If you choose to use Replication Manager for storage replication management (which is highly recommended), install Oracle and ASM in the same home directory.

Dedicated server process

Replication Manager cannot create an application set when connected to the target database using SHARED SERVER. Replication Manager requires a dedicated server process. In the `TNSNAMES.ORA` file, modify the value of `SERVER` as shown below to connect to the target database (this is only needed for the service that is used for the Replication Manager connection):

```
# tnsnames.ora Network Configuration File:
/u01/app/oracle/product/10.2.0/db_1/network/admin/tnsnames.ora
# Generated by Oracle configuration tools.
MTERAC211 =
  (DESCRIPTION =
    (ADDRESS = (PROTOCOL = TCP)(HOST = mteoradb67-vip)(PORT = 1521))
    (CONNECT_DATA =
      (SERVER = DEDICATED)
      (SERVICE_NAME = mterac21)
      (INSTANCE_NAME = mterac211)
    )
  )
```

Conclusion

The EMC Celerra Unified Storage Platform's high-availability features combined with EMC's proven storage technologies provide a very attractive storage system for the Oracle RAC 10g/11g over FCP and NFS. Specifically:

- It simplifies database installation, backup, and recovery.
- It enables the Oracle RAC 10g/11g configuration by providing shared disk.
- The Data Mover failover capability provides uninterrupted database access.
- Redundant components on every level, such as the network connections, back-end storage connections, RAID, and power supplies, achieve a very high level of fault tolerance, thereby providing continuous storage access to the database.
- Celerra with SnapView provides rapidly available backups.
- The overall Celerra architecture and its connectivity to the back-end storage make it highly scalable, with the ease of increasing capacity by simply adding components for immediate usability.

Running the Oracle RAC 10g/11g with Celerra provides the best availability, scalability, manageability, and performance for your database applications.

References

The following documents, located on Powerlink.com, provide additional, relevant information. Access to these documents is based on your login credentials. If you do not have access to the following content, contact your EMC representative:

- *EMC Solutions for Oracle Database 10g/11g for Midsize Enterprises EMC Celerra Unified Storage Platform – Reference Architecture*
- *EMC Solutions for Oracle Database 10g/11g for Midsize Enterprises EMC Celerra Unified Storage Platform – Applied Technology Guide*
- *EMC Solutions for Oracle Database 10g/11g for Midsize Enterprises EMC Celerra Unified Storage Platform Physically Booted Blended FCP/NFS Solution: RecoverPoint with CX Splitters*
- CLARiON CX4 series documentation

The following resources have more information about Oracle:

- The [Oracle Technology Network](#)
- [Metalink](#), the Oracle support website

Appendix A: Sample ks.cfg

```
install
nfs --server=128.222.1.24 --dir=/pdd2/ip-dart-
qa/Solutions/software/Linux/Red_Hat_Enterprise_Linux/AS_4_update_3_-_AMD64-
IntelEM64T
lang en_US.UTF-8
langsupport --default=en_US.UTF-8 en_US.UTF-8
keyboard us
xconfig --card "ATI Radeon 7000" --videoram 8192 --hsync 31.5-37.9 --vsync 50-70 -
-resolution 800x600 --depth 16 --startxonboot --defaultdesktop gnome
network --device eth0 --bootproto dhcp
network --device eth1 --onboot no --bootproto none
network --device eth2 --onboot no --bootproto none
network --device eth3 --onboot no --bootproto none
network --device eth4 --onboot no --bootproto none
rootpw --iscrypted $1$rP2mLD4F$xqJrp/LiSMqOH8HVA1Xg4.
firewall --disabled
selinux --enforcing
authconfig --enablesshadow --enablemd5
timezone America/New_York
bootloader --location=mbr --append="rhgb quiet"
clearpart --all --drives=sda,sdb
part / --fstype ext3 --size=100 --grow --ondisk=sda --asprimary
part swap --size=16384 --ondisk=sdb --asprimary

%packages
@ compat-arch-development
@ admin-tools
@ editors
@ system-tools
@ text-internet
@ x-software-development
@ legacy-network-server
@ gnome-desktop
@ compat-arch-support
@ legacy-software-development
@ base-x
@ server-cfg
@ development-tools
@ graphical-internet
e2fsprogs
sysstat
kernel-smp-devel
kernel-devel
vnc
telnet-server
rdesktop
kernel-smp
tsclient

%post
```