EMC Solutions for Oracle Database 10*g*/11*g* for Midsize Enterprises

Virtualized Solutions

EMC Celerra NS40 Unified Storage Platform

Reference Architecture

EMC Global Solutions

42 South Street
Hopkinton MA 01748-9103
1.508.435.1000
www.EMC.com

Published March, 2009

EMC Solutions for Oracle Database 10*g*/11*g* for Midsize Enterprises Virtualized Solutions
EMC Celerra NS40 Unified Storage Platform Reference Architecture

P/N H6027.1

Contents

About this Document

Chapter 1 About this Solution

Chapter 2 Solution Architecture

Chapter 3 Environment Profile

Glossary

About this Document

This document provides an overview of the architecture of a number of solutions for midsize enterprises that use the Celerra Unified Storage Platform and Oracle Database 10*g* and 11*g* on Linux, in the context of virtualization using VMware ESX.

Purpose

Information in this document can be used as the basis for a solution build, white paper, best practices document, or training.

Information in this document 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

This document is intended for EMC internal personnel, EMC partners, and customers.

Scope

This document describes the architecture of a number of solutions built and tested by EMC at Research Triangle Park, North Carolina.

Implementation instructions and best practices are beyond the scope of this document.

Related documents

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

On EMC.com and EMC Powerlink:

- *EMC Solutions for Oracle Database 10g/11g for Midsize Enterprises Physically Booted Solutions EMC Celerra NS40 Unified Storage Platform Reference Architecture*
- *EMC Solutions for Oracle 10g/11g for Midsize Enterprises EMC Celerra Unified Storage Platform - Best Practices Planning*

- *EMC Solutions for Oracle Database 10g/11g for Midsize Enterprises EMC Celerra Unified Storage Platform - Applied Technology*
- *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 - Applied Technology*
- Celerra and Oracle Database presentations, white papers, and Validation Test Reports

On EMC.com:

- The Solutions for Oracle page

1
About this Solution

This chapter presents the following topics:

Business challenges

Midsize enterprises face the same challenges as their larger counterparts when it comes to managing database environments. These challenges include:

- Rising costs
- Control over resource utilization and scaling
- Lack of sufficient IT resources to deploy, manage, and maintain complex environments at the departmental level
- The need to reduce power, cooling, and space requirements

Unlike large enterprises, midsize enterprises are constrained by smaller budgets and cannot afford a custom, one-off solution. This makes the process of creating a database solution for midsize enterprises even more challenging than for large enterprises.

Technology solution

What is a solution?

Within this document, a solution is defined as a complete stack of hardware and software upon which a customer would choose to run their entire business.

A solution includes all the elements that are needed by the business, these are referred to in this document as “solution components.”

Solution components include:

- Backup and recovery
- Data protection
- Testing and development
- Resiliency
- Performance

Solution requirements

The business challenges that midsize enterprises face require flexible, scalable, cost-effective, and resilient storage solutions. To address these requirements, EMC has developed storage solutions that:

- Implement tiered storage
- Are based on open systems software
- Use cost-effective industry-standard hardware
- Include all of the required elements for a complete database solution
- Are fully documented so that the solution can be built, tuned, and managed successfully and reliably
- Are as dense and power efficient as possible
- Are flexible and manageable and provide significant ease-of-use advantages

EMC Celerra

The heart of the solutions in this document, the EMC® Celerra® Unified Storage Platform, is a remarkably versatile device. Celerra includes a world-class Network Attached Storage (NAS) array combined with the functionality and performance of the leading midrange Storage Area Network (SAN) array. Celerra provides both NAS and SAN functionality and performance without compromise.

Celerra provides:

- NAS through the Network File System (NFS) and Common Internet File System (CIFS) protocols
- iSCSI storage through the Celerra's Data Movers
- SAN storage over the Fibre Channel Protocol (FCP) through the back-end EMC CLARiiON® CX4-240 series storage array

Oracle Database 10*g*/11*g*

Market share

Oracle is the dominant enterprise-class database software product at this point in time. Analyst market share studies demonstrate that Oracle has the largest market share percentage of any of the players in this market. With good reason, as Oracle provides a very reliable, robust, and manageable product. EMC remains committed to creating solutions based upon the Oracle database software stack.

Oracle software stack

The goal of the solutions presented in this document is to provide comprehensive testing, validation, and documentation of complete environments that:

- Include the Oracle software stack
- Are enabled by EMC storage hardware and EMC value-added software

The Oracle software stack covered by the solutions consists of:

- Oracle Enterprise Linux
- Oracle Database
- Automatic Storage Management (ASM)
- Oracle 11*g* Direct NFS (DNFS)

Virtualized solutions

The term “virtualized solution” refers to the way the production database servers are booted. That is, the production database servers in these solutions are booted on virtualized hardware, using a hypervisor such as VMware ESX.

In each of the solutions described in this document, virtualization of all database servers is implemented through the VMware ESX software platform.

The use of a virtualized utility server (such as a test/dev clone server, or a disaster recovery target) is extremely common and is well-supported; it is included in the solutions described in this document.

The following virtualized solutions are described in this document:

Pure NFS solution

All of the storage elements are accessed using the NFS protocol.

Pure NFS with VMware High Availability cluster solution

This solution incorporates a four-node VMware High Availability (HA) cluster. All of the storage elements are accessed using the NFS protocol.

Direct NFS

DNFS, a new feature introduced in Oracle Database 11*g*, 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 Database 11*g* were accessed using the DNFS protocol.

Some of the benefits of DNFS include:

- Asynchronous direct I/O

- Reduced CPU and memory usage
- Consistent performance across operating systems
- Load-balancing and high availability (HA) are managed internally within the DNFS client itself

DNFS is discussed in more detail in "Solution benefits".

Note: This DNFS solution is based on validation of Oracle 11*g* Direct NFS. Oracle 11*g* Release 1 has two known bugs with regard to DNFS resiliency. Refer to "Implementing DNFS: Mandatory fixes and patches" for information on the available patches.

Blended FCP/NFS solution

The high-demand, low-latency storage elements are accessed using Fibre Channel Protocol (FCP) and Oracle ASM. All other storage elements are accessed using the NFS protocol. Table 1 outlines which storage elements are accessed using each protocol.

Table 1 Blended FCP/NFS: Storage elements

Accessed using FCP	Accessed using NFS
Datafiles	Flashback recovery area
Online redo log files	Archive logs
Control files	Disk-based backups
Temp files	

Solution components

A solution component is an element of a solution that provides a discrete validated function. Table 2 describes the solution components that have been validated for solutions described in this document.

Note: "Solution Architecture" provides precise details on the solution components that have been validated for each solution that is described in this document.

Table 2 Solution components

Solution component	Description
Scale-Out OLTP	• 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	• 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)	• 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)	• 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	• 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.

Solution component	Description
Advanced Protect	• Uses additional hardware or 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	• 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	• 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 benefits

Reduced total cost of ownership

CPU usage

In any reasonable configuration, the database server's CPU is the most precious component of the entire architecture. Therefore, the overarching principle of EMC's Oracle Database 10*g* and 11*g* virtualized solutions for midsize enterprises is to free up the database server's CPU (as well as memory and I/O channels) from utility operations such as backup and recovery, disaster recovery staging, test/dev, and cloning. The highest and best use of the database server’s CPUs is to parse and execute the SQL statements that are required by the application user.

Choice of hardware and software

The hardware and software components used in the solutions are standard, high-volume items that can be obtained at reasonable cost, and have abundant exposure in the marketplace. In this way, equipment acquisition and maintenance costs are reduced.

Total cost of ownership is reduced by using:

- Industry-standard hardware
- Open-systems operating system software
- Midrange multi-protocol storage arrays
- Specialized de-duplication storage arrays

- Virtualization of the database server both for production and for utility servers, such as the test/dev target and the disaster recovery target

Ease of use

Simplified implementation

The comprehensive testing program provided by EMC eases the burden of implementation, storage provisioning, and management. This is because the majority of issues that the customer would encounter while implementing the solution have already been thoroughly tested and documented.

Simplified management

DNFS

The use of DNFS simplifies network setup and management by eliminating administration tasks such as:

- Setting up network subnets
- LACP bonding
- Tuning of Linux NFS parameters

Simplified installation and configuration

Blended FCP/NFS architecture

The use of a blended FCP/NFS architecture provides very significant ease-of-use advantages. Provided that the customer has access to a Celerra Unified Storage Platform, the amount of software that the customer is required to install and configure on the database server is significantly reduced.

Normal file-system semantics

The NFS protocol provides access to storage elements such as archive logs and disk-based backups using normal file-system semantics. This is a significantly simpler form of management than that provided through FCP and ASM.

Robust performance and scaling

Performance testing

The performance testing provided by EMC is unique. It utilizes an industry-standard OLTP benchmark, but does so without exotic tunings that are not compliant with best practices. In addition, real-world configurations that the customer is likely to deploy are used. For example, RAID protection is included in all validated configurations. This allows the customer to be reasonably assured that the configuration they choose to run their application will do so predictably and reliably.

FCP and NFS

The use of FCP in the solutions (through the CLARiiON CX4-240 back-end array) allows for high-performance, low-latency I/O for databases that require high-performance storage. However, this is required only for database files that need a high level of performance (datafiles, temp files, control files, and online redo log files).

All remaining database files (archived logs, flashback recovery area, and backup target) can be stored on NFS mounts, thus improving manageability and ease of use. For databases that do not have high-performance storage demands, the NFS-mounted file systems can be used for all database files.

Both blended (combined FCP and NFS) and pure (NFS only) configurations are described in this document.

Direct NFS

The Direct NFS (DNFS) client performs concurrent I/O by bypassing the operating system. The benefits of this are:

- Better performance due to the reduction of memory consumption and CPU utilization
- Consistent NFS performance is observed across all operating systems.

DNFS is optimized for DB workloads and supports asynchronous I/O, which is suitable for most databases; it delivers optimized performance by automatically load balancing across the available paths. Load balancing in DNFS is frequently superior to the conventional Linux kernel NFS (kNFS).

Information lifecycle management

The solutions described in this document use tiered storage. High-performance storage is provided through the CLARiiON CX4-240 back end over FCP and data is stored on FC disks. This is used for files requiring extremely high-performance, low-latency access, such as Oracle datafiles, temp files, control files, and online redo log files.

Other files are stored over NFS on slower, less expensive SATA II disks. These include archived log files, flashback recovery area, and backup target. For databases requiring less performance, all database files can be stored on NFS-mounted file systems.

Virtualization

Advantages over physically booted

Virtualization provides significant advantages over physically booted Oracle in a variety of contexts:

- The use of virtualization for the production database server provides manageability and ease-of-use advantages.
- In a scale-out context, virtualization can provide superior performance and scalability compared to physically booted configurations - even when using hardware identical to that used in the physically booted configuration.

- Utility servers, such as a Test/dev target, Advanced Protect target, and Basic Protect target, are more easily and conveniently managed as virtual machines than as physically booted Oracle database servers. The advantages of consolidation, flexible migration and so on, which are the mainstays of virtualization, apply to these servers as well.

Due to the requirement for RAC qualification, presently there is no support for Oracle 10*g* and 11*g* RAC servers on virtualized devices. For this reason, EMC does not publish such a configuration as a supported and validated solution. However, the use of Oracle Database 10*g* and 11*g* (in single-instance mode) presents far fewer support issues.

VMotion

In addition, the use of VMotion to migrate an Oracle database instance from one piece of hardware to another was achieved in our tests with virtually no performance impact and no downtime on the running Oracle database instance. This provides a very high level of manageability and downtime reduction for tasks such as software and hardware upgrades.

High Availability

The use of a VMware HA cluster provides further high availability advantages to the virtualized solution. EMC has validated the use of a VMware HA cluster with Oracle Database 10*g* and 11*g* in a single-instance scale-out environment. In addition, a VMware HA cluster can be used to automate failover in the event of a hardware or software failure.

Reduced costs

One of the main challenges faced by the customer is to reduce cost by utilizing infrastructure effectively. Virtualization enables reduction in the number of servers and related IT hardware in the data center. The other main feature is the ability to move a running VM production database from one physical sever to another physical server without any downtime.

High availability

The resiliency testing carried out by EMC ensures that the database configuration is reliable. Clustering is used at every major layer of the solution, including the database server, the NAS file server, and the back-end SAN array. By testing the fault tolerance of all of these layers, the ability of the application to withstand hardware failures with no downtime is assured.

Solution component advantages

Scale-Out OLTP

EMC's comprehensive performance-testing program provides assurance to the customer that a given database workload can be run on the configuration predictably and reliably. This is because the performance testing is conducted on real-world hardware with best-practices compliant configurations in terms of the storage, operating system, and database software layers.

Scale-Out OLTP is the use of a real-world online transaction processing (OLTP) workload with a federated database approach in which multiple separate database images, running on separate, non-clustered database servers, are used. This implies that the database workload can be easily partitioned such that not all database data needs to be seen by all customers of the application. A typical example would be software as a service (SaaS).

Basic Backup and Recovery

Basic Backup and Recovery is validated primarily to serve as a baseline for comparison with the Advanced Backup solution, which demonstrates the use of storage layer replication tools (snapshots for example) to perform Oracle database backup.

Advanced Backup and Recovery

The database server's CPU is the most valuable resource in a database configuration. The use of the database server's CPU, memory, and I/O channels to perform backup and recovery operations is therefore wasteful of a precious commodity. Advanced Backup and Recovery off-loads the database server's CPUs in many cases.

The advantages of Advanced Backup are as follows:

Reduced mean time to recovery

Advanced Backup and Recovery with snapshots dramatically improves the mean time to recovery (MTTR) by reducing the time required for the restore operation. Further, as the backup operation has minimal impact on the database server performance, the backup can be run more often. This means that the recovery operation is also optimized since fewer archived logs must be applied.

Reduced storage, power, and space costs

A specialized de-duplication array is also included in this solution. This array provides very dense storage for Oracle RMAN backup files. This reduces the costs of storage, power, and space, in that the same amount of data can be stored on fewer disks.

Basic Protect

Basic Protect, using Oracle Data Guard, is validated as a baseline for comparison with Advanced Protect.

Advanced Protect

Similar to Advanced Backup and Recovery, Advanced Protect frees up the database server's CPU, memory, and I/O channels from the impact of a disaster recovery configuration.

One form of Advanced Protect is provided:

- MirrorView™/A through iSCSI

Advanced Protect using MirrorView/A through iSCSI

The advantages of Advanced Protect using MirrorView/A through iSCSI are as follows:

Simplified deployment

The use of consistency technology allows the customer to deploy a disaster recovery solution with a more simplified Oracle Application Stack on the target side. Since Data Guard is not required on the target side, in many cases Oracle 10*g* or 11*g* Standard Edition can be used.

IP network

The use of IP as the network protocol reduces costs and complexity. No conversion of FCP to IP is required.

Resiliency

Resiliency testing is conducted to insert faults into various layers of the configuration. This includes inserting faults in:

- Network ports on the database server
- Network ports on the storage array
- Database server processor
- Storage array processor

This ensures that the configuration will continue to remain up and running in the event of a hardware failure. No single points of failure can be present in the configuration.

Test/dev

The ability to deploy a writeable copy of the production database is required by many customers. The process of provisioning this copy must create minimal, if any, performance impact on the production database server. Also, absolutely no downtime can be tolerated. The Test/dev solution documented here provides this.

The use of virtualized test/dev single-instance Oracle Database 10*g* or 11*g* servers has been fully qualified by EMC and is documented in this guide.

2

Solution Architecture

This chapter presents the following topics:

Overall architecture

Three virtualized solutions are described in this document:

1. Pure NFS
2. Pure NFS VMware HA cluster, including DNFS for Oracle 11*g*
3. Blended FCP/NFS

Storage protocol

In terms of the storage protocol, the solutions fall into two categories:

Pure NFS

All database objects are stored on an NFS mount.

Pure NFS VMware HA cluster

All database objects are stored on an NFS mount.

In the case of Oracle 11*g*, datafiles, temp files, control files, online redo log files and archive log files are accessed using the DNFS protocol.

Blended FCP/NFS

Datafiles, tempfiles, control files, and online redo log files are stored over FCP and Automatic Storage Management (ASM). All other database objects are stored over NFS.

Target servers

The solutions also include virtualized servers for Test/dev, Basic Protect, and Advanced Protect targets.

Virtualization configuration

Setting up a virtualization configuration involves:

1. Deploying a VMware ESX server or a VMware ESX HA cluster
2. Creating virtual machines (VMs)
3. Allocating memory and CPU resources
4. Configuring VMs
5. Discovering LUNs on the VMs (NFS mounts for the pure NFS solution, Raw Device Mapping (RDM) for FCP)
6. Configuring VMotion for high availability

Virtualized solutions

The following features are common to each of the three solutions described in this document.

Basic architecture

- Oracle Database 10*g* or 11*g* for x86-64 is run on Red Hat Enterprise Linux or Oracle Enterprise Linux on all VMs.
- Client virtual machines run on a VMware ESX server. They are connected to a client network.
- Client and TCP/IP storage networks consist of dedicated network switches and VLANs.
- The HA cluster interconnect and storage networks consist of trunked and virtualized IP connections to balance and distribute network I/O. Jumbo frames are enabled on these networks.
- A Celerra is connected to the VMware ESX servers through the storage network.
- Archived log files, the flashback recovery area, and the backup target are stored on a RAID-protected NFS file system. This NFS file system is stored on low-cost, high-capacity SATA II drives.

Pure NFS virtualized solution

Reference architecture diagram

Figure 1 describes the reference architecture for the pure NFS virtualized solution.

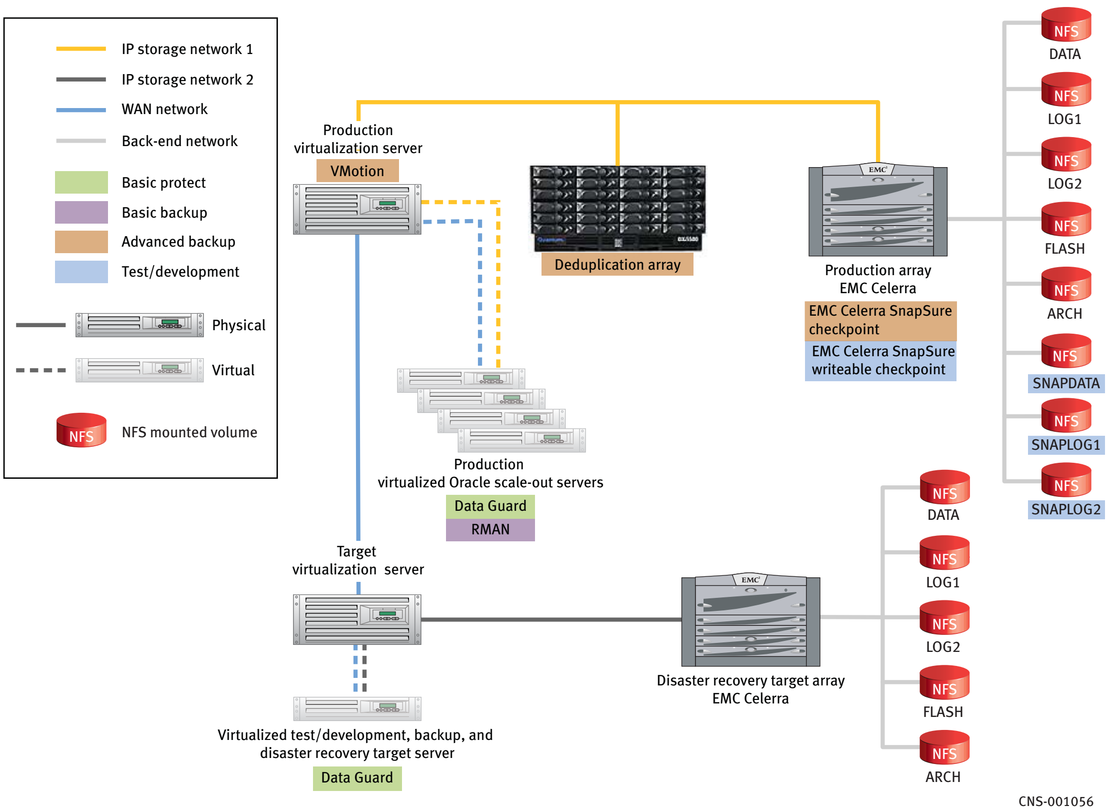

Figure 1 Pure NFS virtualized solution: Oracle single instance on Linux using Celerra

Solution features

In addition to the features described in "Virtualized solutions", the pure NFS solution includes the following elements:

Basic architecture

- One ESX production server is used for all database instances. A second ESX server is used as a VMotion target. For the purpose of high availability, VMotion is used to migrate a VM running the production database from one ESX server to another without any downtime. Both of the ESX servers are connected to both TCP/IP networks.
- TCP/IP and NFS version 3 provide network connectivity and file system semantics for NFS file systems.
- Four database server VMs are created on a single ESX server.
- Each VM is allocated 7 GB of memory and two CPUs.
- The client VMs are connected to the client network.

Storage network architecture

- Oracle database files and online redo log files reside on their own NFS file system. Online redo log files are mirrored across two different file systems using Oracle software multiplexing. Three NFS file systems are used, one file system for datafiles and temp files, and two file systems for online redo log files.
- Oracle controlfiles are mirrored across the online redo log file NFS file systems.
- RAID-protected NFS file systems are designed to satisfy the I/O demands of individual database objects. For example, RAID 5 is sometimes used for the datafiles and temp files, but RAID 1 is always used for the online redo log files (two separate configurations are supported; refer to "Pure NFS RAID configuration" and "Blended RAID configuration").
- The NFS file systems that are used to store the Oracle datafiles, temp files, online redo log files, and control files are stored on FC disks.

Target site

- At the target site, a separate Celerra is connected to the VMware ESX server through the target storage network. The Oracle Database 10*g*/11*g* single-instance target server accesses this network through a virtualized connection.

Validated solution components

Table 3 describes the solution components that have been validated for the pure NFS virtualized solution.

Table 3 Pure NFS: Validated solution components

Solution component	Details
Scale-Out OLTP	Utilizes an industry-standard OLTP database performance benchmark
Basic Backup and Recovery	Oracle Recovery Manager (RMAN)
Advanced Backup and Recovery	EMC Celerra SnapSure™ checkpoints
Basic Protect	RMAN Oracle Data Guard
Resiliency	Every significant layer of the solution has been tested by introducing faults
Test/dev	EMC Celerra SnapSure writeable checkpoints

Blended FCP/NFS virtualized solution

Reference architecture diagram

Figure 2 describes the reference architecture for the blended FCP/NFS virtualized solution.

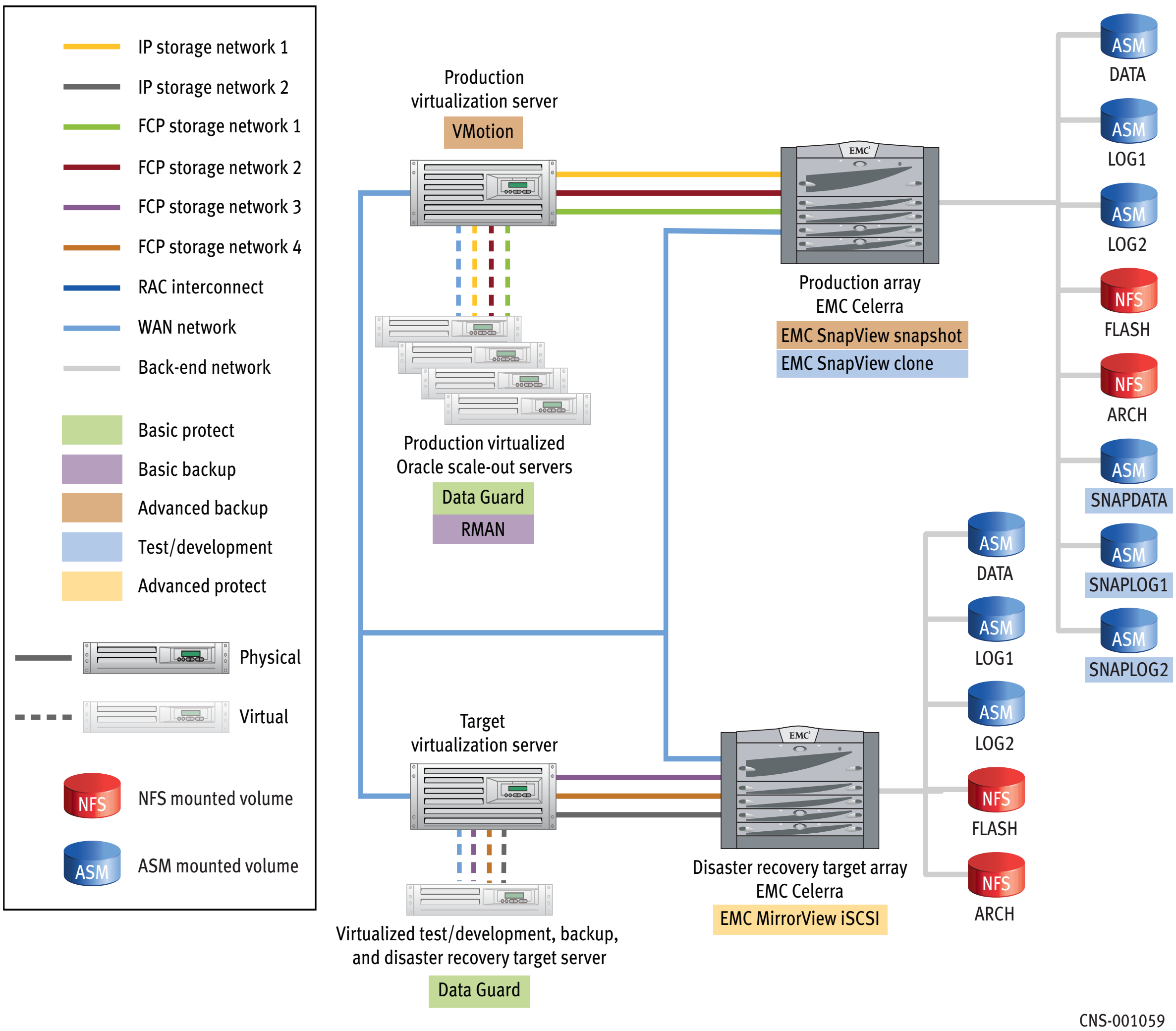

Figure 2 Blended FCP/NFS virtualized solution: Oracle single instance on Linux using Celerra

Solution features

In addition to the features described in "Virtualized solutions", the blended FCP/NFS solution includes the following additional elements:

Basic architecture

- The production and target Celerra Unified Storage Platforms are both connected to the WAN.
- FCP storage networks consisting of dedicated, redundant FCP switches are present at both sites.
- TCP/IP and NFS version 3 provide network connectivity and file system semantics for NFS file systems.
- At the production site, the virtualized Oracle Database 10*g*/11*g* servers are connected to the production FCP storage network using virtualized connections. The production Celerra (actually the CLARiiON CX4-240 back-end array) is also connected to this network.
- One ESX production server is used for all database instances. A second ESX server is used as a VMotion target. For the purpose of high availability, VMotion is used to migrate a VM running the production database from one ESX server to another without any downtime. Both of the ESX servers are connected to both TCP/IP networks.
- Four database server VMs are created on a single ESX server.
- Each VM is allocated 7 GB of memory and two CPUs.
- The client VMs are also connected to the client network.

Storage architecture

- High-performance database objects are accessed over an FCP network using redundant network switches. Oracle ASM is used as the file system and volume manager. ASMLib is also used to virtualize the LUNs on the database server.
- Oracle datafiles, temp files, and online redo log files are stored on separate LUNs that are mounted on the database server using ASM over FCP. Online redo log files are mirrored across two different ASM disk groups using Oracle software multiplexing. Three ASM disk groups are used, one for datafiles, and two for online redo log files.
- The control files are mirrored across the online redo log ASM disk groups.
- Each ASM disk group and its underlying LUNs are designed to satisfy the I/O demands of individual database objects. For example, RAID 5 is used for the datafiles, but RAID 1 is used for the log files. All of these disk groups are stored on FC disks.

- FCP storage networks consist of dedicated, redundant FCP switches. Four separate sets of LUNs (for data and redo logs) are created on the back-end storage and each corresponding set is presented to the VM using RDM.

Target site

- At the target site, the VMware ESX server is connected to the target FCP storage network. The Oracle Database 10*g*/11*g* single-instance target server accesses this network through a virtualized connection. The target Celerra (actually the CLARiiON CX4-240 back-end array) is also connected to this network.
- The LUNs are discovered using Raw Device Mapping (RDM) on the target VMs.

Validated solution components

Table 3 describes the solution components that have been validated for the blended FCP/NFS solution.

Table 4 Blended FCP/NFS solution: Validated solution components

Solution component	Details
Scale-Out OLTP	Utilizes an industry-standard OLTP database performance benchmark
Basic Backup and Recovery	Oracle Recovery Manager (RMAN)
Advanced Backup and Recovery	EMC CLARiiON SnapView™ snapshots
Basic Protect	Oracle Recovery Manager (RMAN) Oracle Data Guard
Advanced Protect	MirrorView/A through iSCSI
Test/dev	EMC CLARiiON SnapView clones

Pure NFS VMware HA cluster solution

Reference architecture diagram

Figure 3 describes the reference architecture for the pure NFS VMware HA cluster solution.

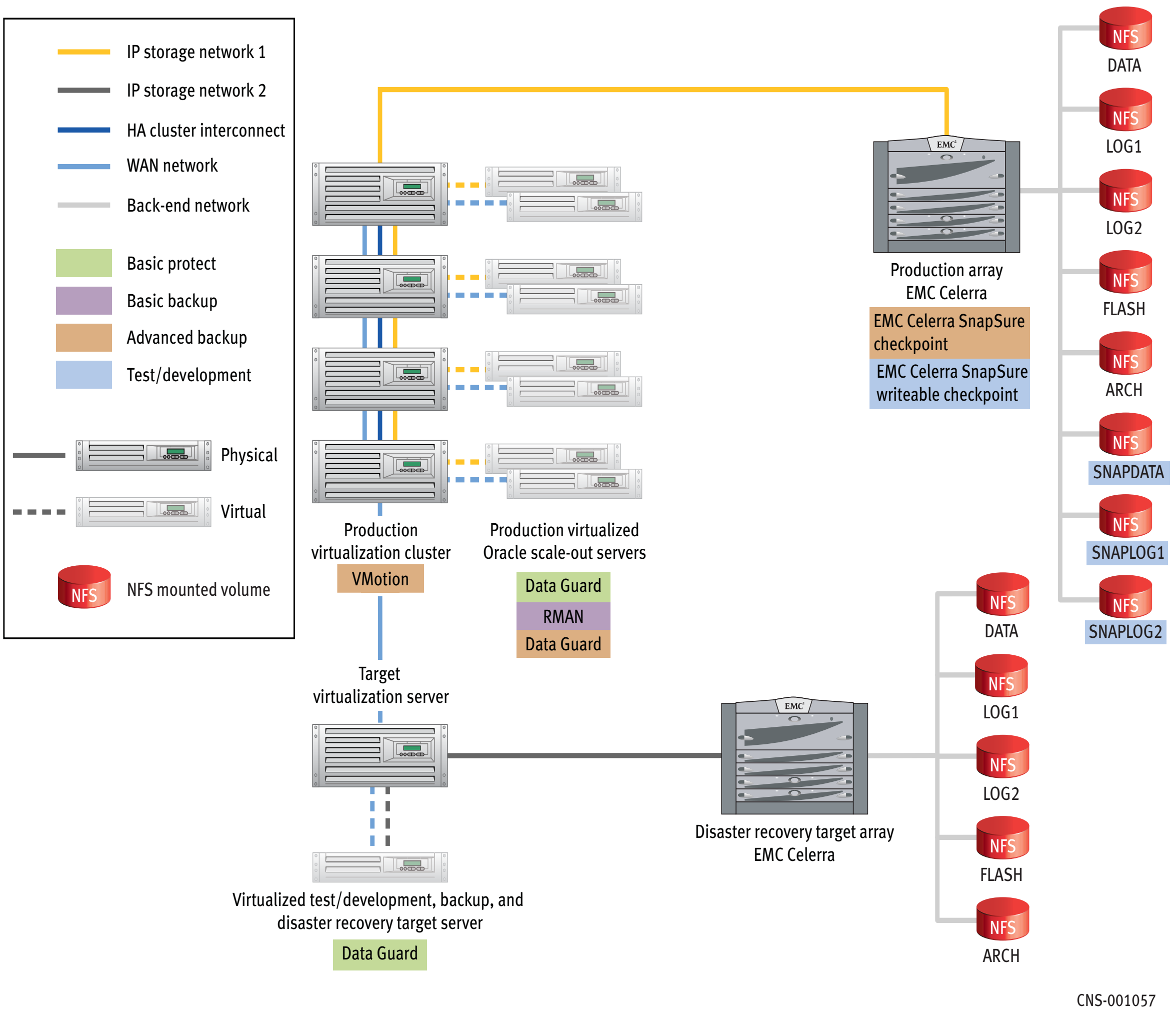

Figure 3 **Pure NFS VMware HA cluster solution: Oracle single instance on Linux using Celerra**

Solution features

In addition to the features described in "Virtualized solutions" and in "Pure NFS virtualized solution", the pure NFS VMware HA cluster solution includes the following additional elements:

Basic architecture

- A four-node VMware HA cluster is used. An isolated network consisting of three trunked Gigabit Ethernet (GbE) connections is used as the HA cluster interconnect.
- TCP/IP and kNFS provide network connectivity and file system semantics for NFS file systems on Oracle 10*g*. DNFS provides file system semantics for the Oracle 11*g* pure NFS solution.
- Two VMs are used for each ESX server.

Storage network architecture

- In the case of Oracle 11*g*, the Oracle database files, temp files, control files, online redo log files, and archive logs are accessed using the DNFS protocol.

Validated solution components

Table 3 describes the solution components that have been validated for the pure NFS VMware HA cluster solution.

Note: Only the Scale-Out OLTP solution component was validated on DNFS. The other solutions components will be validated on DNFS in the next test cycle.

Table 5 Pure NFS VMware HA cluster solution: Validated solution components

Solution component	Details
Scale-Out OLTP	Utilizes an industry-standard OLTP database performance benchmark
Basic Backup and Recovery	Oracle Recovery Manager (RMAN)
Advanced Backup and Recovery	EMC Celerra SnapSure checkpoints
Basic Protect	Oracle Recovery Manager (RMAN) Oracle Data Guard
Test/dev	EMC Celerra SnapSure writeable checkpoints to a virtualized single-instance target

Implementing DNFS: Mandatory fixes and patches

Known bugs

Oracle 11*g* Release 1 has two known bugs with regard to DNFS resiliency. Patches are available for both of these bugs; the details are described below.

Warning: If the appropriate Oracle patches are not applied, the stability and continuity of a running database may be at risk when configured to use DNFS.

EMC does not recommend the implementation of DNFS solutions on a NAS/IP storage architecture unless the appropriate Oracle patches have been installed and configured.

To implement the fixes

1. Ensure that you are running the most up-to-date Celerra DART version.

   The minimum Celerra DART version required for DNFS implementation is 5.6.40.3

2. Enable the **transChecksum** parameter on the Celerra X-Blades that serve the Oracle DNFS clients.

   To enable the **transChecksum** parameter, run the following command on the Celerra Control Station:

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

   This avoids the possibility of TCP Port and XID (transaction identifier) reuse by two or more databases running on the same database server, which could cause data corruption.

3. Download and install the Oracle 11.1.0.7 patch.
4. Download and install Oracle patch 8206872.

See Oracle Metalink for more information on downloading and installing these patches for your Linux operating system.

Network architecture

IP network architecture

Virtual LANs

Virtual LANs (VLANs) logically group devices that are on different network segments, or sub-networks. The solutions use three VLANs to segregate network traffic of different types. This improves throughput, manageability, application separation, high availability, and security.

Client VLAN

The client VLAN supports connectivity between the virtualized production Oracle Database 10*g*/11*g* servers, the virtualized target Oracle Database 10*g*/11*g*, and the client workstations. The client VLAN also supports connectivity between the Celerra and the client workstations to provide network file services to the clients. Control and management of these devices are also provided through the client network.

VMware HA cluster VLAN

The pure NFS VMware HA cluster virtualized solution uses an HA cluster interconnect network on an isolated VLAN consisting of three trunked NICs for each ESX server.

Storage VLAN

The storage VLAN uses the NFS protocol to provide connectivity between servers and storage. In the case of the pure NFS virtualized solution and the blended virtualized solution, one NIC on the ESX server is assigned to each virtual machine. In the case of the VMware HA cluster virtualized solution two NICs are used for each VM for the storage network.

Celerra NS40 Data Mover ports

Figure 4 illustrates the network ports on the rear of two EMC NS40 Data Movers.

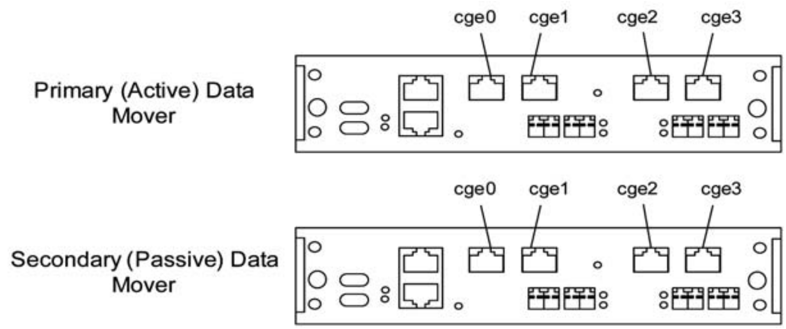

Figure 4 Celerra NS40 Data Mover ports and traffic types

Ports cge0 (the last character is a zero) and cge1 are aggregated and connected to the storage network. They handle all I/O required by the database servers to the datafiles, online redo log files, archived log files, control files, OCR file, and voting disk. Ports cge2 and cge3 are left open for future growth.

Note: Celerra Manager can be used to create Ethernet channels, link aggregations, and fail-safe networks.

FCP network architecture

FCP network characteristics

The following are the characteristics of the FCP network used in testing the blended FCP/NFS solution:

- Fully redundant FCP switches
- Two target ports on each of the CLARiiON CX4-240 storage processors are utilized. Each target port is connected to a separate FCP switch.
- The HBA ports on the ESX server are zoned to all the LUNs on the back-end storage, but only the required LUNs are presented to each VM using RDM.
- Zoning is employed on the FCP switches to ensure that all Oracle hosts have redundant paths to access all LUNs on the CLARiiON CX4-240 array.

EMC CX4-240 storage processor ports

Figure 5 illustrates the network ports on the rear of an EMC CX4-240 storage processor.

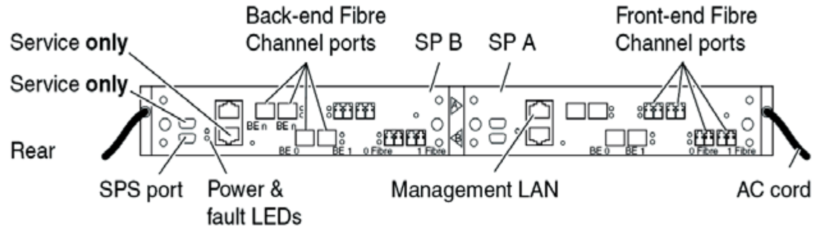

Figure 5 EMC CX4-240 ports

The Celerra Data Movers are always connected to ports 2 and 3 at the front end and FC ports on the back end of the EMC CLARiiON CX4-240. The remaining ports are available to the database servers for FCP connectivity through the redundant FCP storage network switches.

Storage architecture

Pure NFS solutions

NAS storage

Setting up NAS involves:

- Establishing the RAID levels
- Allocating hot spares
- Creating disk volumes
- Combining disk volumes (stripes and slices) into metavolumes, or disk pools, that can span disks
- Creating file systems that can be exported to servers

High availability and failover

EMC Global Solutions have built in high-availability (HA) features. These HA features allow the Celerra to survive various failures without a loss of access to the Oracle database. These features protect against the following:

- Data Mover failure
- Network port failure
- Power loss affecting a single circuit connected to the storage array
- Storage processor failure
- Disk failure

Pure NFS RAID configuration

Two sets of RAID and disk configurations have been tested, as shown in Table 6.

Table 6 Pure NFS solution RAID configurations

Figure	Configuration	Description
Figure 6	Pure NFS configuration 1	3 FC shelf RAID 5/RAID 1 AVM using user-defined storage pools
Figure 7	Pure NFS configuration 2	2 FC shelf RAID 5/RAID 1 AVM using user-defined storage pools

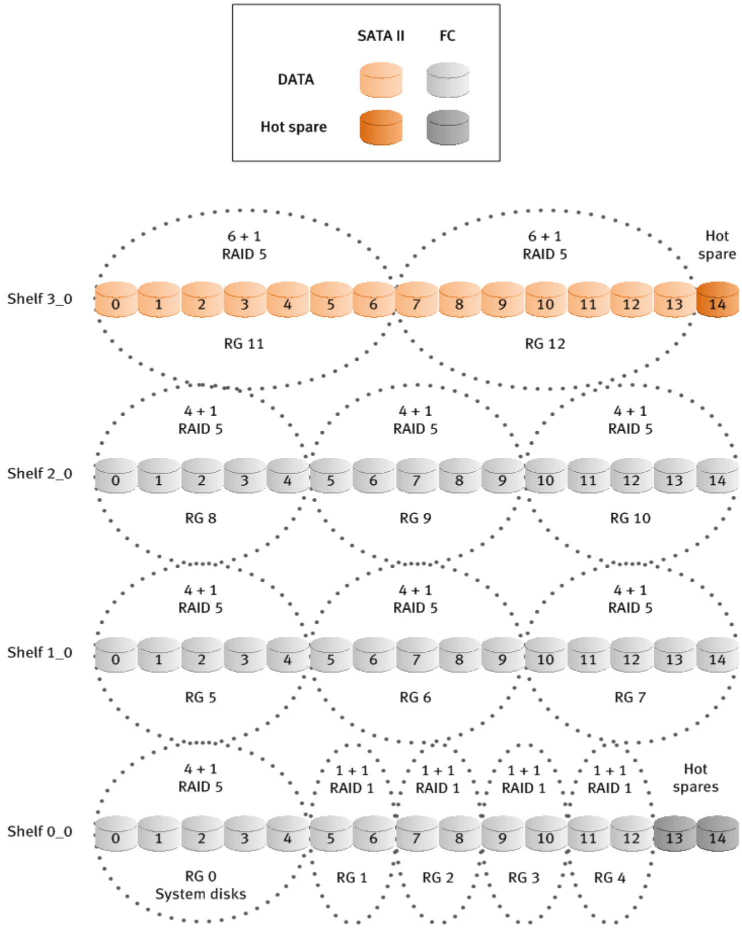

Figure 6 Pure NFS configuration 1: 3 FC shelf RAID configuration

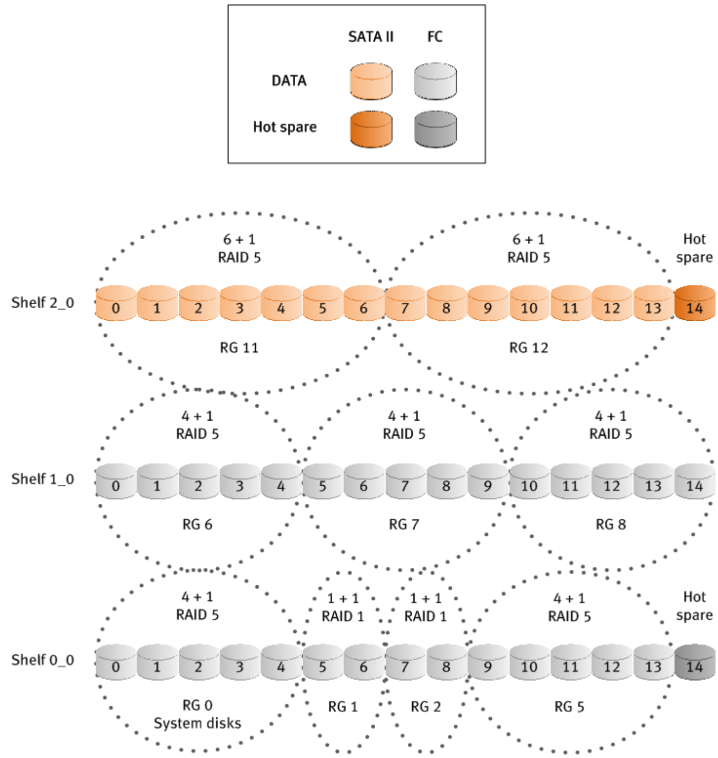

Figure 7 **Pure NFS configuration 2: 2 FC shelf RAID configuration**

Disk volume setup

Once the RAID groups are created, the Celerra automatically creates disk volumes that are accessible to the Data Movers.

Automatic Volume Management

Automatic Volume Management (AVM) is a way for the user to configure Celerra volumes from RAID groups. User-defined pools allow the user to control exactly which RAID groups are used for a given volume. With AVM having user-defined pools, the pool can be automatically expanded if additional RAID groups are added later. Thus, this configuration option provides the greatest flexibility and control for configuring Celerra volumes. AVM with user-defined pools was used to configure all volumes in the blended solution.

Table 7 contains a description of all the user-defined pools used.

File systems, exports, and mount points

The file systems shown in Table 7 were created on the AVM user-defined pools, exported on the Celerra, and mounted on the database servers.

For pure NFS virtualized solutions, the file systems for hosting the databases are located on a single Celerra. The file systems are created as shown in Table 7.

Separate file systems are created for datafs, log1fs, and log2fs that map to each database that is managed by each individual VM. This is to ensure that operations carried out on a particular database, such as snapshot-based backup or restore, do not affect file systems or databases being managed by other VMs.

The same archive and flash file systems are shared by all VMs, but separate directories are created to limit the access to the databases from each VM.

Table 7 File system layout

File system/ export	AVM user-defined pool	Volumes
/vm	Default pool	/VMwareBoot
/datafs	datapool (user-defined storage pool created using datastripe volume)	datastripe (metavol consisting of all available FC 4+1 RAID 5 groups)
/log1fs	log1pool (user-defined storage pool created using log1stripe volume)	log1stripe (metavol using half of the RAID 1 groups)
/log2fs	log2pool (user-defined storage pool created using log2stripe volume)	log2stripe (metavol using half of the RAID 1 groups)
/archfs	archpool (user-defined storage pool created using archstripe volume)	archstripe (metavol using the SATA 6+1 RAID 5 group)
/flashfs	flashpool (user-defined storage pool created using flashstripe volume)	flashstripe (metavol using the SATA 6+1 RAID 5 group)
/snapdatafs	datapool (user-defined storage pool created using datastripe volume)	snapdatafs (SnapSure writeable checkpoint of the datafs volume)
/snaplog1fs	log1pool (user-defined storage pool created using log1stripe volume)	snaplog1fs (SnapSure writeable checkpoint of the log1fs volume)
/snaplog2fs	log2pool (user-defined storage pool created using log2stripe volume)	snaplog2fs (SnapSure writeable checkpoint of the log2fs volume)

Blended FCP/NFS solution

SAN storage

Setting up SAN storage involves:

- Creating RAID groups
- Binding LUNs
- Allocating hot spares
- Creating a storage group
- Assigning hosts to the storage group
- Assigning LUNs to the storage group

High availability and failover

The EMC CLARiiON CX4-240 array has built-in, high-availability features that allow it to survive the following failures without a loss of access to the Oracle Database 10*g*/11*g* database:

- Storage processor failure
- FCP target port failure
- Power loss affecting a single circuit connected to the storage array
- Disk failure

Blended RAID configuration

Two sets of RAID and disk configurations have been tested, as shown in Table 8.

Table 8 Oracle Database 10g/11g FCP solution RAID configurations

Figure	Configuration	Description
Figure 8	Blended configuration 1	3 FC shelf RAID 1-0/RAID 1
Figure 9	Blended configuration 2	3 FC shelf RAID 5/RAID 1

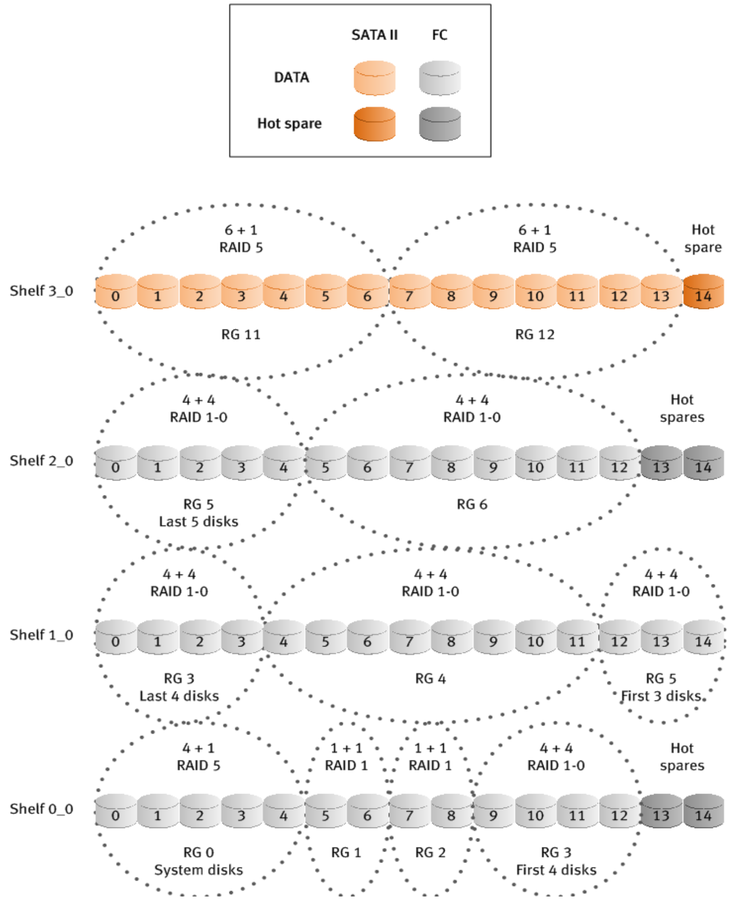

Figure 8 Blended configuration 1: 3 FC shelf RAID 1-0/RAID 1

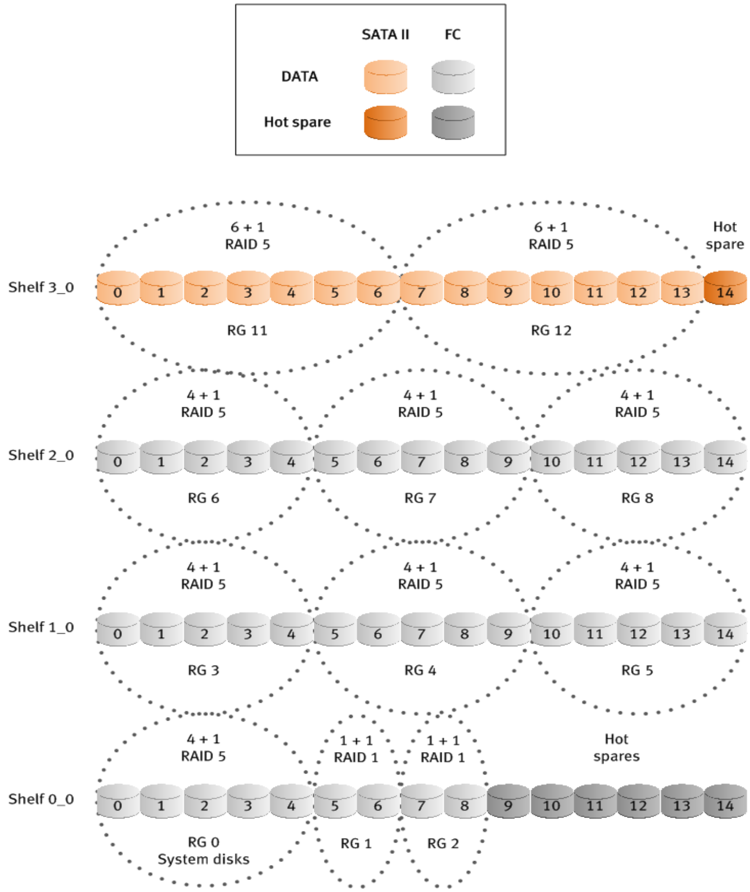

Figure 9 Blended configuration 2: 3 FC shelf RAID 5/RAID 1

LUN setup

Once the RAID groups are created, EMC Navisphere® software is used to create the LUNs that are used to store the database. These LUNs are added to a storage group accessible to the virtualized Oracle database servers.

ASM and NFS

ASM is used to store the database objects requiring high performance. 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 brings down cost in terms of storage, while improving manageability.

Table 9 contains a detailed description of all the database objects and where they are stored.

Table 9 File system layout

File system/ mount point	File system type	Data stored on	Contents
/vm	NFS	/VMwareBoot volume on Celerra	Virtual machine boot images
/RDM_PointerMap	VMFS	LUN 20	VMware Virtual Disk (VMDK) pointer files
+DATA	ASM	LUNs 3 through 8	Oracle datafiles
+LOG1 and +LOG2	ASM	LUNs 1 and 2	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
+CLONEDATA	ASM	LUNs 53 through 58	Cloned Oracle datafiles
+CLONELOG1 and +CLONELOG2	ASM	LUNs 51 and 52	Cloned online redo logs and control file (mirrored copies)

Blended virtualized solution

For the blended virtualized solution, four different sets of DATA and LOG LUNs are created on the same storage. Each set is presented to a single VM using RDM. Similarly, four different directories are created on the archive and flash file systems so that they can be individually mounted on the corresponding VMs.

Database server architecture

Virtualized solutions using a single ESX server

Pure NFS and blended FCP/NFS solutions

Each VMware ESX server has 14 network interfaces.

Table 10 lists each interface and describes its use.

Table 10 VMware ESX server network interface configuration

Interface port ID	Description
vmnic0	Public network
vmnic1	Unused
vmnic2	Unused
vmnic3	Unused
vmnic4	Unused
vmnic5	Unused
vmnic6	Unused
vmnic7	Unused
vmnic8	Unused
vmnic9	Unused
vmnic10	Storage network (VM1)
vmnic11	Storage network (VM2)
vmnic12	Storage network (VM3)
vmnic13	Storage network (VM4)

A single ESX server hosts four VMs, and four independent single instance TPC-C databases are created on each VM. The ESX server has 32 GB of physical memory and eight CPUs. Each VM is allocated 7 GB of memory and two CPUs.

Pure NFS

In the case of the pure NFS solutions all four databases reside on a single EMC Celerra storage array. The file systems are directly mounted using NFS on the VMs.

Each virtualized Oracle Database 10*g*/11*g* server uses the NFS protocol to directly connect to the Celerra storage array. NFS runs on top of TCP/IP over the virtualized Ethernet connection provided by the VMware ESX server.

Blended FCP/NFS

For the blended solution the data and redo logs reside on a single EMC CLARiiON CX4-240 series array, and the archive and flash area reside on a Celerra. A separate set of LUNs are created on the EMC CLARiiON CX4-240 series array and presented accordingly using RDM to the respective VMs. The archive and flash file systems are directly mounted using NFS on the VMs.

NFS client

Each virtualized Oracle Database 10*g*/11*g* server uses the NFS protocol to directly connect to the Celerra storage array. NFS runs on top of TCP/IP over the virtualized Ethernet connection provided by the VMware ESX server.

Virtualized solution using multiple ESX servers

Pure NFS VMware HA cluster

Each ESX server has 10 network interfaces. Two interfaces connect to the storage network, using a link aggregation trunk. Three interfaces connect the server to the HA cluster interconnect network, enabling the heartbeat and other network I/O required by HA cluster. One interface connects to the client network.

Table 11 lists each interface and describes its use.

Table 11 ESX server network interface configuration

Interface port ID	Description
eth0	Client network
eth1	Unused
eth2	Unused
eth3	Unused
eth4	Storage network (virtualized by ESX)
eth5	Storage network (virtualized by ESX)
eth6	Unused
eth7	HA cluster interconnect (virtualized by ESX)
eth8	HA cluster interconnect (virtualized by ESX)
eth9	HA cluster interconnect (virtualized by ESX)

VMware HA cluster

VMware HA cluster is enabled on each of the VMware ESX servers. Two VMs containing the single-instance Oracle database servers run on each VMware ESX server.

NFS client

Oracle 10*g*

Each virtualized Oracle Database 10g server uses the NFS protocol to connect to the Celerra storage array. NFS runs on top of TCP/IP. The storage network virtualized by the VMware ESX servers is used to access the NFS mounts.

Oracle 11*g*

In the case of Oracle 11*g*, the embedded DNFS protocol is used to connect each virtualized Oracle database server to the Celerra storage array.

Table 12 lists each interface and describes its use for the DNFS configuration.

Table 12 DNFS: Database server network interface configuration

Interface port ID	Description
eth0	Client network
eth1	Unused
eth2	Unused
eth3	Unused
eth4	Storage network
eth5	Storage network
eth6	Unused
eth7	HA cluster interconnect (virtualized by ESX)
eth8	HA cluster interconnect (virtualized by ESX)
eth9	HA cluster interconnect (virtualized by ESX)

High availability and failover

VMware ESX virtual switching provides aggregation, load balancing and failover of the TCP/IP connections. Two NICs on each ESX server are used in the storage network. The HA cluster interconnect network is aggregated in a similar manner using three NICs.

Port aggregation using an ESX virtual switch was not used on the DNFS configuration for storage connectivity. Load balancing and high availability were taken care of by the DNFS protocol itself.

Application architecture

Oracle binary files

The Oracle Database 10*g*/11g binary files, including the Oracle ASM software (if used), are stored in the VM's vmdk file that is stored via NFS on a file system mounted directly onto the VMware ESX server.

Pure NFS solutions

Datafiles, online redo log files, archive log files, and temp files reside on Celerra NFS file systems. These file systems are designed (in terms of the RAID level and number of disks used) to be appropriate for each type of file.

Table 13 lists each file or activity type and indicates where it resides.

Table 13 Application file or activity types and locations

File or activity type	Location
Database binary files	vmdk file
Datafiles, tempfiles	/datafs
Online redo log files	Mirrored across /log1fs and /log2fs
Archived log files	/archfs
Flashback recovery area	/flashfs
Control files	Mirrored across /log1fs and /log2fs

Blended FCP/NFS solutions

High-performance files (datafiles, tempfiles, controlfiles, and online redo logfiles) reside on the EMC CLARiiON CX4-240 series array and are accessed through FCP. Oracle ASM is used to manage these files. All other database files (archive log files, backup target, and the flashback recovery area) reside on the Celerra storage array, and are accessed through NFS.

In the case of both ASM/FCP and NFS, these storage areas are designed (in terms of the RAID level and number of disks used) to be appropriate for each type of file.

Table 9, "File system layout," provides more information on the location where each of these objects is stored.

3

Environment Profile

This chapter presents the following topics:

Hardware resources

The hardware used by the solutions is described in Table 14.

Table 14 Hardware specifications

Equipment	Quantity	Configuration
EMC Celerra NS40 Series Unified Storage Platform (includes an EMC CLARiiON CX4-240 back-end storage array)	2	• 2 Data Movers • 4 GbE network connections per Data Mover • 2 or 3 FC shelves • 1 SATA shelf • 30 or 45 73 GB FC disks (depending on configuration) • 15 500 GB SATA disks • 1 Control Station • 2 storage processors
Gigabit Ethernet switch	5 (client, VMware HA cluster interconnect, and storage networks)	• 24 ports per switch
FCP switch	2	• 16 ports • 4 Gb throughput
VMware ESX HA cluster server	4	• 2 2.66 GHz Intel Pentium 4 quad-core processors • 24 GB of RAM • 2 146 GB 15k internal SCSI disks • 2 onboard GbE Ethernet NICs • 2 additional Intel PRO/1000 PT quad-port GbE Ethernet NICs • 2 SANblade QLE2462-E-SP 4 Gb/s dual-port FC HBAs (4 ports in total)
De-duplication array	1	• Capacity: 12.26 TB
Virtualization server	2	• 4 2.86 GHz AMD Opteron quad-core processors • 32 GB of RAM • 2 146 GB 15k internal SCSI disks • 2 onboard GbE Ethernet NICs • 3 additional Intel PRO/1000 PT quad-port GbE Ethernet NICs • 2 SANblade QLE2462-E-SP 4 Gb/s dual-port FC HBAs (4 ports in total)

Software resources

The software used by the solutions is described in Table 15.

Table 15 Software specifications

Software title	Number of licenses
Red Hat Enterprise Linux version 4.5 or Oracle Enterprise Linux version 5.2	1 per virtualized production or target database server
VMware ESX Server 3.5	1 per VMware server
Microsoft Windows Server 2003 Standard Edition	1 per virtual client
Oracle Database 10*g* or 11*g* Standard Edition	1 per virtualized production or target database server
Quest Benchmark Factory for Databases 5.0.1	1 per virtual client
EMC Celerra Manager Advanced Edition version 5.6.37-6	1 per Celerra NS40
EMC Navisphere Agent version 6.26.0.2.24	1 per database server
EMC FLARE® version 6.26 patch level 14	1 per CLARiiON storage processor
EMC Navisphere Management version 6.26	1 per CLARiiON storage processor

Glossary

This glossary defines technical and industry-specific terms used in this document.

A

Advanced Backup and Recovery

A solution component that provides backup and recovery functionality through the storage layer using specialized hardware or software.

Advanced Backup and Recovery has the following benefits:

- Offloads the database server's CPUs from the I/O and processing requirements of the backup and recovery operations
- Superior MTTR through the use of virtual storage layer replication (commonly referred to as snapshots)
- Reduced power, cooling and space requirements through the use of de-duplication

Advanced Protect

A solution component that provides disaster recovery functionality through the storage layer using specialized hardware or software.

Advanced Protect has the following benefits:

- Offloads the database server's CPUs from the I/O and processing requirements of the disaster recovery operations
- Superior failover and failback capabilities
- Reduces the software required to be installed at the disaster recovery target because of the use of consistency technology

B

Basic Backup and Recovery

A solution component that provides backup and recovery functionality through the operating system and the database server software stack.

Basic Backup and Recovery uses the database server's CPUs for all I/O and processing of backup and recovery operations.

Basic Protect

A solution component that provides disaster recovery functionality through the operating system and the database server software stack.

Basic Protect uses the database server's CPUs for all I/O and processing of disaster recovery operations.

blended solution

A solution that uses a combination of the NFS and FCP protocols. A blended solution requires the use of a Celerra Unified Storage Platform.

- FCP is used for high-I/O and low latency database objects (notably the datafiles, temp files, online redo log files, and control files).
- NFS is used for all other database objects (consisting basically of the flashback recovery area, disk-based backups, and archive logs).

A blended solution provides ease-of-use advantages in that the amount of software required to be installed and configured on the database server is reduced, and normal file system semantics can be used for the NFS mounted database storage objects.

D

Direct NFS (DNFS)

A network storage protocol in which the NFS client is embedded in the Oracle 11*g* database kernel.

H

high availability

The use of specialized hardware or software technology to reduce both planned and unplanned downtime

I

Information Lifecycle Management

The use of tiered storage to address the unique I/O and performance requirements of various vintages of data, as the data ages during its useful life.

K

Kernel NFS (kNFS)
A network storage protocol in which the NFS client is embedded in the operating system kernel.

O

online transaction processing (OLTP)
The use of a database system to capture primary business data that represents the ongoing, marginal operations of the business.

An example is POS records from a retail operation or call records from a telephone switch.

P

physically booted solution
A configuration in which the production database servers are directly booted off of a locally attached, hard disk without the use of a hypervisor such as VMware or Oracle VM.

Note: Utility servers (such as test/dev target or disaster recovery target) may still be virtualized in a physically-booted solution.

---

pure solution
A solution in which all database storage objects are accessed using the NFS protocol.

R

resiliency
Testing that is designed to validate the ability of a configuration to withstand faults at various layers.

The layers that are tested include: network switch, database server storage network port, storage array network port, database server cluster node, and storage processor.

S

scale-out
The practice of partitioning a database workload using a federated approach across multiple, separate, non-clustered database servers.

Scale-out is useful for software-as-a-service configurations. It allows scaling by simply adding another database server to the configuration. Scale-out requires a workload in which any given customer of the database application needs to see only a subset of the database data.

scale-up
The use of a clustered or single-image database server configuration. Scaling is provided by increasing the number of CPUs in the database server (in the case of a single-instance configuration) or by adding nodes to the cluster (in the case of a clustered configuration).

Scale-up assumes that all customers of the database will be able to access all database data. It is more difficult to scale a scale-up configuration than it is to scale a scale-out configuration.

solution
A software and hardware configuration upon which a customer would choose to run their entire business, or a given business application.

A solution includes database server hardware and software, IP networks, storage networks, storage array hardware and software, among other components.

solution component
A component of a solution that provides a discrete function, such as backup and recovery, or data protection.

Solution components are typically tested as a separate project within the overall solution validation process.

T

test/dev
The use of storage layer replication (such as snapshots and clones) to provide an instantaneous, writeable copy of a running production database with no downtime on the production database server and with minimal, if any, performance impact on the production server.

Test/dev is a common requirement of many database customers.

V

virtualized solution
A configuration in which the production database servers are virtualized using a hypervisor technology such as VMware or Oracle VM.