EMC²
where information lives®

White Paper

Oracle Database 10*g*/Oracle RAC 10*g* Celerra NS Series NFS

Best Practices Planning

Abstract

This white paper presents the best practices for configuration, backup and recovery, and protection of Oracle Database 10*g* Single Instance and Real Application Clusters on Red Hat Enterprise Linux with EMC® Celerra® NS Series storage arrays.

September 2006

Part Number H2369

Table of Contents

Executive summary

The EMC® Celerra® Network Server is an enterprise-level, high-performance, dedicated network file server that delivers data over the IP network via industry-standard protocols. The Celerra Network Server can be deployed in the existing IP network infrastructure, and using it to share information in the heterogeneous environment is easy and reliable, as if the information were stored on local workstations.
Celerra Network Server, with its advanced features of high availability and scalability, coupled with the proven, flexible, and high-performing technology of EMC's other storage products, is an attractive storage option for an Oracle database.

Oracle over NFS on Celerra Network Server provides the benefits described in Table 1.

Table 1. Benefits of Oracle over NFS on Celerra Network Server

Benefit	Details
Lower total cost of ownership (TCO)	Reduces acquisition, administration, and maintenance costs than equivalent DAS or SAN does
Greater manageability	Eases implementation, provisioning, and volume management
Simplified Real Application Cluster (RAC) implementation	Provides NFS-mounted shared file systems
High availability	Can implement a clustering architecture that provides very high levels of data availability
Increased flexibility	Easily makes databases, or copies of database, available (via remounts) to other servers
Improved protection	Integrates both availability and backup
Benefits of EMC Information Lifecycle Management (ILM)	Implements tiered storage

Introduction

This white paper describes the best practices for running Oracle Database 10*g* Single Instance and Real Application Clusters on a Red Hat Enterprise Linux server with an EMC Celerra Network Server. The topics covered include setup and configuration of:

- Storage
- Network
- Database server hardware and BIOS
- Linux operating system install
- Oracle software install
- Database parameters and settings
- RAC resiliency
- Backup and recovery
- Protection (disaster recovery)
- Cloning
- Migration
- Celerra monitoring

Oracle performance tuning is beyond the scope of this paper. The *Oracle Database Performance Tuning Guide* provides more information on this topic.

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 Database 10g software, basic Red Hat Enterprise 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 the Oracle Database 10g software.

Configuration

Oracle Database 10g Single Instance

The minimum requirements for the solution are:

- Celerra NS Series with DART 5.4
- Oracle Database 10*g* Release 1 Enterprise Edition
- Red Hat Enterprise Linux version 3
- NFS version 3

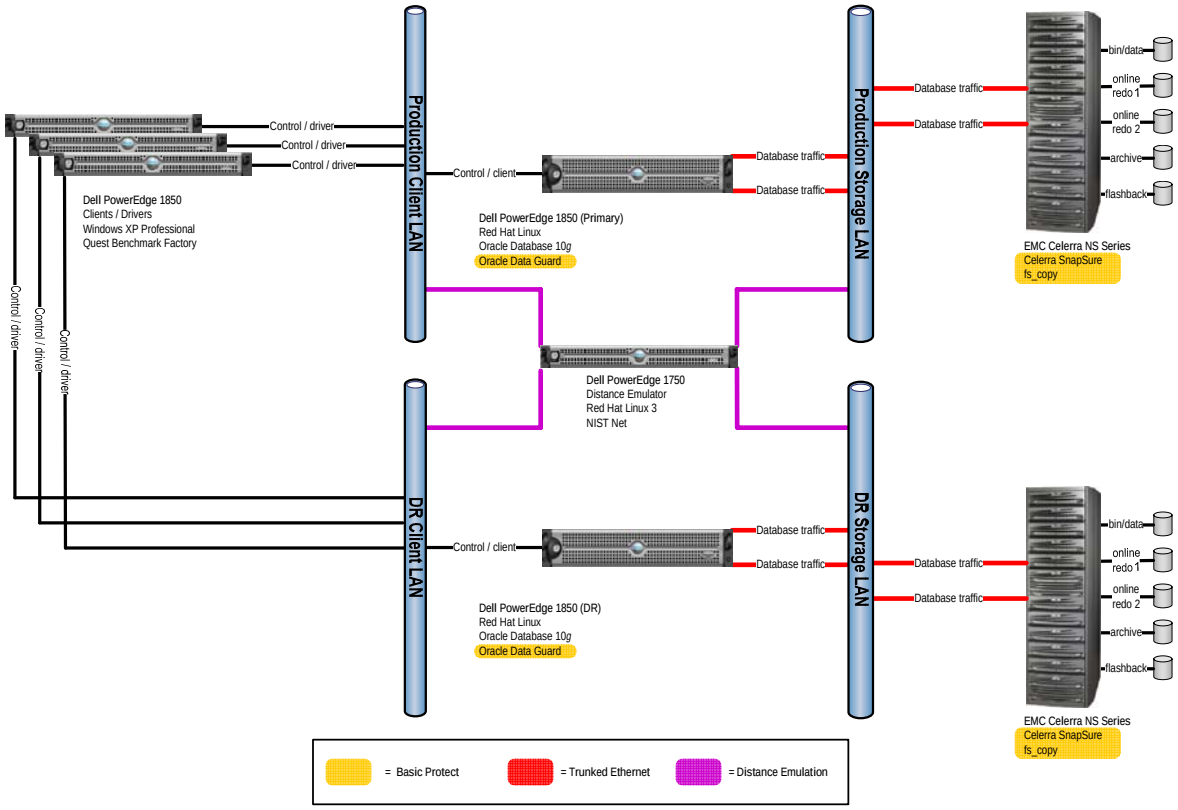

Figure 1. Single instance configuration

Oracle Real Application Clusters 10g

The minimum requirements for the solution are:

- Celerra NS Series with DART 5.5
- Oracle RAC 10*g* Release 2 Enterprise Edition
- Red Hat Enterprise Linux version 4
- NFS version 3

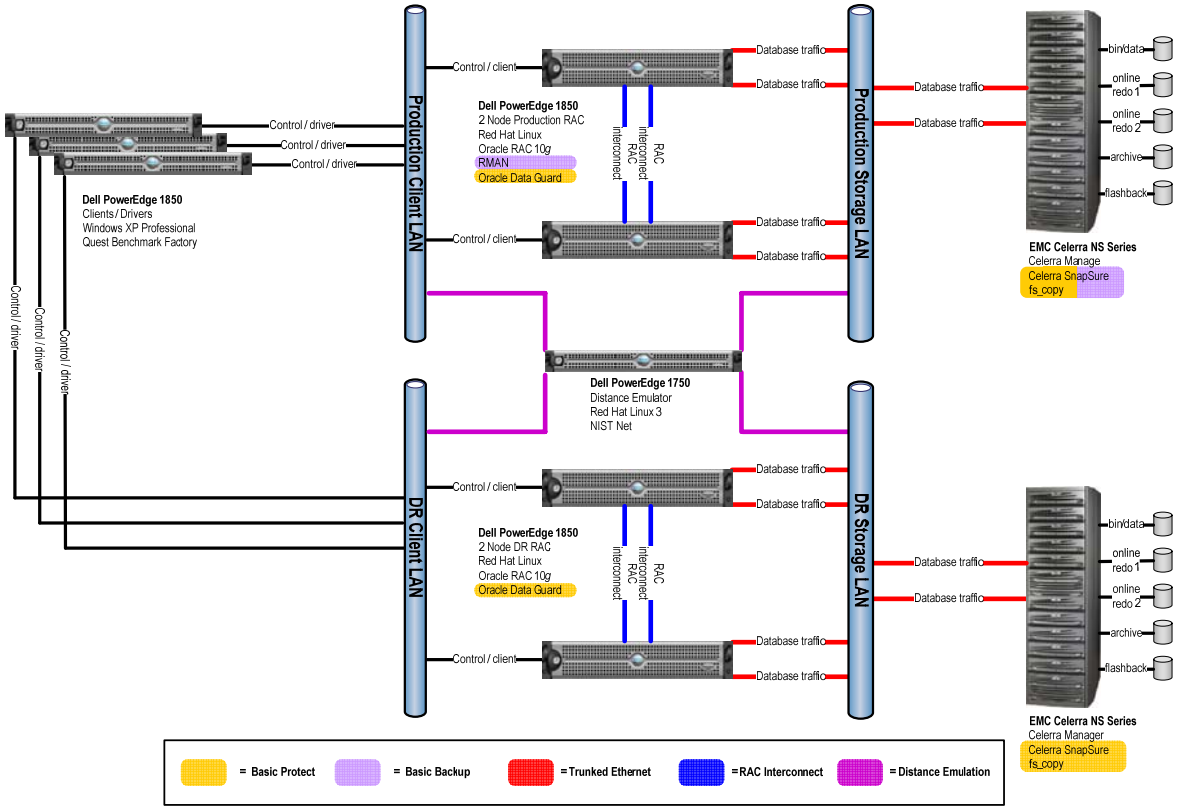

Figure 2. RAC configuration

Storage setup and configuration

Disk drive recommendations

The following are 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.
- Advanced Technology-Attached (ATA) drives have slower response rotational speed and moderate performance with random I/O. However, they are less expensive than Fibre Channel drives for the same or similar capacity. ATA drives are therefore the best option for storing archived redo logs and the flashback recovery area.

RAID groups

Table 2 summarizes general recommendations for RAID types corresponding to different Oracle file types:

Table 2. Recommendations for RAID types corresponding to Oracle file types

Description	RAID 1 / FC[1]	RAID 5 / FC	RAID 3 / ATA
Datafiles	OK[2]	Recommended [3]	Avoid
Control files	OK	OK	Avoid
Online redo logs	Recommended[4]	Avoid	Avoid
Archived logs	OK	OK	Recommended [5]
Flashback recovery area	OK	OK	Recommended
OCR file / voting disk	OK[6]	OK	Avoid

The tables in the "Volume management" section contain the storage templates that can be used for Oracle Database 10*g* databases on a Celerra Network Server. That section can help you determine the best configuration to meet your performance needs.

File system guidelines

EMC recommends that you create a minimum of five file systems for databases. Table 3 gives recommendations on these files:

[1] Celerra RAID 1 with striping is basically RAID 1+0, that is, a stripe of mirrors. Celerra provides striping on RAID 1 volumes.

[2] In some cases, if an application creates a large amount of temp activity, placing your temporary tablespace datafiles on RAID 1 devices instead of RAID 5 devices may provide a performance benefit due to RAID 1 having superior sequential I/O performance. The same is true for undo tablespaces as well if an application creates a lot of undo activity. Further, an application that creates a large number of full table scans or index scans may benefit from these datafiles being placed on a RAID 1 device.

[3] RAID 5 is generally recommended for database files, due to storage efficiency. However, if the write I/O is greater than 30 percent of the total I/O, then RAID 1 (with Celerra striping) may provide better performance, as it avoids hot spots and gives the best possible performance during a disk failure. Random write performance on RAID 1 can be as much as 20 percent higher than on RAID 5.

[4] 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 simple mirroring (RAID 1). Further, RAID 1 provides the best data protection, and protection of online redo log files is critical for Oracle recoverability.

[5] In some cases, placing archived redo log files on RAID 1 may be appropriate. RAID 1 for archived redo logs will provide better mean time to recovery, as the sequential read performance is superior to RAID 3. However, due to storage efficiency, RAID 3 may be chosen. This is a tradeoff, and must be determined on a case-by-case basis.

[6] Three copies of the voting disk are required by the Oracle installer if normal redundancy is used. We placed one copy on each of our online redo log volumes (RAID 1) and the third copy on the datafile volume (RAID 5). You should use FC disks for these files as unavailability of these files for any significant period of time (possibly due to disk I/O performance issues) may cause one or more of the RAC nodes to reboot and fence itself from the cluster.

Table 3. Recommendations for file systems

File system	Contents	Mount point	RAID level	Disk type
datafs	Oracle datafiles, voting disk copy 1	/u02	5	FC
log1fs	Online redo log files, OCR file copy 1, voting disk copy 2	/u03	1	FC
log2fs	Online redo log files, OCR file copy 2, voting disk copy 3	/u04	1	FC
archfs	Archived log files	/u05	3	ATA
flashfs	Flashback recovery area	/u06	3	ATA

In Oracle RAC 10*g* the main file components of Oracle Clusterware are the Oracle cluster repository (OCR file) and the voting disk. Due to the nature of these files, they must be placed on shared devices. EMC recommends that you place these files on NFS mounted volumes.

Best practices for file system design dictate that a file system should consist entirely of volumes 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 database file system:

- RAID levels
- Disk types
- Disk rotational speeds

EMC also recommends that if you need to replace the entire contents of an existing file system (such as a full database restore) you perform a full file system rebuild. This will produce a better data layout, resulting in better sequential I/O performance.

Number of shelves

For high performance, EMC recommends that you use a minimum of two Fibre Channel shelves and one ATA shelf to store Oracle databases on a Celerra Network Server. The most common error people make when planning storage is designing for capacity rather than for performance. The most important single storage parameter for performance is disk latency. High disk latency is synonymous with slower performance; low disk counts leads 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.

Volume management

EMC recommends that you use Manual Volume Management (MVM) for high-performance Oracle databases on a Celerra Network Server. MVM provides you with the flexibility of precise placement of file systems on particular disks or on particular locations on specific disks. It also allows you to stripe the workload across the storage processors on the back-end storage array, typically EMC CLARiiON®. While using MVM, select the LUNs that alternate ownership by the two storage processors. Figure 3, Figure 4, and Figure 5 contain example layouts for Automatic Volume Management (AVM) and MVM 2 FC shelf mixed RAID 1 / RAID 5 configurations, plus an MVM 3 FC shelf mixed RAID 1 / RAID 5 layout.

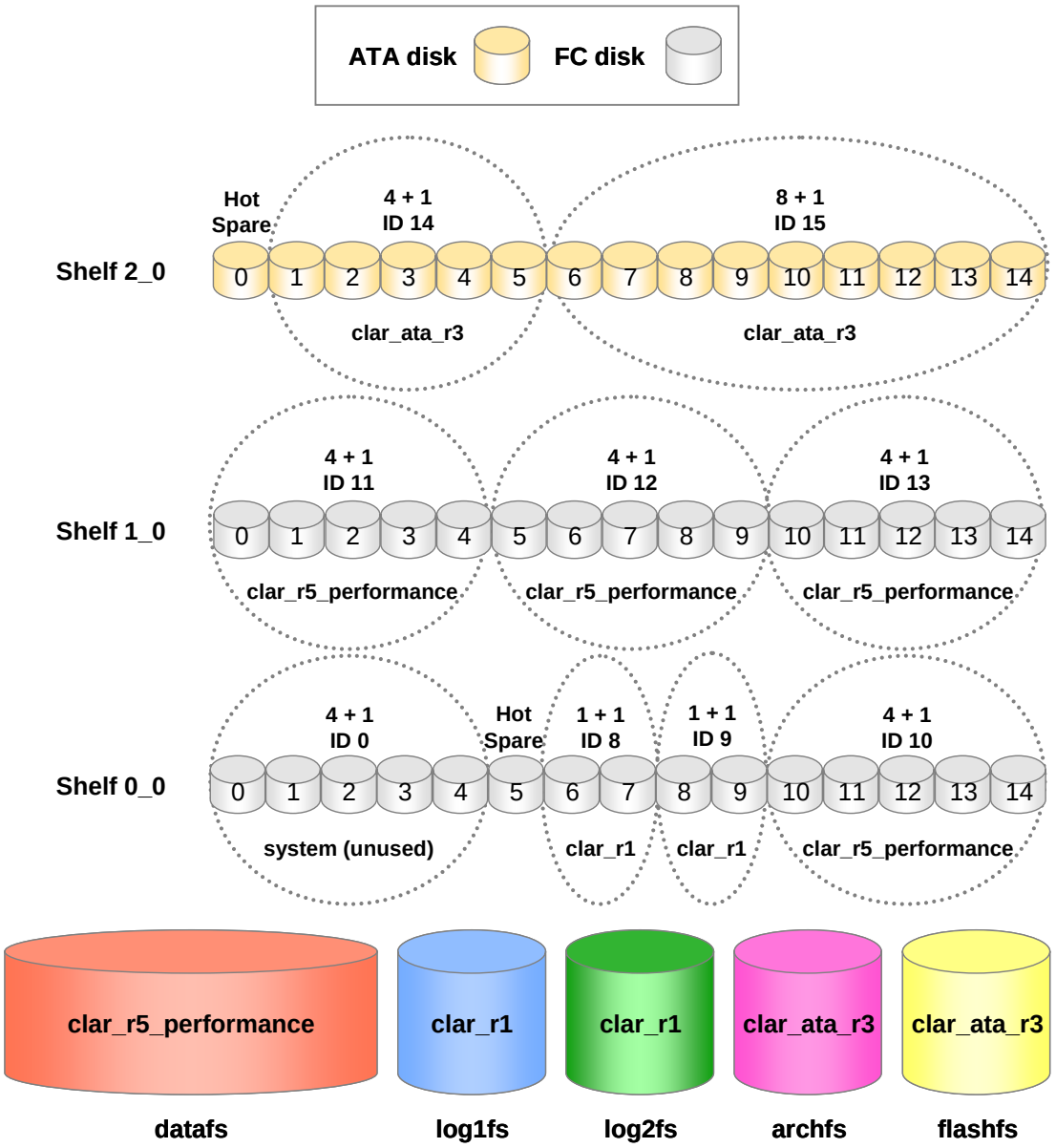

Figure 3. AVM FC shelf mixed RAID 1 / RAID 5 configuration

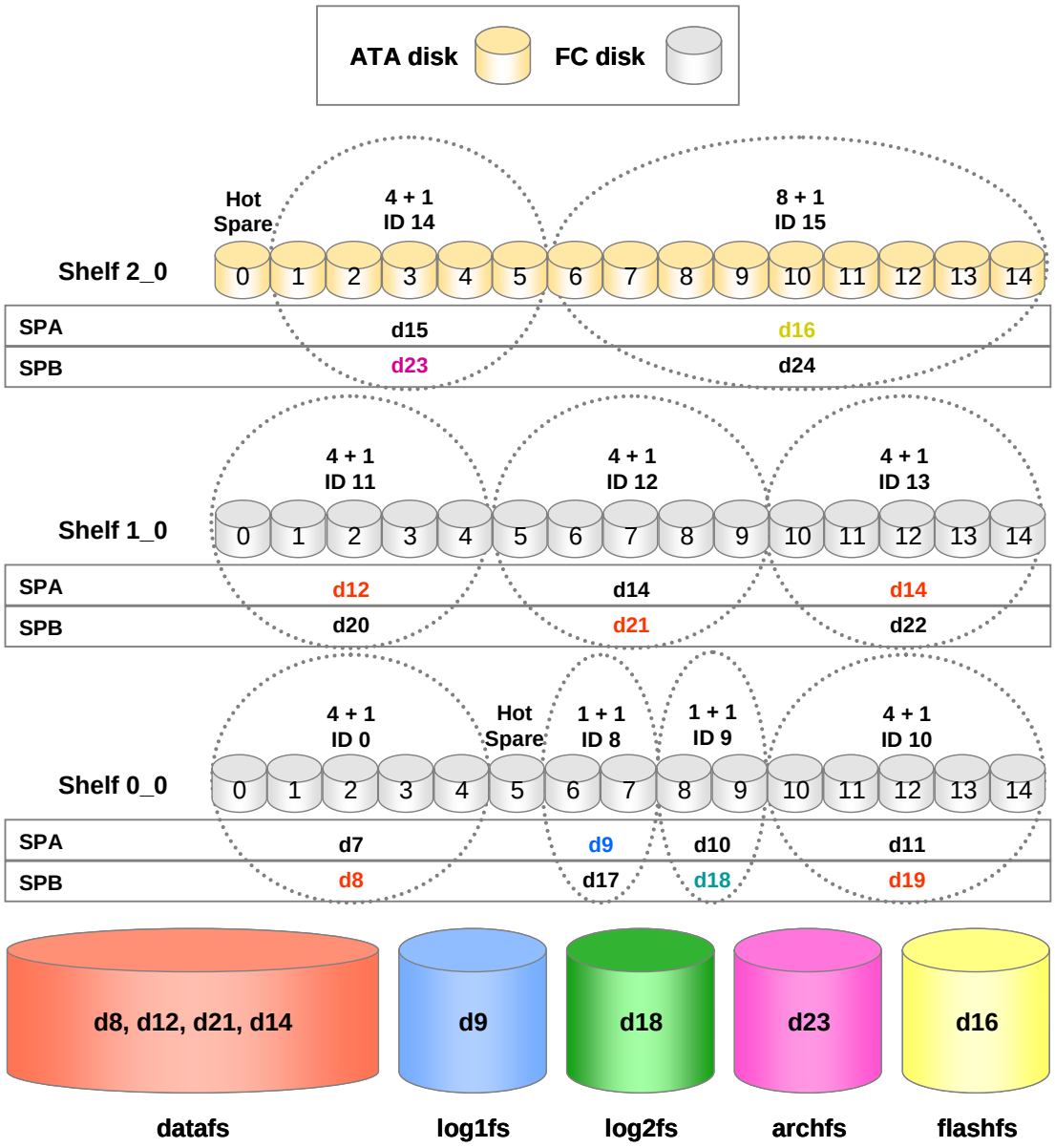

Figure 4. MVM 2 FC shelf mixed RAID 1 / RAID 5 configuration

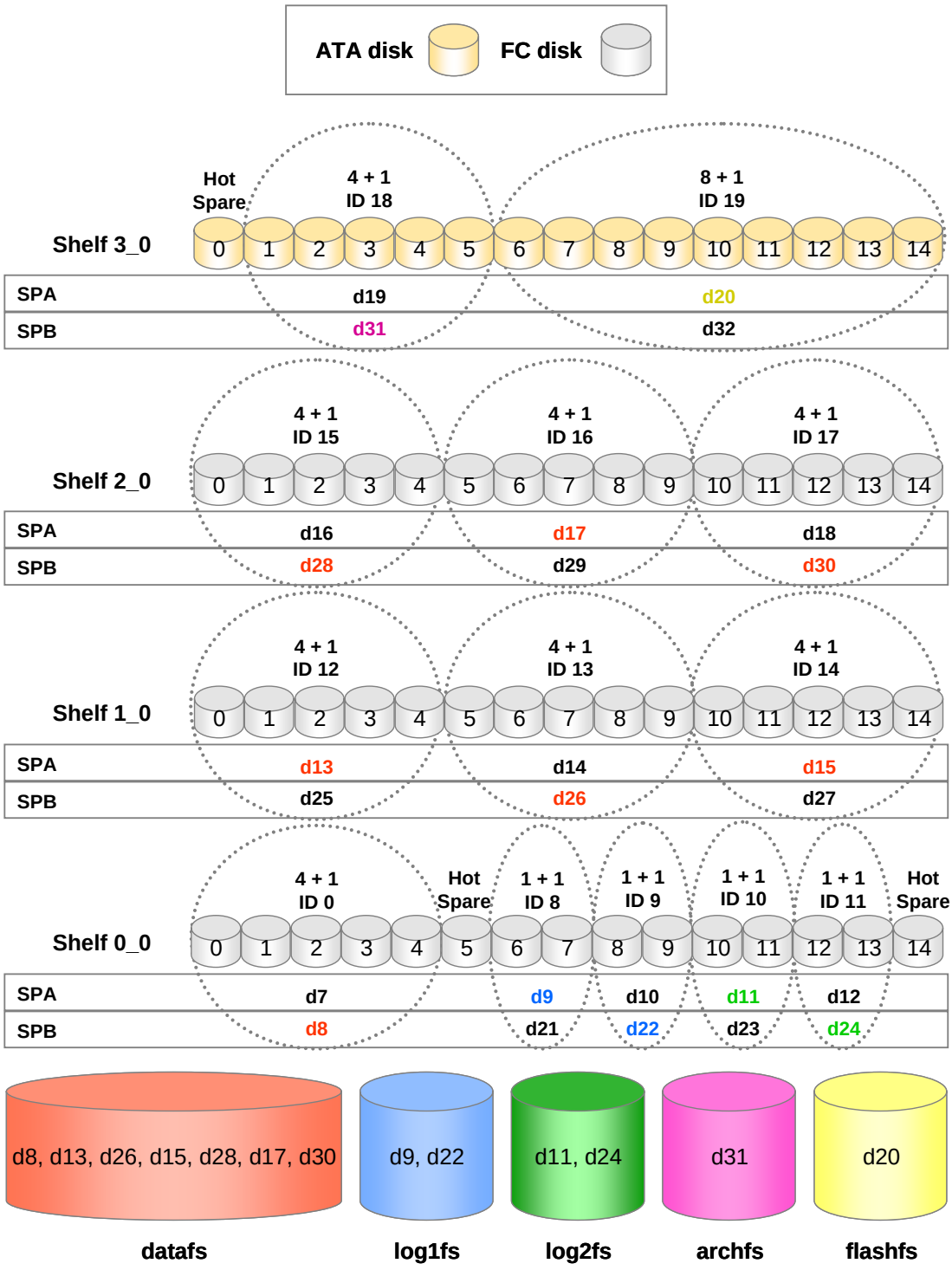

Figure 5. MVM 3 FC shelf mixed RAID 1/RAID 5 configuration

Note that in the MVM templates, the d-vols for each RAID group have been chosen to alternate between the storage processors on the CLARiiON back end. This provided the best data concurrency resulting in better performance.

However, for simplicity you may decide to use AVM, as shown in Figure 3. AVM provides a simple way to automatically create and manage volumes. This eliminates the need to manually create stripes, slices, or metavolumes, which is required by MVM. Further, AVM does this while supporting high availability.

If using AVM, you should ideally allow at least 20 spindles to create a file system for a RAID 5 configuration and at least four spindles for a RAID 1 configuration. Table 4 contains the configurations that EMC found to be optimal for storing Oracle databases.

Table 4. Optimal configurations for storing Oracle databases

Oracle type	Config	MVM / AVM	FC Shelves [7]	RAID groups for datafiles (RAID 5 4+1)	RAID groups for log files (RAID 1 1+1)	Usage
Single Instance	1	AVM	2	4	2	Simplicity / ease of use
	2	MVM	2	5	2	Performance / scalability
RAC	3	AVM	2	4	2	Simplicity / ease of use
	4	MVM	2	5	2	Performance / scalability
	5	MVM	3	7	4	Maximum performance / scalability

For Oracle Database 10*g* Single Instance, EMC recommends configuration 2 for best performance and scalability, and configuration 1 for ease of management. For Oracle RAC 10*g*, configuration 4 is recommended over configuration 3 for performance and scalability. For maximum scalability of a four-node Oracle RAC 10*g* installation, configuration 5 is recommended.

Stripe size

EMC recommends a stripe size of 32 KB for all types of database workloads. This is true for both AVM and MVM.

The default stripe size for MVM is 32 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, issue the following command on the Celerra Control Station:

```
nas_cmd @nas_profile –modify clar_r5_performance_vp –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 Network Server. This feature offers redundancy at the file-server level, allowing continuous data access. It also helps build a fault-resilient RAC architecture.

[7] All configurations assume a single ATA shelf, containing the archfs and flashfs file systems, and configured with one RAID 3 4 +1 RAID group and one RAID 3 8+1 RAID group.

EMC recommends that you set up auto-policy for the Data Mover, so should a Data Mover fail, 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. 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 timeout 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" to 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 upon 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 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
```

NFS threads

EMC recommends that you use the default 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.

Network setup and configuration

Gigabit connection

EMC recommends that you use Gigabit Ethernet for the network connections between the database servers and the Celerra Network Servers. Never use 100BaseT. Also 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 5 describes the database server network port setup.

Table 5. Database server network port setup

VLAN ID	Description	CRS setting
1	Client network	Public
2	RAC interconnect	Private
3	Storage	Do not use

Network port configuration

EMC recommends that you spread the NFS I/O over two out of the four available Celerra network ports. The architecture diagrams show the recommended configuration.

EMC also recommends that you set speed and duplex settings to *auto* on all ports. This is one of the single most common (and easily resolvable) performance issues observed today.

Security

EMC recommends that you place the Celerra Network Server and the Oracle Database 10*g* servers in their own private segregated storage network, so that the traffic between them can be isolated. Also the IP forwarding function available on Linux should be disabled on database servers, so the servers won't act as routers for other networked computers inappropriately attempting to reach the Data Movers.

The network in which the external interface of the Control Station resides should also be on a segregated subnet, secured by the firewall, as the Control Station is the most important administrative interface to the Celerra Network Server. The only computers that can reach the Control Station should be the database servers.

Any other security policies implemented in your organization should be employed to secure the environment.

Jumbo frames

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 storage or RAC interconnect networks. For Oracle Database 10*g* RAC 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). Celerra Data Movers support MTU sizes of up to 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 reduces the number of frames needed for every large I/O request and thus reduces the host CPU needed to generate a large number of interrupts for each application I/O. The benefit of jumbo frames is primarily a complex function of the workload I/O sizes, network utilization, and Celerra Data Mover CPU utilization, so it is not easy to predict.

Detailed instructions on configuring jumbo frames in an Oracle Database 10g on Red Hat Linux environment is contained in the *Oracle Database 10g/Oracle RAC 10g Celerra NS Series NFS Applied Technology Guide.* For information on using jumbo frames with the RAC Interconnect, please see Note 300388.1 on Metalink.

Database servers setup and configuration

BIOS

Dell 1850 servers were used in our testing. These servers were pre-configured with the A06 BIOS. Upgrading the BIOS to the latest version (A07 as of the time of this publication) resolved a raft of issues, including hanging reboot problems and networking issues.

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

Hyperthreading

Intel hyperthreading technology allows multithreaded 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 worse than performance without it. For this reason, EMC recommends disabling hyperthreading. There are two ways to disable hyperthreading: in the kernel or by the BIOS. Intel recommends disabling hyperthreading in the BIOS because it is cleaner than doing so in the kernel. Please refer to your server vendor's documentation for instructions.

Memory

EMC's Oracle RAC 10*g* testing was done with servers using two configurations with respect to memory:

- 8 GB of RAM
- 12 GB of RAM

On identical configurations with respect to the rest of the architecture, 12 GB provided far better scalability than 8 GB. This recommendation is for 64-bit environments; on 32-bit environments, the memory constraints are lower.

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 your 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 possible is entirely appropriate.

ASM

ASM is a storage-oriented solution which is supported and promoted by both EMC and Oracle. However, using ASM in conjunction with NFS is not an optimal solution for the following reasons:

1. Complexity. You must configure a large NFS file as a character device in order to use ASM over NFS. This must be linked into the /dev directory. These steps add complexity and management overhead to the configuration
2. Performance. When running ASM with NFS you are running a file system on top of a block device which is stored in a file system. In this case, the NFS file system is not helping you. However, you must still accept the CPU and other performance overhead of NFS. You get no benefit from this.

Thus, ASM should not be used with an NFS solution, such as the Celerra NS Series.

Linux setup and configuration

NFS mount point parameters

For optimal reliability and performance, the NFS client options in Table 6 are recommended. The mount options are listed in the /etc/fstab file.

Table 6. 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 is not responding. 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 ,wsize=32768	Always	This option sets the number of bytes NFS uses when reading/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 lock	nolock	Single instance only	This setting disables the lockd process.
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).
No attribute caching	noac	RAC only	This mount setting is needed for Oracle Database 10g quorum devices (voting Disks and OCR) to avoid corruption issues. It is also required for datafile and online redo log file volumes.
Timeout	timeo=600	Always	This sets the time (in tenths of a second) the NFS client is going to wait for the request to complete.

Protocol overhead

Typically, in comparison to 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. In such environments, protocol CPU consumption does not affect transaction response times. EMC testing has confirmed this.

Kickstart installation and required rpms

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, one can create a single file 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.

Here are the steps for kickstart installation:

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.

The "Appendix" 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, these packages, and no others, should be installed 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- 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- 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 Celerra Network Server, we recommend the initialization options in Table 7 contained in the spfile or init.ora file for the Oracle instance.

Table 7. 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. The NFS rsize and wsize should also be a multiple of this value. Under no circumstances should the NFS rsize and wsize be smaller than the database block size.	
Direct I/O	FILESYSTEM_IO_OPTIONS=directio
This setting enables direct 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.	
Multiple database writer processes	DB_WRITER_PROCESSES=2*n
Oracle currently does not support async I/O over NFS. To compensate for performance loss due to disabling asynchronous I/O, EMC recommends enabling multiple database writers. Set this to a value 2*n, where n is the number of CPUs.	
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 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.	

Recommendation for control file and log files

Control files

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

EMC recommends that your database has a minimum of two control files located on separate physical disks. One way to multiplex your control files is to store a control file copy on every disk drive that stores members of the redo log groups, if the redo log files are multiplexed.

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 failure of the database being able to recover. The best practice to multiplex your online redo log

files is to place members of a redo log group on different disks. To understand how redo log and archive log files can be placed, refer to the architecture diagram.

Basic backup and recovery

The best practice for backup of Oracle Database 10g is to perform approximately six logical storage backups per day, on a four-hour interval, 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 SnapSure checkpoint backups within the RMAN catalog. Refer to *Oracle Database 10g/Oracle RAC 10g Celerra NS Series NFS Applied Technology Guide* for more information on how to configure this solution. 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 RMAN using a full strategy otherwise. Refer to *Oracle Database 10g/Oracle RAC 10g Celerra NS Series NFS Applied Technology Guide* for more information. Further, EMC recommends that the RMAN backup be to an ATA disk configuration rather than to tape.

Mean time to recovery is optimized by this approach. You have the ability to restore instantly from a SnapSure checkpoint 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 that you need to restore from physical backup, the use of ATA disk will improve restore time.

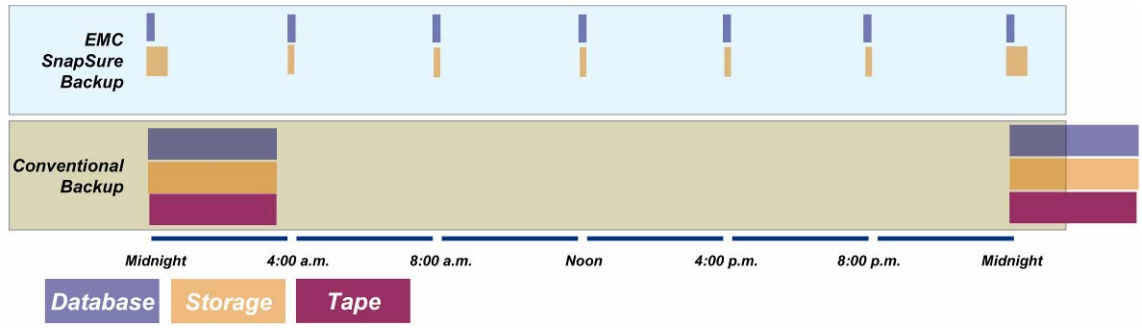

Figure 6. Multiple restore points using EMC SnapSure

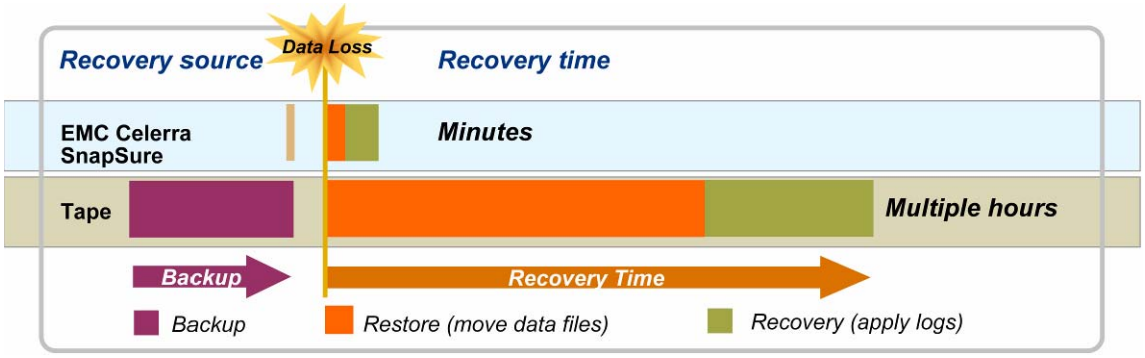

Figure 7. Rapid restore and recovery

Comparing logical storage backup and flashback database

With Oracle Database 10*g*, 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. 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-weight than 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 8, the best practice for disaster recovery of an Oracle Database 10*g* over NFS is to use Celerra fs_copy for seeding the disaster recovery copy of the production database, and then use 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 *Oracle Database 10g/Oracle RAC 10g Celerra NS Series NFS Applied Technology Guide*.

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

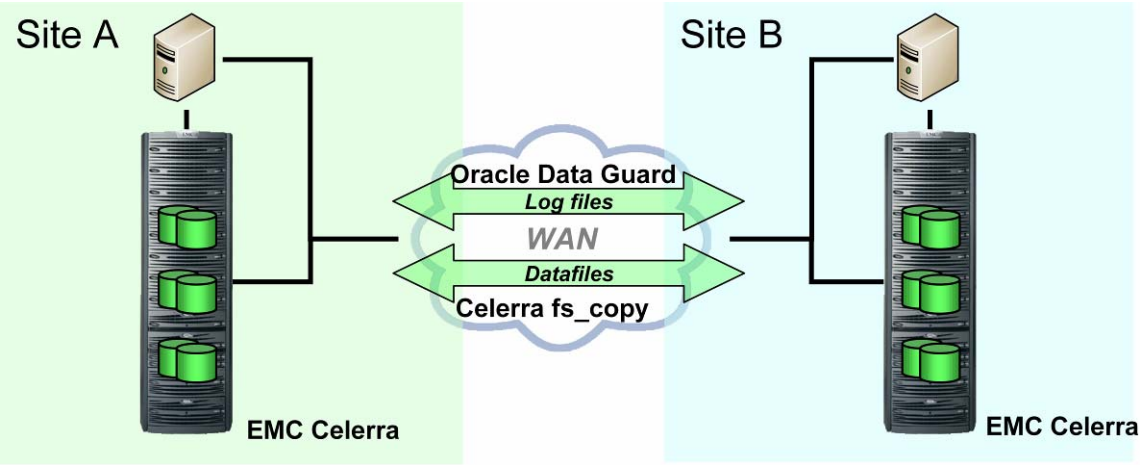

Figure 8. Remote disaster recovery for business protection

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 of the 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 writable clone of a production Oracle Database 10*g* over NFS is to take a hot backup of the database using Celerra 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 make an incremental clone for all other iterations, copying only the changed data in order to update the clone. This is recommended for larger databases as well as for ongoing or continuous need to clone the production database. Refer to *Oracle Database 10g/Oracle RAC 10g Celerra NS Series NFS Applied Technology Guide* for detailed steps on both of these methods.

Managing and monitoring Celerra Network Server

Celerra Manager

The Celerra Manager is a web-based graphical user interface (GUI) for remote administration of a Celerra Network Server. Various tools within Celerra Manager provide the ability to monitor the Celerra Network Server. 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 Oracle Enterprise Manager monitoring plug-in for EMC Celerra Network Server.

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 changed
- Centralize all of the monitoring information in a single console
- Enhance service modeling and perform comprehensive root cause analysis

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

Conclusion

The Celerra Network Server's high-availability features combined with EMC's proven storage technologies provide a very attractive storage system for Oracle Database 10*g* over NFS. Specifically:

- It simplifies database installation, cloning, migration, backup and recovery.
- It simplifies Oracle RAC 10g configuration by providing NFS-mounted shared file systems.
- The Data Mover failover capability provides uninterruptible 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 SnapSure and fs_copy not only provide data protection and rapidly available backups, but also simplify database migration and cloning.

- 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 Oracle Database 10*g (*Single Instance and RAC*)* with Celerra over NFS provides the best availability, scalability, manageability, and performance for your database applications.

References

The resources listed below have more information about EMC Celerra Network Server and Oracle.

- On EMC Powerlink (http://Powerlink.EMC.com):
  - *Managing Celerra Volumes and File Systems Manually*
  - *Oracle Database 10g/Oracle RAC 10g Celerra NS Series NFS Applied Technology Guide*
  - Celerra Network Server documentation
- The Oracle Technology Network
- Metalink, the Oracle support website

Appendix

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 --enableshadow --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
```