Backup and Recovery Performance and Best Practices for Exadata Cell and the Oracle Exadata Database Machine

Oracle Database Release 11.2.0.1 and Earlier

Oracle Maximum Availability Architecture White Paper
February 2011

Maximum Availability Architecture

Oracle Best Practices For High Availability

ORACLE

Executive Overview

The strategic integration of Oracle Exadata Database Machine and Oracle Maximum Availability Architecture (MAA) operational and configuration best practices (Exadata MAA) provides the best and most comprehensive Oracle database availability solution.

One of the key operational aspects of deploying an Exadata Database Machine is to ensure that database backups are performed and Oracle Database can be restored if disaster strikes. This white paper is based on Oracle Database release 11.2.0.1 and earlier releases, and describes the best practices for setting up the optimal backup and recovery strategy to protect your mission-critical data.

This paper provides:

- Key performance observations and rates
- Recommended tape-based backup solution
- Tape-based configuration best practices
- Recommended disk-based backup solution
- Disk-based configuration best practices
- Restore and recovery best practices

Introduction

Oracle Database has very sophisticated and scalable backup technologies. These technologies work especially well on Exadata Database Machine with its high bandwidth InfiniBand network and Exadata Storage Server Grid. Combined, these Oracle technologies provide database-aware storage services, such as the ability to offload database backup processing from the database server to storage, while remaining transparent to SQL processing and database applications. A case in point is the Exadata Database Machine Full Rack configuration described in this paper, for which tape-based backups achieved over 7 TB/hour for full backups and an effective backup rate of 10 to 70 TB/hour for incremental backups. Higher full backup rates are possible by using more media servers and tape drives.

For disk-based backups, full image copies achieved over 7 TB/hour and an effective backup rate of 10 to 48 TB/hour was attained for incremental backups.

Database restore rates achieved over 24 TB/hour, and database redo apply can attain 2.1 TB/hour (637 MB/sec).

These rates are achieved with a database CPU utilization of less than 10% on the targeted database servers, leaving plenty of CPU bandwidth for concurrent user workloads.

To put these backup, restore, and redo apply rates in perspective, a 30 TB database which consumes 100 TB of data, when mirroring and temp are accounted for, can be backed up in less than 5 hours when performing a full backup or in less than 1 hour for an incremental backup. Restoring the same complete backup can take less than 2 hours and redo apply can recover 2 TB of redo in an hour. On Exadata Database Machine, the factor that limits tape backup and restore rates is typically the hardware outside the Database Machine, not the Database Machine or the Oracle backup and recovery software.

The following technologies help to attain these backup rates:

- Exadata Database Machine is a complete package of software, servers, storage, and networking for all data management, including data warehousing, transaction processing, and consolidated mixed application workloads:
  - The InfiniBand fabric provides an extremely high performance network for transferring backup data from storage servers to database servers and then to tape media servers, and vice versa.
  - Not only are the transfer rates high, but the overhead on CPU resources used for the InfiniBand network transfers is very low compared to other solutions.
  - Oracle Exadata Storage Server Software (Exadata Cell) has very highly optimized disk I/O capabilities. Each Exadata Cell can achieve a disk transfer rate of over 1,500 MB/sec.
  - Exadata Cell has offload capabilities that further speed up incremental backups.
  - Its offload capability combines with RMAN block change tracking to efficiently perform large I/Os at the storage-tier level, returning only individual changed blocks for incremental backups and increasing the backup performance of the system.
  - Exadata Cell is a storage product optimized for use with Oracle Database applications. It is the storage building block of Exadata Database Machine. Exadata delivers outstanding I/O and SQL processing performance for online transaction processing (OLTP), data warehousing (DW), and a consolidation of mixed workloads. Exadata Cell is also integrated with other Oracle features including Oracle RAC, Oracle Data Guard, Oracle Flashback technologies, and Oracle ASM.
  - Exadata Cell and native Oracle Database compression capabilities can significantly reduce the overall database size and the I/O.
  - OLTP table compression can provide compression rates with a factor of 2x to 4x, such that a table normally requiring 100 GB of disk space would require only 25 GB to 50 GB of disk space. Exadata Hybrid Columnar Compression (EHCC) can provide compression

rates with a factor of 10x to 50x. Thus, the same table that requires 100 GB disk space could require only 2 GB to 10 GB of disk space, depending on the data. The rates described here are after compression. This means that the user data backup and restore rates can be an order of magnitude higher than the physical rates described here. For example, when 10x EHCC compression of user data is taken into account, the backup rate goes from 7 TB/hour of physical data to 60 TB/hour of user data.

- Oracle Recovery Manager (RMAN) provides the native backup and recovery infrastructure within the Oracle database, enabling optimized data protection in Exadata environments:
  - Backup, restore, and recover operations are performed using standard RMAN commands.
  - Backup and recovery combined with Exadata Cell provides a data aware and comprehensive solution to:
    - Prevent block corruptions from being written to disk.
    - Allow automatic and manual repair options using the Oracle ASM automatic repair mechanism, the Exadata storage built-in corruption checks, and the Oracle generic block corruption parameters.
  - RMAN can parallelize backup operations across all database nodes and Exadata storage cells. This allows all the disks, all the network connections, and all the CPUs in the system to contribute to performing backup operations.
  - Block change tracking allows incremental backups to run very quickly and efficiently. With block change tracking, only the areas of the database that have been modified since the last incremental backup or full backup are read from disk.
- Oracle Secure Backup is a centralized tape backup management solution for the entire IT environment including file system and the Oracle Database. With built-in RMAN integration, Oracle Secure Backup delivers the fastest Oracle Database backup to tape. Some important backup optimizations are available only with Oracle Secure Backup and RMAN:
  - Unused block compression eliminates the time and space usage needed to backup blocks that are allocated to tablespaces but are not currently used by tables.
  - Undo optimization eliminates the time and space usage needed to back up undo that is not required to recover the current backup.
  - These optimizations can provide substantial savings in backup time and tape costs.

Key Performance Observations and Rates

This paper describes the backup and recovery performance testing of Exadata Database Machine Full Rack, Half Rack, and Quarter Rack across various configurations:

- The disk backup and restore testing was performed with image copy formats using a fast recovery area located on Exadata Cell, and using differing degrees of RMAN parallelism.
- The tape backup and restore testing was performed using Oracle Secure Backup Release 10.3, with 2 media servers attached to a Sun StorageTek SL500 tape library and 14 LTO-4 tape drives via a Storage Area Network (SAN).

The following sections provide tables that summarize the performance results for disk-based and tape-based backup and restore performance.

Performance Results for Disk-Based Backup and Restore

TABLE 1 : SUMMARY OF MEASURED DISK-BASED BACKUP AND RESTORE PERFORMANCE

FULL DATABASE BACKUP TO DISK USING IMAGE COPIES[1]			
Instances and Channels	**Quarter Rack**	**Half Rack**	**Full Rack**
1 instance, 2 RMAN channels	1003 MB/sec or 3.4 TB/hr	1003 MB/sec or 3.4 TB/hr	1003 MB/sec or 3.4 TB/hr
2 instances, 2 RMAN channels per instance	1418 MB/sec or 5.1 TB/hr	1897 MB/sec or 6.5 TB/hr	1897 MB/sec or 6.5 TB/hr
2 instances, 4 RMAN channels per instance	1589 MB/sec or 5.4 TB/hr	2081 MB/sec or 7.1 TB/hr	2081 MB/sec or 7.1 TB/hr
FULL DATABASE INCREMENTAL BACKUP TO DISK (10% CHANGE) USED DAILY			
2 instances, 2 RMAN channels per instance	Measured effective backup rate 10 to 46 TB/hr, depending on workload		
FULL DATABASE RESTORE FROM DISK			
All available instances, 2 RMAN channels per instance	1644 MB/sec or 5.6TB/hr	3988 MB/sec or 13.7TB/hr	7000 MB/sec or 24TB/hr

For disk-based backups and incremental backups, use two database instances, and two or four RMAN channels per instance for optimal performance. For the two database instances designated as backup servers, less than 10% CPU and less than 40% I/O bandwidth were used.

[1] Note that a full database backup to disk using image copies only occurs once, because subsequent disk backups are performed using incremental backup and merge.

For database restores, all database instances were used if existing database files were available to avoid the initial data file allocations before copying the data from backup. For this restore case, the CPU utilization was less than 5%. For a restore operation where the data files do not pre-exist, then restore rates will be comparable to the backup rates shown in Table 1, and the recommendation is to use two database instances and two or four RMAN channels per instance for optimal restore performance.

Performance Results for Tape-Based Backup and Restore

TABLE 2 : SUMMARY OF TAPE-BASED BACKUP AND RESTORE PERFORMANCE

FULL DATABASE BACKUP TO TAPE			
Instances and Tape Drives	**Quarter Rack**	**Half Rack**	**Full Rack**
All instances, 14 tape drives, 1 RMAN channel per tape drive	2509 MB/sec or 8.6 TB/hr or 179 MB/sec per tape drive	2509 MB/sec or 8.6 TB/hr or 179 MB/sec per tape drive	2509 MB/sec or 8.6 TB/hr or 179 MB/sec per tape drive
FULL DATABASE INCREMENTAL BACKUP TO TAPE (10% CHANGE)			
All instances, 14 tape drives, 1 RMAN channel per tape drive	Measured effective backup rate 10 to 70 TB/hr		
FULL DATABASE RESTORE FROM TAPE			
All instances, 14 tape drives, 1 RMAN channel per tape drive[2]	1800 MB/sec or 6.1 TB/hr or 128 MB/sec per tape drive	2271 MB/sec or 7.8 TB/hr or 162 MB/sec per tape drive	2271 MB/sec or 7.8 TB/hr or 162 MB/sec per tape drive

In Table 2, the limiting factor for tape-based backup and restore performance was the number of tape drives, except for a Quarter Rack tape restore. For all tape-based backup and restore operations into existing files, CPU utilization on the database nodes was less than 10% and on the media server nodes it was less than 20%. We also recommend using all available database instances for these tape-based backup and restore operations.

[2] It is expected that the restore rate achieved for the Database Machine Full Rack is also applicable for a Database Machine Half Rack. However, this rate should not be expected for a Database Machine Quarter Rack; the Quarter Rack rate will be less because it is limited by the number of disks.

The database restore rates in Table 2 are based on restore operations into existing files for Exadata Database Machine Full Rack. For a restore operation where the data files do not pre-exist, then restore rates will be comparable to the backup rates shown in Table 1, assuming the number of tape drives and media servers is sufficient. For these restore operations, the recommendation is to use two database instances and two or four RMAN channels per instance for optimal restore performance.

Estimating Tape Backup Throughput Rates

Answer the following questions and then use the worksheet examples shown in Table 3 to quickly evaluate potential architectural bottlenecks and assess what tape backup architecture is necessary to meet your requirements.

Questions:

- What is the slowest network segment between Exadata Database Machine and the tape libraries?
- What is the media server ingestion or transfer rate?

  (Media Servers are described in "Media Management Software for Tape Backup")
- What is the tape drive transfer rate?
- How many tape drives do you plan to use?
- What is the expected tape compression ratio?

Table 3 shows tape-based backup flow, and highlights the throughput of each component:

- The content shown in **bold** typeface depicts example data. However, you can use the values in the Estimated Rate column for each backup estimate.
- The Throughput Rate (in the last column of Table 3) is determined using the following equation:

  Quantity x Estimated Rate = Throughput Rate

TABLE 3 : EXAMPLE TAPE BACKUP WORKSHEET

FLOW	COMPONENTS	QUANTITY	ESTIMATED RATE (GB/SEC)	THROUGHPUT RATE (GB/SEC)
1	Exadata Cell	**14**	1[3]	**14**
2	Database Server	**8**	2.0[4]	**16**
3	Network to Media Server:			
	a) Media Server InfiniBand HCA using TCP/IP	**2**	2	**4**
	b) Media Server GigE NICS	4	0.12	**0.48**
4	Media Server to Tape Library SAN[5] Links	4	0.8	**3.2**
5	Tape Drives	**14**	**0.17**	**2.3**
			Use 1 GB/sec for SAS or 700 MB/sec for SATA-based Exadata Cell	

For example, using the throughput rates shown in the last column of Table 3, the following bottlenecks will be reached first if using InfiniBand fabric to media server:

1. Media server to tape library SAN transfer rate
2. Number of tape drives

To increase the backup rate, add more host bus adaptors (HBAs) to the media server or tape library, and increase the number of tape drives.

[3] SAS-based Exadata Cell supports up to 1.5 GB/sec data bandwidth. SATA-based Exadata Cell supports up to 0.85 GB/sec data bandwidth. Measured values were used for the estimates shown in Table 3.

[4] Each server has Dual-ported Quad Data Rate (QDR) InfiniBand Host Channel Adapters (HCAs) (40 Gb/s). However, while performing backups over InfiniBand using the TCP/IP protocol, the achievable throughput is approximately 2 GB/sec.

[5] Storage Area Network (SAN).

If using Gigabit Ethernet (GigE) network between the media server and Exadata Database Machine as shown in the row labeled "3b) Media Server GigE NICS" in Table 3, then the Media Server GigE NICs throughput rate will be your main bottleneck.

Media Management Software for Tape Backups

RMAN optimizes Oracle Database backups to disk and RMAN is integrated with media management software for backup to tape. Media management software[6] is the software layer that facilitates RMAN backups to tape.

Oracle Secure Backup, the media management software used for the testing described in this white paper, is highly scalable with a client/server architecture in which all hosts in the backup domain have one or more Oracle Secure Backup roles. The roles include:

- **Administrative server** is the host that stores configuration information and catalog files for hosts in the administrative domain. A backup domain includes only one administrative server. One administrative server can service every client on your network. The administrative server runs the scheduler which starts and monitors backups within the administrative domain.
- **Media servers** are computers or servers to which at least one tape device is connected. A media server is responsible for transferring data to or from the tape devices that are either directly attached or Storage Area Network (SAN) attached.
- **Client** is any computer or server whose files Oracle Secure Backup backs up or restores. All computers to be backed up within the domain will be assigned the client role during installation along with additional roles as defined by the user, including the media server and/or administrative server roles.

Any computer in the domain can act as the administrative server. In practice, it is common to have an additional computer host the Oracle Secure Backup catalog, RMAN catalog, and the Oracle Enterprise Manager repository.

Backing up and restoring file system data is not addressed in this paper beyond protecting the Oracle Secure Backup catalog. For more information about configuring file system backup and restore, see the Oracle Secure Backup documentation at http://download.oracle.com/docs/cd/E14812_01/doc/doc.103/e12834/toc.htm

[6] The media management software described in this paper is Oracle Secure Backup. However, the configuration best practices are applicable to other media management software products.

Tape-Based Backup Strategy

Some of the key benefits to tape-based backup strategies include:

- Exadata Storage and tape-based backups provide extremely fast backup and restore rates.
- Tape-only solutions isolate faults from Exadata Storage Servers.
- Exadata Database Machine capacity and bandwidth are maximized.

However, disk-based backup may be a better solution if you require better recovery times for data and logical corruptions and for some Tablespace Point in Time (TSPITR) scenarios. For more information about disk-based backup solutions, see the "Disk-Based Backup Strategy" section in this white paper.

For tape-based backup solutions, the recommended strategy is to perform the following backups:

- Weekly RMAN level 0 (full) backups of the database
- Daily cumulative RMAN incremental level 1 backups of the database
- Daily backups of the Oracle Secure Backup catalog

To scale backup rates using the InfiniBand fabric, while still maintaining high availability, use the following strategies:

1. Start with at least two media servers. There can be a total of two Host Channel Adapters (HCAs)—only one active port—per media server, bonded for high availability.
2. Add tape drives until the HCA of each media server's bandwidth is saturated.
3. Use all Oracle Database server instances for backups.
4. Use one RMAN channel per tape drive.

If you are using a Gigabit Ethernet (GigE) network between the media server and Exadata Database Machine, you will be bound by the number of GigE connections times 120 MB/sec. In Exadata Database Machine Full Rack, there are eight HA-bonded GigE ports available or one HA-bonded GigE port per database node.

Figure 1 shows the Oracle MAA recommended architecture.

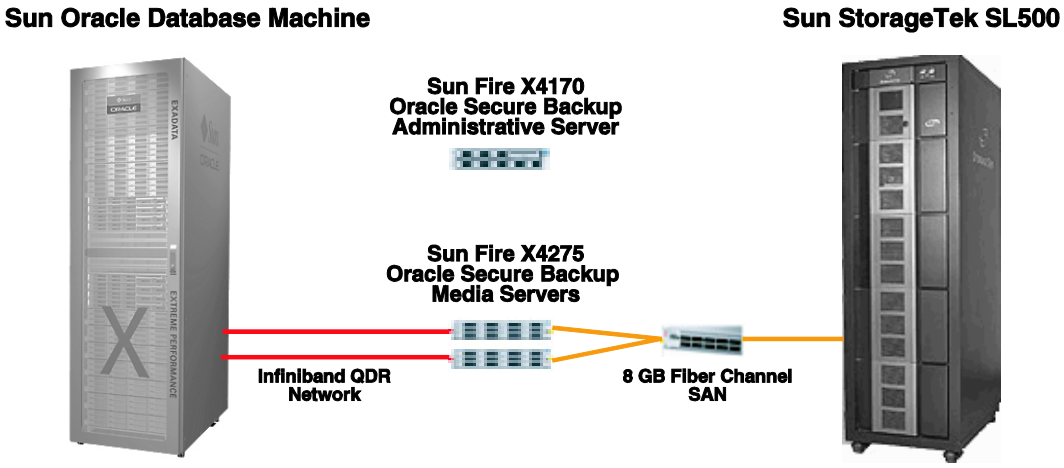


Figure 1 Oracle Exadata Database Machine Hardware System Configuration

Figure 1 represents a tape backup solution with two or more media servers. To connect media servers directly to the existing InfiniBand fabric, you can use the six HA-bonded ports available for Exadata Database Machine. For example, you can use ports 5B, 6A, 6B, 7A, 7B, and 12A for any size (Full Rack, Half, Rack, or Quarter Rack) Exadata Database Machine. Any other available ports should be reserved in the event that you want to upgrade Exadata Database Machine Half Rack to a Full Rack.

Each media server requires an InfiniBand QDR HCA or the recommended dual-ported InfiniBand QDR HCA. The network protocol used for backups over InfiniBand is the standard TCP/IP protocol, so it is transparent to the backup software on the database servers and the media servers. The backup software operates identically whether you use InfiniBand or a Gigabit Ethernet network.

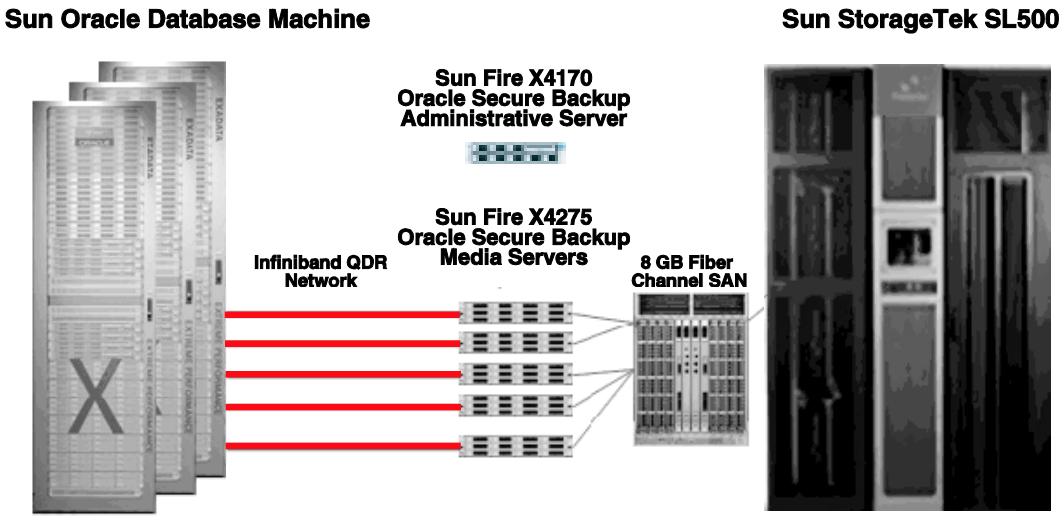


Figure 2: Larger Backup Configuration with Multiple Media Servers

Figure 2 has an effective maximum data transfer rate of 10 GB/sec, or 35 TB/hour. Allocate a sufficient number of tape drives so a media server can attain maximum backup and restore rates. For example, if a tape drive backup rate is 240 MB/sec of compressed data[7], then use at least 8 tape drives to achieve the maximum data transfer rate of one media server's HCA.

Figure 3 shows the recommended Exadata storage server layout for tape-based backup and recovery. In the figure, the faster (outer) 80% of the disk is assigned to the DATA area, and the slower (inner) 20% of the disk is assigned to the fast recovery area (RECO) area. This layout can be configured automatically during deployment.

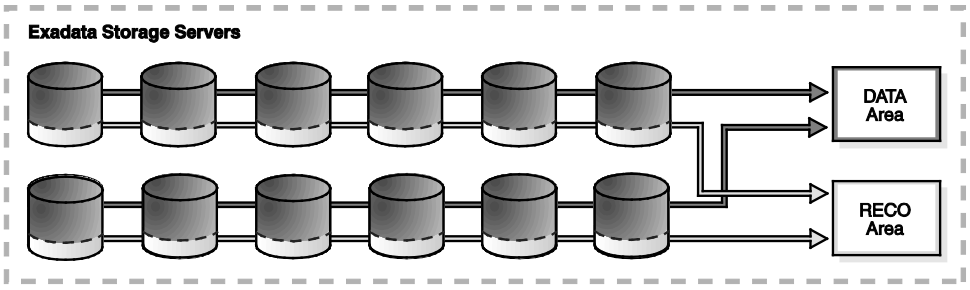


Figure 3: Exadata Storage Server Grid Disk Layout for Tape Based Backup and Recovery

[7] An LTO4 Tape Drive is capable of writing approximately 240 MB/s compressed data to tape, while an LTO3 Tape Drive is capable of writing approximately 160 MB/s of compressed data to tape.

Best Practices for Tape-Based Configurations

This section provides:

Database Configuration Best Practices

- **Use Oracle Secure Backup for low-cost, fast, and MAA-validated tape backups.**
  - Oracle Secure Backup provides the fastest database backup to tape due to its tight integration with RMAN.
  - If you are backing up to tape using Oracle Secure Backup, then the unused-block optimization capability is enabled. However, if the backup is made directly to tape using a third-party media management product, then this does not have any effect because the unused-block optimization is available only with Oracle Secure Backup.
- **Configure the Preferred Network Interface (PNI) to direct the Oracle Secure Backup traffic over the InfiniBand network interface.**

```
ob> lspni (List Preferred Network Interface)
mediaserver1:
    PNI 1:
        interface:    mediaserver1-ib
        clients:      dbnode1, dbnode2, dbnode3, dbnode4,
dbnode5, dbnode6, dbnode7, dbnode8
    PNI 2:
        interface:    mediaserver1
        clients:      adminserver
```

- **Configure one RMAN channel per tape drive and add tape drives to scale backup rates.**

  Typical tape drive backup rates are between 100 MB/sec and 240 MB/sec, depending on the drive type and compression options. Note that tape drive compression becomes less effective when backing up tables that are compressed at the database level. Backup performance scales when you add more tape drives and RMAN channels, assuming there is available throughput on the media server.

A single RMAN channel in Exadata Database Machine can stream data at a rate of 749 MB/sec from Oracle Database Machine to the media server. Therefore, backup performance is limited by tape drive throughput.

- **Configure Gigabit Ethernet or InfiniBand.**
  - Use Gigabit Ethernet for smaller databases:
    - Using a dedicated interface for the transport eliminates the impact to the client access network. Typically, a dedicated backup network is in place. The maximum throughput with the Gigabit Ethernet network is 120 MB/sec times the <*Number of Database Servers*>.
    - For a Exadata Database Machine Full Rack, it is possible to achieve throughput as high as 960 MB/sec.
    - For more information, see the "Configuring the Gigabit Ethernet (GigE) Network to Media Server" section in this white paper.
  - Use InfiniBand for the best backup rates, especially targeting larger databases that require fast backup rates and low CPU overhead.

    The graph in Figure 4 shows the CPU Utilization from one of the database nodes involved in the backup using Oracle Secure Backup when running over InfiniBand Network.

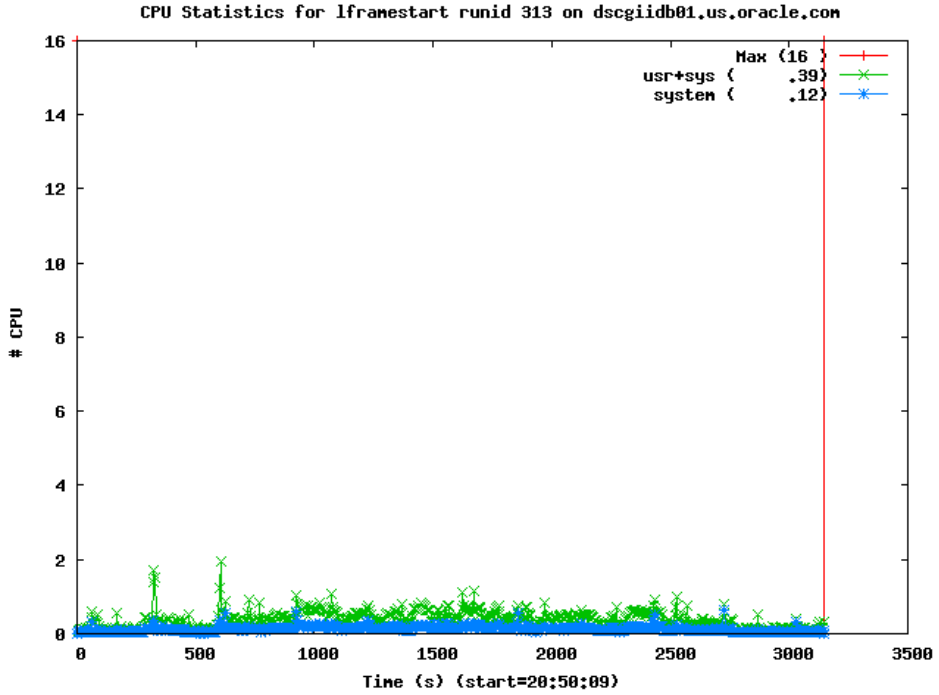


Figure 4: Exadata Storage Server Grid Disk Layout for Tape Based Backup and Recovery

For more information, see the "Configuring InfiniBand Network to Media Server" section in this white paper.

- **Configure Oracle RAC Service for backup running on all database instances.**

  CPU utilization on the database and media server nodes was less than 10% when spread across all database instances.

- **Use SQL*Net service load balancing to distribute RMAN channels evenly among the allocated instances.**

  1) Create a service that runs on the selected nodes in the cluster:

     ```
     srvctl add service -d <dbname> -s <service name> -r
     <instance1>,<instancen>
     ```

  2) When running RMAN, use the service name in the connect string for the "target" parameter:

     ```
     rman target sys/<password>@<scan_address>/<service_name> catalog …
     ```

- **Use RMAN incremental backups and block change tracking.**

  Enable block change tracking to achieve fast incremental backups. Block change tracking allows RMAN to avoid scanning blocks that have not changed when creating incremental backups. Also, when performing incremental backups of databases on Exadata Cell, additional block inspection is offloaded from the database servers.

  Block change tracking provides the greatest benefit for databases where fewer than 20% of the blocks are changed daily. You may still benefit by using block change tracking with change rates greater than 20% but testing is recommended to ensure that backup times are reduced.

- **Set `DB_RECOVERY_FILE_DEST_SIZE` to bound space in the fast recovery area.**

  The database writes archived redo log files and any additional recovery files to the fast recovery area. These include any disk backup files such as level 0 image copies and level 1 backup sets as well as Flashback log files (if enabled). It is important that the value of the `DB_RECOVERY_FILE_DEST_SIZE` parameter be set to less than the total free space in the disk group and the free space should take account of at least one disk failure and preferably one Exadata Cell failure.

  Additionally, if multiple databases are sharing the fast recovery area, you must ensure that the sum of the space allocated to the different databases is less than the free space in the disk group.

- **Use an external RMAN recovery catalog.**

  See the *Oracle Database Backup and Recovery User's Guide* for more information about the RMAN repository.

- **Consider using I/O Resource Management or Database Resource Manager to manage system resources on Exadata Database Machine.**

  If Exadata Database Machine's CPU or I/O resources must be prioritized between application workload and backups, then use I/O Resource Manager or Database Resource Manager. See the *Oracle Exadata Storage Server Software User's Guide* for instructions about setting up and configuring I/O Resource Manager.

- **Tune the network communication if you are using a third-party media management vendor.**

  If you are using a non Oracle media management vendor, then contact the vendor for their configuration best practices. Most vendors test and validate their own products with Exadata and Exadata Database Machine and can recommend how to exploit the full potential of the InfiniBand or GigE networks. Note that there is no special certification for RMAN and the Media Management Vendor (MMV) to work with Exadata. The MMV is required to certify with Oracle Database 11*g* release 2 (11.2) and Oracle Enterprise Linux. Exadata is not a key in this certification.

RMAN Configuration Commands and Backup Scripts

- Use the following configuration commands to parallelize backups across all database nodes, allowing all the disks, network connections, and system CPUs to be leveraged for increased performance.

RMAN configuration changes:

```
CONFIGURE DEFAULT DEVICE TYPE TO SBT;
CONFIGURE DEVICE TYPE SBT PARALLELISM <# Of Tape Drives>;
```

- Use the following backup scripts to automate weekly and daily backups.

RMAN script for weekly backup:

```
run {
  backup incremental level 0 database;
  backup archivelog all not backed up;
   }
```

RMAN script for daily backup:

```
run {
  backup cumulative incremental level 1 database;
  backup archivelog all not backed up;
   }
```

Configuring InfiniBand Network to Media Server

With the available InfiniBand ports in Exadata Database Machine, media servers can be directly connected to the InfiniBand fabric by adding an InfiniBand Quad Data Rate (QDR) host channel adapter (QDR HCA) to the media server. For high availability, connect the HCA to two different database machine InfiniBand switches to eliminate the switch as a single point of failure. This provides transparent failover if connectivity is lost to one of the ports.

Follow these best practices.

Note: The examples included with each bullet are based on a media server running on the Linux operating system.

- **Configure bonding of the InfiniBand interfaces on the media server.**

  The following example of bonding `ib0` and `ib1` is specific to a Linux environment:

  1. Modify the /etc/modprobe.conf file to add the following two lines to the bottom of the file. This adds another bonding alias and options.

     ```
     alias bond1 bonding
     options bonding max_bonds=2
     ```

     The file will be similar to the following example. This file assumes bonding was previously established on `bond0`:

     ```
     alias eth0 tg3
     alias scsi_hostadapter cciss
     alias scsi_hostadapter1 ata_piix
     alias scsi_hostadapter2 usb-storage
     alias ib0 ib_ipoib
     alias ib1 ib_ipoib
     alias bond0 bonding
     alias bond1 bonding
     options bonding max_bonds=2
     ```

  2. Create the `/etc/sysconfig/network-scripts/ifcfg-bond1` file, as follows.

     ```
     DEVICE=bond1
     USERCTL=no
     BOOTPROTO=none
     ONBOOT=yes
     IPADDR=<IP Address for bond1 within the same subnet as the
             existing InfiniBand network >
     NETMASK=<Netmask must be the same as the existing InfiniBand
             network >
     NETWORK=<Network calculated using ipcalc-n
     ```

```
    ip_address netmask>
GATEWAY=<Gateway IP address>
BONDING_OPTS="mode=active-backup miimon=100
    downdelay=5000 updelay=5000"
IPV6INIT=no
```

3. Make copies of the current `ib0` and `ib1` configuration files. Ensure the copied files do not start with `ifcfg-ib0`. Prefix the file name with `backup-` or a similar word, and do not add a suffix such as `-backup`. For example:

```
cd /etc/sysconfig/network-scripts/
cp ifcfg-ib0 backup-ifcfg-ib0
cp ifcfg-ib1 backup-ifcfg-ib1
```

4. Modify the current `ib0` and `ib1` configuration files so they are configured to act as slaves to the bond1 interface. The files should appear as follows:

```
* File ifcfg-ib0:
DEVICE=ib0
USERCTL=no
ONBOOT=yes
MASTER=bond1
SLAVE=yes
HOTPLUG=no
BOOTPROTO=none
MTU=65520


* File ifcfg-ib1:
DEVICE=ib1
USERCTL=no
ONBOOT=yes
MASTER=bond1
SLAVE=yes
HOTPLUG=no
BOOTPROTO=none
MTU=65520
```

5. Restart the system.

6. Log in as the `root` user after the system restarts to verify that NIC bonding is running correctly.

```
# cat /proc/net/bonding/bond1
Ethernet Channel Bonding Driver: v3.2.3 (December 6, 2007)
```

```
Bonding Mode: fault-tolerance (active-backup) (fail_over_mac)
Primary Slave: None
Currently Active Slave: ib0
MII Status: up
MII Polling Interval (ms): 100
Up Delay (ms): 5000
Down Delay (ms): 5000

Slave Interface: ib0
MII Status: up
Link Failure Count: 1
Permanent HW addr: 80:00:00:48:fe:80

Slave Interface: ib1
MII Status: up
Link Failure Count: 1
Permanent HW addr: 80:00:00:49:fe:80
```

- **Update OpenFabrics Enterprise Distribution on the media server.**

  You must use an OpenFabrics Enterprise Distribution (OFED) version that is compatible with the version found in Exadata Database Machine in the media server. You can download the OFED from My Oracle Support Note 888828.1.

- **Configure InfiniBand IPoIB connected mode for best performance.**

  No changes are required to the database servers of Exadata Database Machine running Exadata 11*g* Release 2 (11.2.0.1) and later. However, for custom configurations, you must evaluate the following settings.

  The following commands assume a Linux operating system.

  1. Verify that Connected Mode is enabled on the system, as follows:

     ```
     # cat /sys/class/net/ib0/mode
     connected
     # cat /sys/class/net/ib1/mode
     connected
     ```

     If the status is "Datagram," then proceed to step 2 and step 3.

  2. Edit the `/etc/ofed/openib.conf` file and search for `SET_IPOIB_CM` and change its value to specify "`yes`":

     ```
     # Enable IPoIB Connected Mode
     SET_IPOIB_CM=yes
     ```

  3. Reboot the server and re-verify the connected mode again, following the instructions in step 1.

- **Configure `MTU Size=65520` on InfiniBand for faster data transmission.**

  No changes are required to the database servers of Exadata Database Machine running Exadata 11*g* Release 2 (11.2.0.1) and later releases. However, for custom configurations, you must evaluate the following settings:

  1. Edit the `/etc/sysconfig/network-scripts/ifcfg-ib*` and the `/etc/sysconfig/network-scripts/ifcfg-bond0` files to add an entry for `MTU=65520`. For example:

     ```
     MTU=65520
     ```

  2. Verify that the MTU size is 65520, as follows:

     ```
     # ifconfig ib0 | grep MTU
       UP BROADCAST RUNNING SLAVE MULTICAST MTU:65520 Metric:1
     # ifconfig ib1 | grep MTU
       UP BROADCAST RUNNING SLAVE MULTICAST MTU:65520 Metric:1
     # ifconfig bond0 | grep MTU
       UP BROADCAST RUNNING MASTER MULTICAST MTU:65520 Metric:1
     ```

  3. Reboot the server and verify the MTU size again, following the instructions in step 1.

- **Configure the media server to use the InfiniBand network.**

  To direct the backup and restore traffic over the InfiniBand fabric, configure the media management software to favor InfiniBand. Note that each media management software type has its own method of enabling this configuration.

  For instance, Oracle Secure Backup has the concept of a **preferred network interface**, which can be set on the media server for a specific list of clients. Other media management software requires this configuration to be defined when the software is installed. See your media management software for information about how to direct traffic over a particular network.

Configuring the Gigabit Ethernet (GigE) Network to Media Server

When connecting the media servers to Exadata Database Machine through Ethernet, connect the `eth3` interfaces from each database server directly into the data center network. For high availability, multiple network interfaces on the database servers and multiple network interfaces on the media server can be bonded together. In this configuration, configure the `eth3` interface as the preferred or primary interface and configure `eth2` as the redundant interface.

If throughput is a concern, then connect both `eth2` and `eth3` interfaces from each database server directly into the data center's redundant network. The two interfaces can then be bonded together in a redundant and aggregated way to provide increased throughput and redundancy.

Follow these best practices:

- **Configure the Gigabit Ethernet switch configuration.**

  For optimal throughput and availability, configure hardware Link Aggregation in the gigabit switch. The Link Aggregation Control Protocol (LACP)[8] is defined as part of IEEE 802.1AX-2008 standard. Other software enabled bonding options are available within the operating system of the database servers and media server, which may also be used.

  If you are using LACP, then ensure that LACP is supported and configured on the Ethernet switch for Src XOR Dst TCP/UDP Port. See your vendor's Gigabit switch documentation for information about configuring source and destination port load balancing.

  On a Cisco 4948 switch, use the following commands to implement Src XOR Dst TCP/UDP Port:

```
swi-2(config)#port-channel load-balance src-dst-port
swi-2#wr mem
swi-2#sh etherchannel load-balance
EtherChannel Load-Balancing Operational State (src-dst-port):
Non-IP: Source XOR Destination MAC address
  IPv4: Source XOR Destination TCP/UDP (layer-4) port number
  IPv6: Source XOR Destination IP address
swi-2#
```

  Additionally, if you are using LACP, then when you configure the `ifcfg-bond1` file, change the `BONDING_OPTS` setting to `mode=4`.

- **Configure the database server Gigabit Ethernet.**

  No specific changes need to be made to the database servers. However, to obtain higher backup rates, create a Dual-Port Gigabit Ethernet Configuration. See the *Oracle Exadata Storage Server Software User's Guide* for information about bonding multiple interfaces together on Database Server Nodes (database nodes) in Exadata Database Machine.

  If you are using LACP, when configuring the `ifcfg-bond1` file, change the `BONDING_OPTS` parameter setting to `mode=4`.

[8] See http://en.wikipedia.org/wiki/Link_aggregation.

- **Configure the media server Gigabit Ethernet.**

  The following recommendations are applicable only for media servers running Oracle Enterprise Linux Version 5.3 (or later) or RedHat Enterprise Linux Version 5.3 (or later). If your specific media server is running a different operating system, contact your vendor for the appropriate Gigabit configuration.

  As with the database server Gigabit Ethernet configuration, no specific changes must be made to the media servers. However, to obtain higher backup rates, create a Multiple-Ported Gigabit Ethernet Configuration. The steps to configure bonding on the media server are the same as on the Database Servers. See the *Oracle Exadata Storage Server Software User's Guide* for a detailed procedure.

Configuring Persistent Bindings for Tape Devices

In SAN environments, you must configure persistent bindings so the device address does not change. If the device address changes, the media servers cannot access the device unless you update the device configuration within Oracle Secure Backup. Therefore, it is very important that your environment maintains consistent device addresses.

Persistent bindings are not configured within Oracle Secure Backup but they are a part of your infrastructure setup. You may configure persistent bindings through the HBA or the operating system. The configuration steps may vary by platform and vendor. See My Oracle Support Note: 971386.1 for an example of creating persistent bindings for device attachments.

Backing up the Oracle Secure Backup Catalog

The Oracle Secure Backup catalog maintains backup metadata, scheduling, and configuration details for the backup domain. Just as it is important to protect the RMAN catalog or control file, the Oracle Secure Backup catalog should be backed up on a regular basis. In Oracle Secure Backup, the catalog backup has been pre-configured:

- **Media family**: `OSB_Catalog_MF` writes all catalog backups to same tape or tapes.
- **Job summary**: `OSB-CATALOG-SUM` sends e-mail showing a daily report status of catalog backup to users.
- **Dataset**: `OSB-CATALOG-DS` defines all directories and files to backup for file system backups.
- **Schedule:** `OSB-CATALOG-SCHED` shows the schedule for the catalog backup.

The primary catalog backup configuration settings have been defined with only one step remaining which requires user intervention: Edit the `OSB-CATALOG-SCHED` triggers to specify when to perform the backup.

Disk-Based Backup Strategy

Some of the key benefits to a disk-based backup strategy include:

- Better recovery times for data and logical corruptions and for some Tablespace Point in Time (TSPITR) scenarios
- Ability to using backups directly with no restore

For disk-based backup solutions, Oracle recommends the following:

- Use a fast recovery area
- Perform an initial RMAN level 0 (full) backups
- Perform daily RMAN incremental level 1 backups
- Roll incremental backups into full backup and delay by 24 hours

To scale backup rates for disk:

1. Start with one instance and two RMAN channels.
2. Add another instance and two more RMAN channels for performance and availability.

Optimal backup rates were achieved with 2 database instances and 2 to 4 RMAN channels per instance. For the 2 database instances designated as backup servers, less than 10% of CPU and less than 40% of I/O bandwidth was used. During backup operations, I/O intensive parallel queries may want to avoid these designated backup servers.

Figure 5 shows the recommended Exadata Storage Server Grid disk group layout. In the figure, the faster (outer) 40% of the disk is assigned to the DATA area, and the slower (inner) 60% of the disk is assigned to the fast recovery area (RECO) area. This can be configured automatically during deployment.

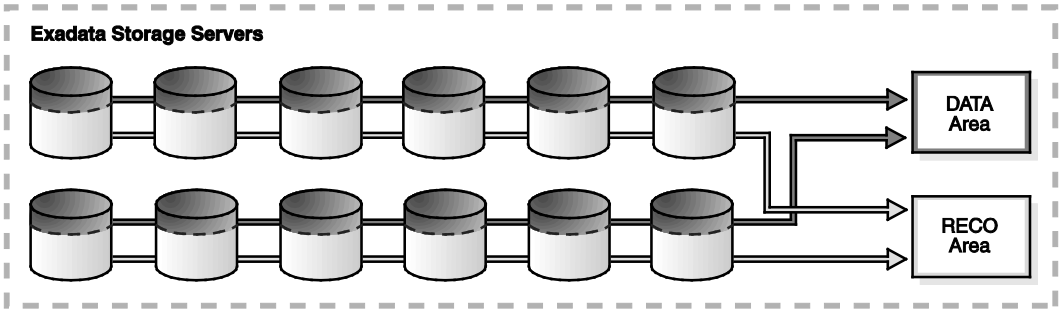
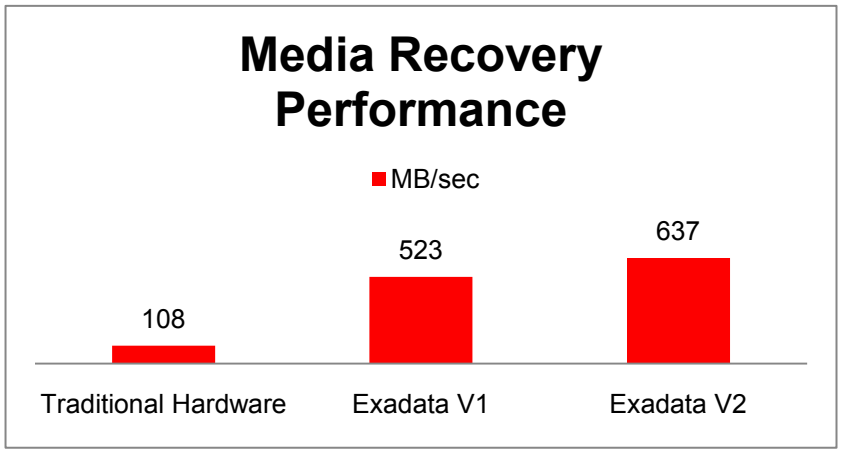


Figure 5: Exadata Storage Server Grid Disk Group Layout for Disk-Based Backup and Recovery

Another strategy is to purchase additional SATA Exadata storage specifically to store the fast recovery area. This allows the application to leverage the full Exadata Database Machine storage grid, allows the use of lower-cost storage for backups, and provides better failure isolation by

using separate backup hardware. To reserve more space and bandwidth for the DATA disk group, Oracle recommends using a tape-based backup solution, or at the very least, a hybrid approach where full database backups go to tape and incremental disk backups go to the fast recovery area.

When applying redo data on Exadata Database Machine, MAA testing achieved rates greater than 600 MB/sec while applying changes from ETL loads, and greater than 200 MB/sec for OLTP-type workloads. For example, our Beehive Collaborative Application applied redo records at 300 MB/sec and they are using Exadata.



Best Practices for Disk-Based Configurations

This section provides:

- Database Configuration Best Practices
- RMAN Commands and Configuration

Database Configuration Best Practices

- **Use RMAN incremental backups and block change tracking.**
  - To reduce backup time and resources, perform nightly incremental backups to the fast recovery area and merge them into the image copy backup on regular basis. If recovery is needed, then the copies can be directly used as normal data files and recovered to a consistent point, without the need for a restore operation, thus significantly reducing overall recovery time.
  - For fast incremental backups, enable block change tracking. Block change tracking allows RMAN to avoid scanning blocks that have not changed, when creating incremental

backups. Also, when performing incremental backups of databases on Exadata storage, additional block inspection is offloaded from the database servers.

Block change tracking provides the greatest benefit for databases where fewer than 20% of the blocks are changed daily. You may still benefit by using block change tracking with change rates greater than 20%, but testing is recommended to ensure that backup times are reduced.

- **Set the initialization parameter _file_size_increase_increment=2143289344**

  Set this parameter to optimize the space used when incremental (level 1) backups are taken on the fast recovery area.

- **Set the initialization parameter _backup_ksfq_bufsz=4194304**

  Set this parameter on HP Oracle Database Machine and Oracle Exadata Database Machine systems running Oracle Database release 11.2.0.1 and earlier releases. This parameter optimizes the RMAN read operations that occur during an RMAN backup.

  Note: This parameter should be reset or removed when the system is upgraded to release 11.2.0.2 or later releases.

- **Set the initialization parameter _backup_ksfq_bufcnt**

  Set this parameter on Oracle Exadata Database Machine systems running Oracle Database release 11.2.0.1 and earlier releases. This parameter controls the number of buffers and therefore memory used by individual RMAN channels during backup and restore operations.

  - On a quarter rack configuration this parameter should be set to 32.
  - On a Half Rack, Full Rack, or multiple rack configuration this parameter should be set to 64.

  Note: Do not set this parameter on HP Oracle Database Machine systems running release 11.2.0.1 and earlier releases.

  Note: This parameter should be reset or removed when the system is upgraded to release 11.2.0.2 or later releases.

- **Use a backup Oracle Net service for better performance and high availability.**

  Configure an Oracle Service to run against designated backup servers in the cluster. The service is used by the RMAN `BACKUP` command. RMAN automatically spreads the backup load evenly among target the instances offering the service. For example:

```
$ srvctl add service -d <db_unique_name> \
 -s <service_name> \
 -r <list of preferred instances>
$ srvctl start service -d <db_unique_name> \
 -s <service_name>
```

```
Connect to RMAN using the service name:
$ rman target sys/<sys_password>@<service_name>
```

- **Use two to four RMAN channels per instance.**

  In most cases, two RMAN channels per database server are sufficient. You may get some incremental gains with four channels per instance. For the highest throughput, allocate a total of eight RMAN channels. During backup operations, sufficient CPU resources are available for production usage because fewer than two CPU cores are used for each participating backup server if all eight RMAN channels are used. Listener load balancing distributes the connections between the two instances.

- **Set `DB_RECOVERY_FILE_DEST_SIZE` to bound space in the fast recovery area.**

  The database writes archived redo log files and any additional recovery files to the fast recovery area. These include any disk backup files such as level 0 image copies and level 1 backup sets as well as Flashback log files (if enabled). It is important that you set the value of this parameter to less than the total free space in the disk group, which must take account of at least one disk failure and preferably one Exadata Cell failure.

  Additionally, if multiple databases are sharing the fast recovery area, ensure that the sum of the space allocated to the different databases is less than the free space in the disk group.

- **Use an external RMAN recovery catalog repository.**

  See the *Oracle Database Backup and Recovery User's Guide* for more information about the RMAN repository.

- **Avoid using third-party storage solutions.**

  Exadata Database Machine and Exadata are fundamentally different from traditional third-party solutions implemented external to Oracle. Using third-party storage is not recommended because the performance and complexity will vary dramatically, and the MAA best practices may not apply to non Oracle storage solutions. However, if you are using third-party storage then:

  - Use IP-based protocols, such as iSCSI or NFS.
  - Use an intermediate server that acts as an iSCSI or NFS server with SAN-based storage.
  - When using a SAN host bus adapter (HBA), network rates may limit the backup rate in which case the best practice recommendation is to implement an intermediate server.

RMAN Configuration Commands and Backup Script

- Use the following configuration commands to parallelize backups across two database nodes allowing all the disks and network connections to be leveraged for increased performance. Use two or four RMAN channels per instance or a maximum of eight RMAN channels total.

RMAN configuration changes

```
CONFIGURE DEFAULT DEVICE TYPE TO DISK;
CONFIGURE DEVICE TYPE DISK PARALLELISM 8;
```

- Use the following backup script to automate RMAN backups and automation of applying the previous incremental backups.

RMAN script:

```
run {
  recover
     copy of database
     with tag 'Disk_Backup';
  backup
     incremental level 1
     for recover of copy
     with tag 'Disk_Backup'
     database;
   }
```

Restore and Recovery Best Practices

Restore rates of up to 23 TB/hour were achieved when avoiding the initial data file allocations. If existing files are present prior to the need to execute a restore database operation, do not delete the data files but issue the restore database command to take advantage of this optimization. The following sections describe how to:

- Restore into existing data files
- Restore into a new Oracle ASM disk group

Restore into Existing Data Files

With pre-existing files, create a restore service that runs across all database instances and use two RMAN channels per database instance for Half Rack and larger, and use four RMAN channels per database instance for Quarter Rack. If no existing data files are present prior to the restore operation, create a restore service with only two database instances and use two to four RMAN channels per database instance. For tape-based restores, the number of channels is restricted by the number of tape drives.

For example,

1. Create an Oracle service for restore for all available database nodes.

   The service is used by the RMAN restore command. RMAN automatically balances the restore load among the targeted instances. For example:

   ```
   $ srvctl add service -d <db_unique_name> \
   ```

```
 -s <service_name> \
 -r <list of preferred instances>
$ srvctl start service -d <db_unique_name> \
 -s <service_name>
```

2. Connect to RMAN using the service name:

   ```
   $ rman target sys/<sys_password>@<service_name>
   ```

   - For tape-based channels = 14 tape drives

     ```
     CONFIGURE DEFAULT DEVICE TYPE TO SBT;
     CONFIGURE DEVICE TYPE SBT PARALLELISM 14;
     ```

   - For disk-based channels = *n* times 2 where *n* = <*the number of instances*>

     ```
     CONFIGURE DEFAULT DEVICE TYPE TO DISK;
     CONFIGURE DEVICE TYPE DISK PARALLELISM 16;
     ```

3. Issue the following RMAN command:

   ```
   RUN {
     restore database;
     recover database;
     }
   ```

4. To restore the Oracle Secure Backup catalog, see the Oracle Secure Backup documentation at

   http://download.oracle.com/docs/cd/E14812_01/doc/doc.103/e12834/catalog_recovery.htm#BABIJEIH

Restore into a New Oracle ASM Disk Group

For Oracle ASM disk groups, create a restore service that runs on a maximum of two database instances and allocate a total of eight RMAN channels per database instance for the restore.

Offload Backups with Oracle Data Guard

Oracle Data Guard and standby databases deliver a high return on investment when used to offload backups from the primary database. Both disk and tape-based backups can be performed using a physical standby database. Data Guard physical standby databases support all Oracle datatypes and features, including Exadata Hybrid Columnar Compression, and it can support the very high transaction volume driven by Exadata Database Machine. You can also use Oracle Active Data Guard to offload incremental backups to a standby database and greatly reduce backup times and the impact on the primary database.

Additional benefits of Oracle Data Guard include a high return on investment when used for queries, reports, testing, or rolling database upgrades, automatic block repair (with zero impact on applications), and other maintenance, while also providing disaster protection. Oracle Data Guard is the Oracle MAA prescribed disaster-recovery solution to protect mission critical databases residing on Exadata Database Machine and the Exadata Storage Server.

For more information, see the "Oracle Data Guard: Disaster Recovery for the Oracle Exadata Database Machine" white paper on the MAA Web site at http://www.oracle.com/technetwork/database/features/availability/maa-wp-dr-dbm-130065.pdf

Monitoring and Troubleshooting

Monitoring RMAN

When the RMAN job is executed the job transcript is written to `stdout` by default, but the output can be redirected to a log file that can be analyzed for errors and warnings as well as to review backup piece names written too. Additionally, RMAN uses the `NLS_DATE_FORMAT` environment variable to report times in hours / minutes and seconds, that can be useful to monitor run times.

Monitoring and Troubleshooting Oracle Secure Backup

Where you begin troubleshooting depends on the problem reported:

- Primary Oracle Secure Backup resources:
  - Backup/restore Job transcript and/or properties
  - Daemon (process) logs
  - Device logs
- External environmental areas to review:
  - Operating System configuration settings
  - Confirm Oracle Secure Backup user has the correct Operating System privileges to perform backup and restore operations
  - Confirm the tape device is accessible to the host

Job Transcripts

Job transcripts are usually the first place to begin providing detailed error messages for Oracle Secure Backup jobs. Job transcripts may be viewed using the Web Tool or command-line interface, `obtool`.

Obtool Commands:

```
ob>  lsjob -A           Listing of all jobs useful to obtain Job ID
ob>  catxcr -fl0 <jobid>     Displays entire transcript for the job
```

Job Properties

Job properties are another resource to determine what caused corresponding job issue. In some circumstances the job transcripts may not contain information such as when:

- Backup/restore job failed before it began
- Parent job scheduled when the backup/restore request issued but child job, which actually transferred data never, began

Job Properties and Logs

Job properties and logs may provide information describing why the job failed and are available from the Web Tool or the command line. To access properties from the command line, use the command:

```
ob> lsjob --log <jobid>
```

Monitoring TCP/IP Traffic

Oracle Secure Backup sends the data across the TCP/IP stack. To verify the backup rates, you can view the "`sar`" output either in real time or historically to see the transfer rates achieved from the database servers to the media servers or vice versa.

Conclusion

Oracle Exadata Database Machine and the best practices described in this white paper allow you to backup, restore, and recover Oracle Database extremely fast, using either disk-based backup or tape-based backup and restore mechanisms. Exadata Database Machine and Exadata Storage Server benefit from their native integration with Oracle Database and other Oracle features. The best practices provide an optimal solution for Oracle databases of any size, and the recommendations are straightforward to implement using standard RMAN commands.

The solutions in this white paper to backup Oracle databases residing on Exadata Database Machine and the Exadata Storage Server are prescribed and validated by Oracle MAA.

Appendix A – Test Environment

The MAA test environment used in this white paper was comprised of the following hardware:

- One Exadata Database Machine
- One Sun Fire X4170 server for the Oracle Secure Backup administrative server
- Two Sun Fire X4275 servers for Oracle Secure Backup media servers
- One Sun StorageTek SL500 with 14 LTO4 tape drives
- One Brocade 5100 8Gbit Fiber Channel switch

Backup and Recovery Performance and Best Practices for Exadata Cell and the Oracle Exadata Database Machine
February 2011
Author: Andrew Babb
Contributing Authors: Lawrence To, Steve Fried, Steve Wertheimer, Douglas Utzig, Michael Nowak, Tim Chien, Viv Schupmann (editor)

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
oracle.com

Oracle is committed to developing practices and products that help protect the environment

0109