

Oracle Exadata Storage Server Configuration and Administration

ORACLE®

Copyright © 2009, Oracle. All rights reserved.

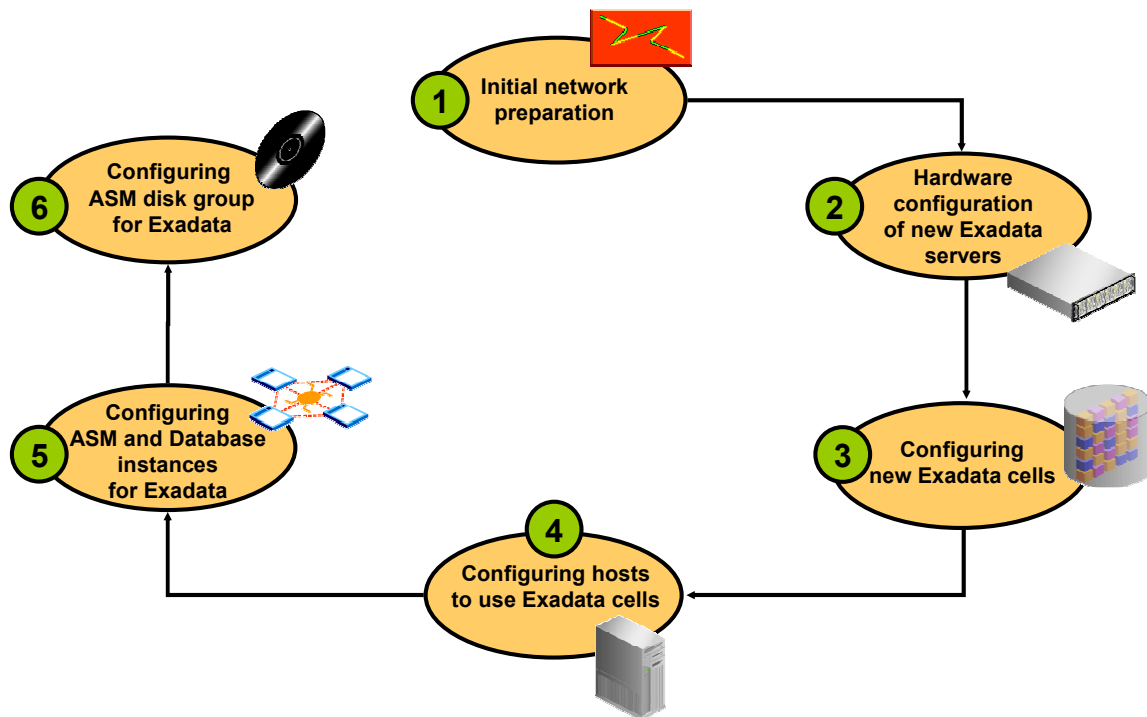
Objectives

After completing this lesson, you should be able to:

- Perform the initial Exadata Storage Server boot sequence
- Create and configure an Exadata cell
- Create and configure ASM disk groups using Exadata storage
- Use the Exadata administration tool CellCLI
- Handle probable maintenance scenarios

ORACLE

Exadata Installation and Configuration Overview



Exadata Installation and Configuration Overview

Oracle Exadata Storage Server ships with all hardware and software pre-installed, however it is necessary to configure Exadata for your environment. This slide introduces a general overview of the configuration tasks.



Initial Network Preparation

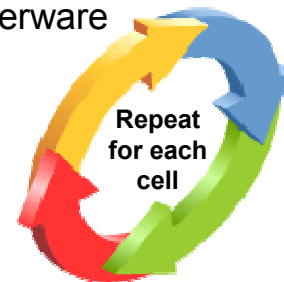
1

For each storage cell, assign the following IP addresses:

- One IP address for the bonded InfiniBand port.
- One IP address for administration network access.
- One IP address for lights out management.

Note these network configuration recommendations:

- Set up a fault-tolerant, private network subnet for the InfiniBand network.
 - Use the InfiniBand network for Oracle Clusterware
- Assign a block of IP addresses for each network type.
 - Do not allocate IP addresses ending in .0, .1, or .255
- Define your storage cells to DNS.



ORACLE

Initial Network Preparation

Each storage cell contains the following network ports:

1. One dual-port InfiniBand card for high-speed, high-volume data transfer: Each Exadata cell is designed to be connected to two separate InfiniBand switches for high availability. The dual port card is only for availability reasons because each port is capable of transferring the full data bandwidth generated by the storage cell. You will need to assign one IP address to the bonded InfiniBand interface during the initial configuration of the storage cell.
2. Gigabit Ethernet ports for general administration network access to the cell operating system: Sun Oracle Exadata Storage Server comes with four Gigabit Ethernet ports. However, only connect one port to a switch, and configure it for network access. You will need to assign one IP address to the cell for network access during the initial configuration of the storage cell.
3. One gigabit Ethernet port for lights out management: Sun Oracle Exadata Storage Server uses Sun Integrated Lights Out Manager (ILOM). You should assign one IP address to the cell for ILOM during the initial configuration of the storage cell. You can access the remote management functionality with a Java-enabled Web browser at the assigned IP address.

Initial Network Preparation (continued)

Note the following network configuration and IP address recommendations:

- It is recommended that the InfiniBand network should be a dedicated private network subnet for Exadata cells and database server hosts. Multiple InfiniBand switches are recommended to eliminate the switch as a single point of failure.
- The InfiniBand network should be used for network and storage communication when using Oracle Clusterware. Use the following command on your clusterware hosts to verify the private network for Oracle Clusterware communication is using InfiniBand:

```
oifcfg getif -type cluster_interconnect
```

- The Reliable Data Socket (RDS) protocol should be used over the InfiniBand network for database server to cell communication and Oracle Real Application Clusters (RAC) communication. Check the database alert log to verify the private network for Oracle RAC is running the RDS protocol over the InfiniBand network. The following message should be in the log:

```
cluster interconnect IPC version: Oracle RDS/IP (generic)
```

- Choose non-overlapping blocks of IP addresses. Select non-overlapping IP subnet blocks, one for each network interface port type (InfiniBand storage network, general administration network and lights out management network), so that they do not overlap with each other, and do not overlap other IP subnet blocks that are already in use by any of the machines on the network..

As mentioned before, the InfiniBand storage network should be a dedicated private subnet for Exadata cells and database server hosts.

Typically, the general administration network and the lights out management network will be on the same subnet and may share that subnet with other hosts. For example, on the 192.168.200 subnet, you might assign the block of IP addresses between 192.168.200.31 and 192.168.200 the following address blocks could be defined:

- 192.168.200.32/27
- 192.168.200.64/27

The first address block could be assigned the general administration network and the second address block could be assigned to the lights out management network. Other hosts in the 192.168.200 subnet would be allocated IP addresses outside both of these blocks.

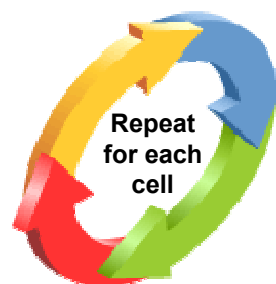
- Do not allocate addresses that end in .0, .1, or .255, or those that would be used as broadcast addresses for the specific netmask that you have selected. For example, avoid addresses such as 192.168.200.0, 192.168.200.1, and 192.168.200.255.
- If a domain name system (DNS) is required, then set up your DNS to help reference cells and interconnections. Exadata Cells do not require DNS. However, if DNS is configured in your network, then set up your DNS with the appropriate IP address and host name of Exadata Cell.



Hardware Configuration of New Exadata Cell

2

1. Check all physical connections.
2. Power on the Exadata cell.
3. Answer questions during boot sequence:
 - Domain Name Service (DNS) server IP addresses
 - Time preference (time region and location)
 - Network Time Protocol (NTP) servers
 - Ethernet and InfiniBand IP addresses, netmasks, gateway and hostnames
 - Remote management configuration details
4. Change the initial passwords for the root, celladmin, and cellmonitor users.



Hardware Configuration of New Exadata Cell

The slide lists the general steps to configure the Exadata storage cell:

1. Check all the physical connections to the Exadata storage cell. It is important that all the physical network connections are correct prior to configuring the cell. Check also that both power supplies are connected and that you have a keyboard, video display and mouse.
2. Power on the cell to boot its operating system.
3. Answer the configuration questions when you are prompted by the cell. The slide lists the information that you need to provide.
4. After you successfully perform the previous step, the login screen is displayed. Change the initial passwords for the root, celladmin, and cellmonitor users to more secure passwords. The initial password for root is welcome1. The initial password for the cellmonitor and celladmin users is welcome.

Note: The cellmonitor user is set up with privileges that enable you to view Exadata Cell objects only. You must be logged in as the celladmin user to perform administrative tasks.

Note: Configuration of the Sun Oracle Exadata Storage Server differs slightly from the configuration of an Exadata server based on HP hardware. In a HP-based Exadata server the remote management configuration was conducted through the BIOS setup utility rather than during the first boot configuration process.



Answering Questions During Boot Sequence

```
...
Network interfaces
Name State      IP address      Netmask      Gateway      Hostname
eth0  Linked
eth1  Unlinked
eth2  Unlinked
eth3  Unlinked
ib0   Linked
ib1   Linked
Warning. Some network interface(s) are disconnected. Check cables and switches and retry
Do you want to retry (y/n) [y]: n
Nameserver: mynameserv.company.com
Add more nameservers (y/n) [n]: n
Setting up local time...
Select country by number, [n]ext, [1]ast: 230
Select zone by number, [n]ext: 17
Selected timezone: America/Denver
Is this correct (y/n) [y]: y
The current ntp server(s):
Do you want to change it (y/n) [n]: y
Fully qualified hostname or ip address for NTP server. Press enter if none: ntpl.company.com
Continue adding more ntp servers (y/n) [n]: n
...
```

ORACLE

3 - 7

Copyright © 2009, Oracle. All rights reserved.

Answering Questions During Boot Sequence

The next four slides show an example of the initial configuration process for the Sun Oracle Exadata Storage Server. On each slide, the blue text indicates a user input.

The configuration commences during the server boot sequence. The output from the initial part of the boot sequence is not shown. This slide commences at the beginning of the interview phase where user input is required.

In this slide settings are made for the DNS name server, timezone and NTP server.

Notice that the interview phase commences with a warning indicating that a number of network interfaces are disconnected. As shown above, it is safe to ignore this warning because each Sun Oracle Exadata server comes equipped with four ethernet ports however only one (eth0) is used. So it is normal for eth1, eth2 and eth3 to be disconnected. Always make sure though that the required network interfaces (eth0, ib0 and ib1) are correctly linked.



Answering Questions During Boot Sequence

```
...
Network interfaces
Name State      IP address      Netmask      Gateway      Hostname
eth0  Linked
bond0 ib0,ib1
Select interface name to configure or press Enter to continue: eth0
Selected interface. eth0
IP address or none: 10.XXX.XXX.XXX
Netmask: 255.255.248.0
Gateway (IP address or none): 10.XXX.XXX.1
Fully qualified hostname or none: cell101.company.com
Continue configuring or re-configuring interfaces? (y/n) [y]: y
Network interfaces
Name State      IP address      Netmask      Gateway      Hostname
eth0  Linked      10.XXX.XXX.XXX  255.255.248.0  10.XXX.XXX.1  cell101.company.com
bond0 ib0,ib1
Select interface name to configure or press Enter to continue: bond0
Selected interface. bond0
IP address: 192.168.50.76
Netmask: 255.255.255.0
Fully qualified hostname or none: cell101-priv.company.com
Continue configuring or re-configuring interfaces? (y/n) [y]: y
Network interfaces
Name State      IP address      Netmask      Gateway      Hostname
eth0  Linked      10.XXX.XXX.XXX  255.255.248.0  10.XXX.XXX.1  cell101.company.com
bond0 ib0,ib1  192.168.50.76  255.255.255.0  cell101-priv.company.com
Select interface name to configure or press Enter to continue: <enter>
...
```

ORACLE

3 - 8

Copyright © 2009, Oracle. All rights reserved.

Answering Questions During Boot Sequence (continued)

In this slide the configuration phase continues with settings being specified for the Ethernet network (eth0) that supports administrative access to the storage server, along with the InfiniBand network (bond0) that supports the main storage network.

Notice that the InfiniBand interface is named bond0 and uses bonding between physical InfiniBand interfaces ib0 and ib1. Bonding provides the ability to transparently fail over from ib0 to ib1 or from ib1 to ib0 if connectivity to either interface is lost.

If you choose not to configure each interface in the list, that interface will not be configured, and it will not be started at system startup. In this case the cell will be inoperable. You can later configure, or change, cell network settings using the `ipconf` utility.



Answering Questions During Boot Sequence

```
...
Select cononical hostname from the list below
1: cell01.company.com
2: cell01-priv.company.com
Canonical fully qualified domain name: 1
Select default gateway interface from the list below
1: eth0
Default gateway interface: 1

Canonical hostname: cell01.company.com
Nameservers: mynameserv.company.com
Timezone: America/Denver
NTP servers: ntp1.company.com
Default gateway device: eth0
Network interfaces
Name State      IP address      Netmask          Gateway           Hostname
eth0  Linked      10.XXX.XXX.XXX  255.255.248.0    10.XXX.XXX.1      cell01.company.com
bond0 ib0,ib1    192.168.50.76   255.255.255.0
Is this correct (y/n) [y]: y
...
```

ORACLE

Answering Questions During Boot Sequence (continued)

In this slide the network configuration is finalized with the specification of the canonical hostname and default gateway. Both of these settings map to the ethernet network providing administrative access to the cell.



Answering Questions During Boot Sequence

```
...
Do you want to configure basic ILOM settings (y/n) [y]: y
Loading basic configuration settings from ILOM ...
ILOM Fully qualified hostname [cell01-ilom.company.com]: cell01-ilom.company.com
ILOM IP address [10.XXX.XXX.YYY]: 10.XXX.XXX.YYY
ILOM Netmask [255.255.248.0]: 255.255.248.0
ILOM Gateway [10.XXX.XXX.1]: 10.XXX.XXX.1
ILOM Nameserver or none [mynameserv.company.com]: mynameserv.company.com
ILOM Use NTP Servers (enabled/disabled) [enabled]: enabled
ILOM First NTP server. Fully qualified hostname or ip address or none [ntp1.company.com]: ntp1.company.com
ILOM Second NTP server. Fully qualified hostname or ip address or none [none]: none

Basic ILOM configuration settings:
Hostname           : cell01-ilom.company.com
IP Address         : 10.XXX.XXX.YYY
Netmask            : 255.255.248.0
Gateway            : 10.XXX.XXX.1
DNS servers        : mynameserv.company.com
Use NTP servers     : enabled
First NTP server    : ntp1.company.com
Second NTP server   : none
Timezone (read-only) : America/Denver

Is the correct (y/n) [y]: y
...
```

ORACLE

3 - 10

Copyright © 2009, Oracle. All rights reserved.

Answering Questions During Boot Sequence (continued)

Configuration completes with settings for Integrated Lights Out Manager (ILOM).

If you choose not to configure ILOM at this time you can use the `ipconf` utility to do so later.

After the user interview phase is completed, the Exadata storage server finalizes its system startup process. The output from the remaining system startup activities is not shown in the slide. Finally, a login prompt is displayed.

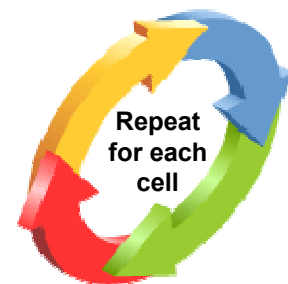
As mentioned before, after you successfully configure the cell you should login and change the initial passwords for the `root`, `celladmin`, and `cellmonitor` users to more secure passwords. The initial password for `root` is `welcome1`. The initial password for the `cellmonitor` and `celladmin` users is `welcome`.



Configuring a New Exadata Cell

3

1. Run performance tests on the cell with `CALIBRATE`.
2. Configure the cell.
3. Confirm cell disks.
4. Create grid disks.



ORACLE

3 - 11

Copyright © 2009, Oracle. All rights reserved.

Configuring a New Exadata Cell

As part of the initial boot configuration, the Cell Server software is started with a basic configuration. In addition, the physical disks and flash modules that Exadata uses for data storage are configured as cell disks. By default, all the flash-based cell disks are allocated to Exadata Smart Flash Cache, and all the cell disks based on physical disks are available to be allocated into grid disks according to your needs.

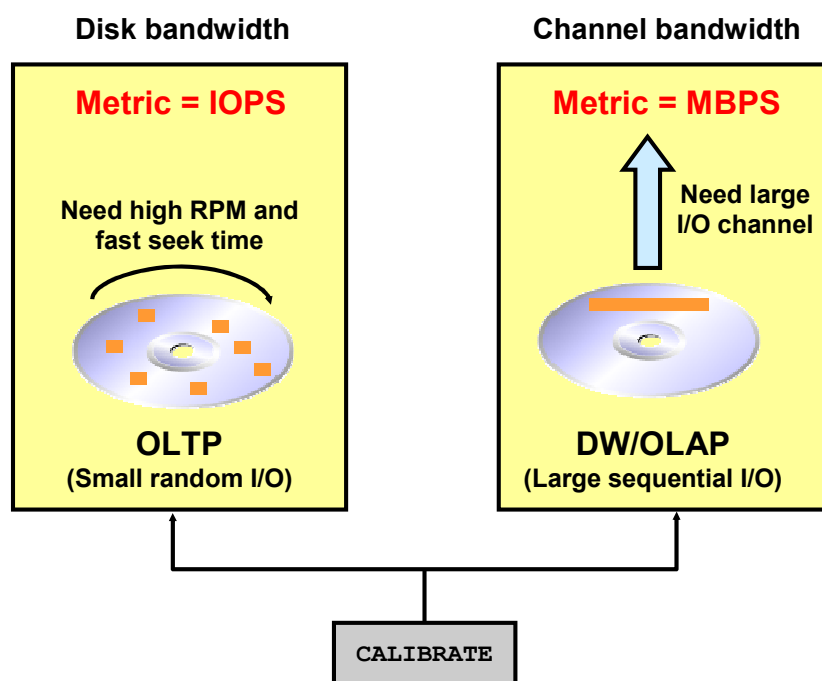
At this point you are ready to finalize the configuration of the Exadata cell software. Following is a summary of the recommended procedure, all the steps are executed using CellCLI:

1. As the `root` user, run performance tests on the cell with the `CALIBRATE` command.
2. As the `celladmin` user, configure the cell server software with the `ALTER CELL` command.
3. As `celladmin`, confirm the cell disks by using the `LIST CELLDISK` command.
4. As `celladmin`, create the grid disks on each cell disk of the storage cell by using the `CREATE GRIDDISK` command.

Repeat this process on each Exadata Storage Server.



Important I/O Metrics for Oracle Databases



Important I/O Metrics for Oracle Databases

The **CALIBRATE** command runs raw performance tests on Exadata disks and flash modules. This enables you to measure two important database metrics: IOPS and MBPS.

The database I/O workload typically consists of small random I/O and large sequential I/O. Small random I/O is more prevalent in an OLTP application environment in which each server process reads a data block into the buffer cache for updates and the changed blocks are written in batches by the DBWn process. Large sequential I/O is common in an OLAP application environment. The OLTP application performance depends on how fast a small I/O is serviced, which depends on how fast the disk can spin and find the data. Large I/O performance depends on the capacity of the I/O channel that connects the server to the storage array; large I/O throughput is better when the capacity of the channel is larger.

- **IOPS (I/O per second):** This metric represents the number of small random I/O that can be serviced in a second. The IOPS rate mainly depends on how fast the disk media can spin. The IOPS rate from a storage array can be increased either by adding more disk drives or by using disk drives with a higher RPM (rotations per minute) rate.
- **MBPS (megabytes per second):** The rate at which data can be transferred between the computing server node and the storage array depends on the capacity of the I/O channel that is used to transfer data. More data can be transferred through a wider pipe.



Testing Performance Using CALIBRATE

```
[root@cell101 ~]# cellcli
CellCLI: Release 11.2.1.2.0 - Production on Mon Nov 02 16:42:06 PST 2009

Copyright (c) 2007, 2009, Oracle. All rights reserved.
Cell Efficiency ratio: 1.0

CellCLI> calibrate force
Calibration will take a few minutes...
Aggregate random read throughput across all hard disk luns: 1601 MBPS
Aggregate random read throughput across all flash disk luns: 4194.49 MBPS
Aggregate random read I/Os per second (IOPS) across all hard disk luns: 4838
Aggregate random read I/Os per second (IOPS) across all flash disk luns: 137588
Controller read throughput: 1615.85 MBPS
Calibrating hard disks (read only) ...
Lun 0_0 on drive [20:0      ] random read throughput: 152.81 MBPS, and 417 IOPS
Lun 0_1 on drive [20:1      ] random read throughput: 154.72 MBPS, and 406 IOPS
...
Lun 0_10 on drive [20:10     ] random read throughput: 156.84 MBPS, and 421 IOPS
Lun 0_11 on drive [20:11     ] random read throughput: 151.58 MBPS, and 424 IOPS
Calibrating flash disks (read only, note that writes will be significantly slower).
Lun 1_0 on drive [[10:0:0:0]] random read throughput: 269.06 MBPS, and 19680 IOPS
Lun 1_1 on drive [[10:0:1:0]] random read throughput: 269.18 MBPS, and 19667 IOPS
...
Lun 5_2 on drive [[11:0:2:0]] random read throughput: 269.15 MBPS, and 19603 IOPS
Lun 5_3 on drive [[11:0:3:0]] random read throughput: 268.91 MBPS, and 19637 IOPS
CALIBRATE results are within an acceptable range.
```

ORACLE

Testing Performance Using CALIBRATE

The CALIBRATE command enables you to verify the disk and flash memory performance before the cell is put online. You must execute this command while being logged as the root user at the operating system level.

The CALIBRATE FORCE command allows you to run the tests when Cell Server is running. If you do not use the FORCE option, then Cell Server must be shut down. Running CALIBRATE at the same time as the Cell Server will impact performance which is why it is not recommended during normal operation.

Because the Cell Server software is running immediately after the initial boot sequence, you must either shut down the Cell Server software or execute the CALIBRATE FORCE command.

CALIBRATE FORCE is acceptable in this circumstance because the cell is not yet running a user workload, so there is no work to disrupt. In the above example, which shows a typical output for SAS disks, the results matched expectations. A message will alert you if the performance measurements are substandard.



Configuring the Exadata Cell Server Software

```
[celladmin@cell01 ~]$ cellcli
CellCLI: Release 11.2.1.2.0 - Production on Mon Nov 02 17:46:13 PST 2009

Copyright (c) 2007, 2009, Oracle. All rights reserved.
Cell Efficiency ratio: 1.0

CellCLI> ALTER CELL smtpServer='my_mail.example.com', -
          smtpFromAddr='exadata.cell01@example.com', -
          smtpPwd=<email_address_password> -
          smtpToAddr='jane.smith@example.com', -
          notificationPolicy='critical,warning,clear', -
          notificationMethod='mail'

Cell cell01 successfully created
Starting CELLSRV services...
The STARTUP of CELLSRV services was successful
CellCLI>
```

ORACLE

3 - 14

Copyright © 2009, Oracle. All rights reserved.

Configuring the Exadata Cell Server Software

The settings provided during the initial boot sequence configure the hardware and cell operating system. In addition, the Cell Server software is configured using the `CREATE CELL` command. By default, the cell name is set to the network host name of the Exadata server and the `INTERCONNECT1` attribute is set to `bond0` which is the InfiniBand storage network interface.

You can change the name of the cell or configure the optional Cell Server attributes by using the `ALTER CELL` command.

The slide shows an example `ALTER CELL` command that configures email notification. This facility sends email messages to the administrator of the storage cell whenever critical, warning, and clear alerts are detected by the cell. In addition to email notification it is possible to configure notification using SNMP.

Note: After the initial boot configuration, Restart Server (RS) and Management Server (MS) should be running. If not, then an error message will display when using the CellCLI utility. In that case, run the following commands to start the RS and MS services:

```
ALTER CELL STARTUP SERVICES RS
ALTER CELL STARTUP SERVICES MS
```



Confirming Cell Disks

```
CellCLI> LIST CELLDISK
      CD_00_cell101      normal
      CD_01_cell101      normal
      CD_02_cell101      normal
      ...
      CD_09_cell101      normal
      CD_10_cell101      normal
      CD_11_cell101      normal
      FD_00_cell101      normal
      FD_01_cell101      normal
      FD_02_cell101      normal
      ...
      FD_13_cell101      normal
      FD_14_cell101      normal
      FD_15_cell101      normal

CellCLI>
```

ORACLE

3 - 15

Copyright © 2009, Oracle. All rights reserved.

Confirming Cell Disks

After the Exadata cell is configured, there are 12 disk-based cell disks and 16 flash-based cell disks created with default names. In most cases, you can use the default cell disk names.

The above example shows the use of the `LIST CELLDISK` command to display the disk-based and flash-based cell disks. Check that the command shows a status of `normal` for all the cell disks and take note of the cell disk names for future reference.

If desired, you should now configure your cell disks to enable the creation of interleaved grid disks. Use the following command to alter all the cell disks to enable interleaving:

```
ALTER CELLDISK ALL HARDDISK INTERLEAVING='normal_redundancy'
```



Creating Grid Disks

```
CellCLI> CREATE GRIDDISK ALL PREFIX=data, size=300G
```

```
GridDisk data_CD_00_cell101 successfully created
```

```
...
```

```
GridDisk data_CD_11_cell101 successfully created
```

```
CellCLI> CREATE GRIDDISK ALL PREFIX=fra
```

```
GridDisk fra_CD_00_cell101 successfully created
```

```
...
```

```
GridDisk fra_CD_11_cell101 successfully created
```

```
CellCLI> LIST GRIDDISK
```

```
data_CD_00_cell101      active
```

```
...
```

```
data_CD_11_cell101      active
```

```
fra_CD_00_cell101       active
```

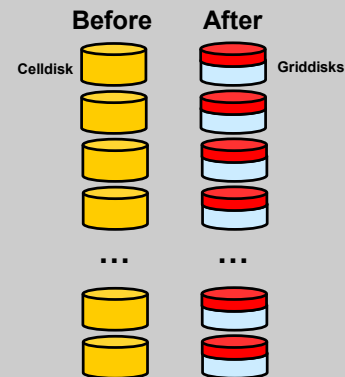
```
...
```

```
fra_CD_11_cell101       active
```

```
CellCLI> exit
```

```
[celladmin@cell101 ~]$
```

Use fastest
disk portion



ORACLE

3 - 16

Copyright © 2009, Oracle. All rights reserved.

Creating Grid Disks

After cell disks are created, you can create grid disks using the `CREATE GRIDDISK` command. In the above example, the `ALL PREFIX` option is used to automatically create one grid disk on each cell disk. When the `ALL PREFIX` option is used, the generated grid disk names are composed of the grid disk prefix followed by an underscore (`_`) and then the cell disk name.

It is best practice to use the ASM disk group name as the prefix name for the corresponding grid disks. In the example, prefix values `DATA` and `FRA` are the names of the ASM disk groups that will be created. Grid disk names must be unique across all cells within a single deployment. By following the recommended naming conventions for naming the grid and cell disks you will automatically get unique names.

The optional `SIZE` attribute specifies the size of your grid disk. If omitted, the grid disk will automatically consume all the space remaining on the corresponding cell disk.

The `LIST GRIDDISK` command shows the grid disks that are created.

Note that for cell disks which are not enabled for interleaving, the first grid disk created on each cell disk uses the outermost portion of the disk. In this area each track contains more data than the inner tracks resulting in higher transfer rates and better performance. Because the best available offset is chosen automatically in chronological order of grid disk creation, you should first create those grid disks expected to contain the most frequently accessed data, and then create the grid disks that will contain the relatively colder data.



Configuring Hosts to Access Exadata Cells

4

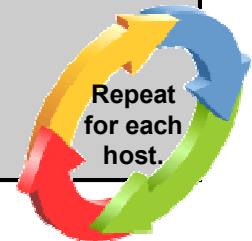
1. Create the following directory and files:

```
# mkdir -p /etc/oracle/cell/network-config
# chown oracle:dba /etc/oracle/cell/network-config
# chmod ug+wx /etc/oracle/cell/network-config

$ cd /etc/oracle/cell/network-config

$ cat - > /etc/oracle/cell/network-config/cellinit.ora
ipaddress1=192.168.50.23/24

$ cat - > /etc/oracle/cell/network-config/cellip.ora
cell="192.168.51.27"
cell="192.168.51.28"
cell="192.168.51.29"
```



2. Restart database and ASM instances.

ORACLE

3 - 17

Copyright © 2009, Oracle. All rights reserved.

Configuring Hosts to Access Exadata Cells

After your Exadata cells are configured, the database server hosts must be configured with the `cellinit.ora` and the `cellip.ora` files to use the cells:

- The `cellinit.ora` file contains all the database server IP addresses that connect to the storage network. The IP addresses are specified in Classless Inter-Domain Routing (CIDR) format. For example, `192.168.50.23/24` corresponds to all the IP addresses between `192.168.50.1` and `192.168.50.254`.
- The `cellip.ora` file contains the IP addresses that are used by storage cells to send data to the database server host. These IP addresses correspond to the bonded InfiniBand interface (`bond0`) on the cells.

Restart the database and the Oracle ASM instances on the database server host after you finish creating the `cellinit.ora` and `cellip.ora` files. After the files have been configured they should not be edited while your database or ASM instances are running.

Note: Only Linux x86-64 database servers are supported to access Exadata storage.



Configuring ASM and Database Instances for Exadata

5

- Both Database and ASM installation software must be at least 11.2.0.1.
- Use ASM to store OCR and voting disks on Exadata.
- ASM instance initialization parameter:
 - `ASM_DISKSTRING='o/*/*`
- Database instances initialization parameters:
 - `COMPATIBLE='11.2.0.0.0'`



ORACLE

3 - 18

Copyright © 2009, Oracle. All rights reserved.

Configuring ASM and Database Instances for Exadata

Oracle Database and Oracle Grid Infrastructure 11g Release 2 (11.2.0.1) or greater must be installed on the database server host before you can access Exadata version 2 cells from ASM and database instances.

If you are using Oracle Clusterware, it is recommended that you place the Oracle Cluster Registry (OCR) and voting disks on ASM. Alternatively, you can use existing conventional storage, or you can configure and use a set of iSCSI devices on your storage cells for this purpose.

To ensure that ASM discovers Exadata grid disks, set the `ASM_DISKSTRING` initialization parameter. A search string of the form `o/<cell IP address>/<grid disk name>` is used to discover Exadata grid disks. Wildcards may be used to expand the search string. For example, to discover all the available Exadata grid disks use the following search string:

```
ASM_DISKSTRING='o/*/*
```

To discover a subset of available grid disks having names that begin with `data` use:

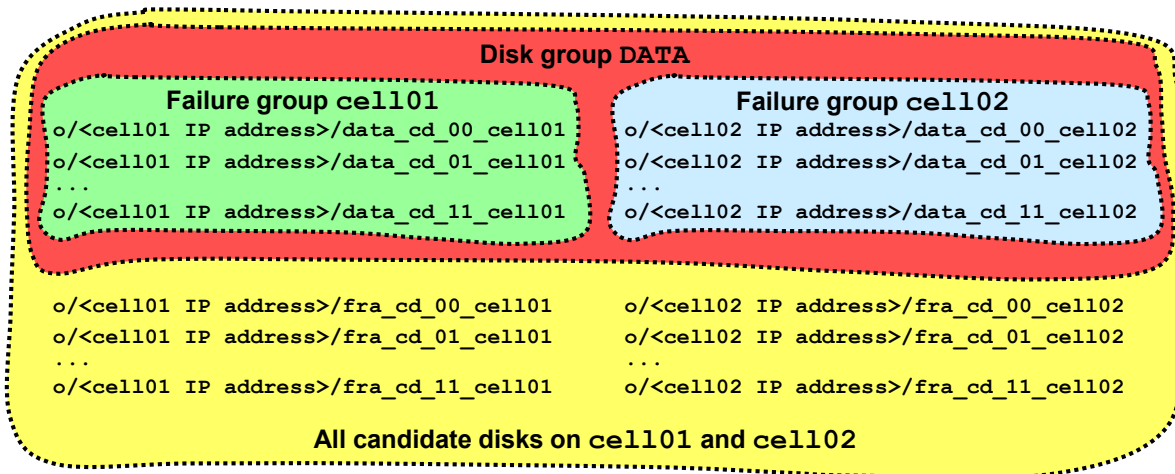
```
ASM_DISKSTRING='o/*/data*'
```

To configure a database instance to access cell storage, ensure that the `COMPATIBLE` parameter is set to `11.2.0.0.0` or later in the database initialization file. Note that Database Configuration Assistant (DBCA) 11.2.0.1 does not set the `COMPATIBLE` initialization parameter to `11.2.0.0.0` by default and you must set this parameter on the Initialization Parameters page.



Configuring ASM Disk Groups for Exadata

6



```
CREATE DISKGROUP data NORMAL REDUNDANCY
DISK 'o/*/data*'
ATTRIBUTE 'compatible.rdbms' = '11.2.0.0.0',
          'compatible.asm' = '11.2.0.0.0',
          'cell.smart_scan_capable' = 'TRUE',
          'au_size' = '4M';
```

ORACLE

3 - 19

Copyright © 2009, Oracle. All rights reserved.

Configuring ASM Disk Groups for Exadata

You can now create ASM disk groups from your ASM instance. An ASM disk group can include Exadata grid disks and conventional disks. However, to enable smart scan processing, all the disks in an ASM disk group must be Exadata grid disks, and the following disk group attribute settings must be used:

```
'compatible.rdbms' = '11.2.0.0.0'
'compatible.asm' = '11.2.0.0.0'
'cell.smart_scan_capable' = 'TRUE'
```

In addition, it is recommended that you set the AU_SIZE disk group attribute value to 4M to optimize disk scanning.

The example on the slide shows candidate ASM disks from two Exadata cells, cell101 and cell102. The CREATE DISKGROUP statement specifies all of the candidate ASM disks having names that start with data. By default, ASM failure groups corresponding to each cell are automatically defined. As a result, two failure groups are automatically created using corresponding grid disks from each cell. By default, the failure group names correspond to the cell names.

Optional Configuration Tasks

- Configure Exadata storage security
- Configure I/O Resource Management (IORM)

ORACLE

3 - 20

Copyright © 2009, Oracle. All rights reserved.

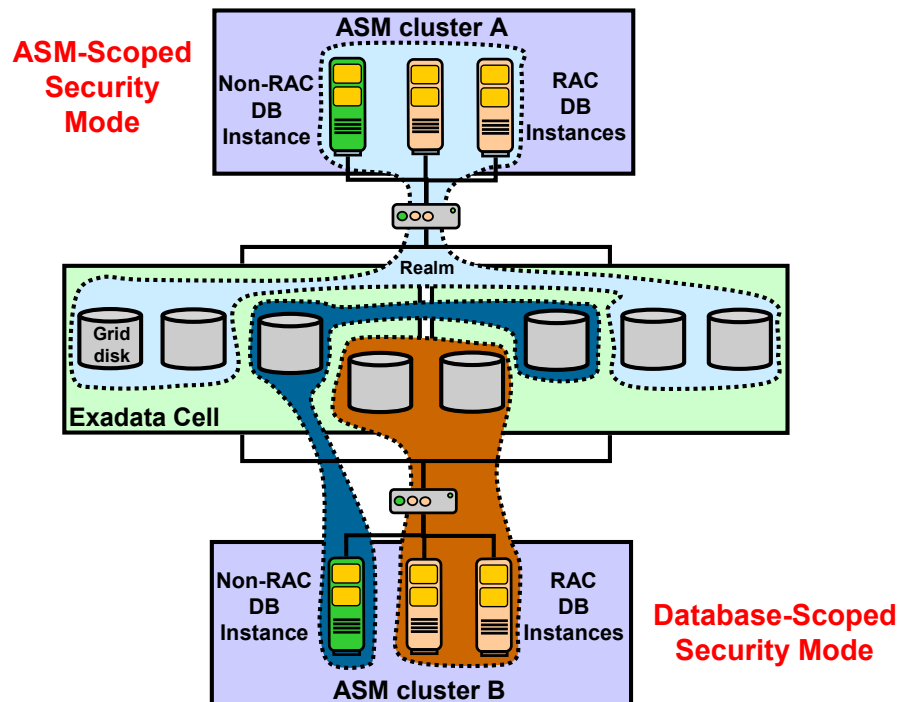
Optional Configuration Tasks

After you complete the cell configuration, you can perform the following optional tasks on the storage cell:

- Configure Exadata storage security.
- Configure I/O Resource Management (IORM). IORM is the subject of lesson five in this course.

Note: Repeat each configuration task on each relevant storage cell.

Exadata Storage Security Overview



Exadata Storage Security Overview

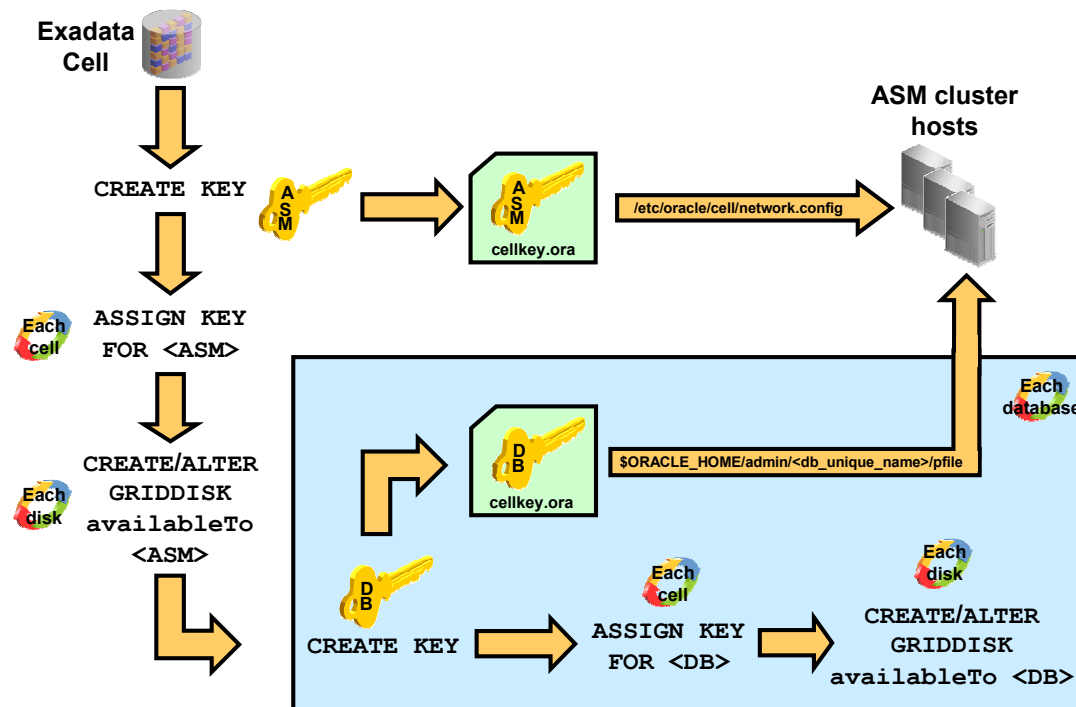
Exadata storage security is implemented by controlling which ASM clusters and database clients can access specific grid disks on storage cells. A collection of grid disks in the same security domain is referred to as a storage realm:

- To set up security so that all database clients of an ASM cluster have access to specified grid disks, configure ASM-scope security.
- To set up security so that specific database clients of an ASM cluster have access to specified grid disks, configure database-scope security.

Both concepts are illustrated in the slide. ASM cluster A shares two grid disks per cell with all of its database clients. ASM cluster B shares one grid disk per cell to store the single instance database, and another two grid disks to store the RAC database.

Note: By default, none of these security modes are implemented. This situation is called open security where all database clients can access all grid disks. Open security is useful for test or development databases with lower security requirements. Open security does not require any configuration and as long as the database hosts are well secured, you could also use this mode for your production databases.

Exadata Storage Security Implementation



Exadata Storage Security Implementation

The slide briefly describes the steps to configure ASM-scoped and database-scoped security. It is important to realize that you must set up ASM-scoped security first if you want to set up database-scoped security.

To implement ASM-scoped security perform the following:

1. Shut down your ASM and database instances.
2. Generate a security key using the `CREATE KEY CellCLI` command. Run this command once only on any cell.
3. Construct a `cellkey.ora` file using the generated security key. Copy the `cellkey.ora` file into the `/etc/oracle/cell/network-config/` directory on every host in your ASM cluster.
4. Use the `ASSIGN KEY` command to assign the security key to the Oracle ASM cluster on all the cells that you want the Oracle ASM cluster to access. The ASM cluster name is determined by the `DB_UNIQUE_NAME` initialization parameter setting.
5. Enter the Oracle ASM cluster name in the `availableTo` attribute with the `CREATE GRIDDISK` or `ALTER GRIDDISK` command to configure security on the grid disks on all the cells that you want the Oracle ASM cluster to access. At the conclusion of this step, each grid disk has an association with the ASM cluster that is allowed to use the disk.
6. Restart your ASM and database instances.

Exadata Storage Security Implementation (continued)

After you have configured and tested ASM-scoped security you can proceed to set up database-scoped. Perform the following procedure for each database you want to configure with database-scoped security:

1. Shut down your ASM and database instances.
2. Generate a security key using the `CREATE KEY` CellCLI command. Run this command once only on any cell.
3. Construct a `cellkey.ora` file using the generated security key.
Copy the `cellkey.ora` file into the
`$ORACLE_HOME/admin/<db_unique_name>/pfile/` directory on every host running your database.
4. Use the `ASSIGN KEY` command to assign the security key to the database on all the cells that you want the database to access. The database name is determined by the `DB_UNIQUE_NAME` initialization parameter setting.
5. Enter the database name in the `availableTo` attribute with the `CREATE GRIDDISK` or `ALTER GRIDDISK` command to configure security on the grid disks on all the cells that you want the database to access. At the conclusion of this step, each grid disk has an association with the ASM cluster and specific database that is allowed to use the disk.
6. Restart your ASM and database instances.

Note: For more information, including examples and further details, refer to the *Oracle Exadata Storage Server Software User's Guide 11g Release 2 (11.2)*.



Cell Maintenance Overview

- **Planned maintenance**
 - **Examples**
 - Patch, or upgrade Exadata server software
 - **Procedure Overview**
 1. Offline the corresponding ASM failure group.
 2. Execute your planned maintenance operation.
 3. Online the ASM failure group again.
- **Unplanned maintenance**
 - **Examples**
 - Disk failure, Cell hardware failure, or CELLSRV process failure
 - **Procedure Overview**
 1. Remedy the failure.
 2. Bring online or re-create the affected ASM failure groups.

ORACLE

3 - 24

Copyright © 2009, Oracle. All rights reserved.

Cell Maintenance Overview

After your Exadata storage servers are configured and in use the administrative focus shifts to ongoing monitoring and maintenance. Maintenance operations can be broadly divided into two categories: planned and unplanned.

Planned maintenance will most likely involve patching or upgrading the Exadata storage server software environment but it may also include replacing or upgrading an item of hardware prior to a failure. For example, you might replace one of the storage server power supplies as a planned maintenance operation because it is no longer functioning but its failure has not impacted on the operations of the cell.

If you are following the recommended pattern of maintaining ASM redundancy using failure groups on multiple Exadata cells, then planned maintenance operations can be undertaken with minimal disruption to normal processing. In essence, you need to:

1. Take offline the failure groups associated with the cell being maintained. You should define a maintenance window which provides ample time for the maintenance operation. This can be achieved by specifying a timeout clause in your `ALTER DISKGROUP ... OFFLINE` statement or by setting the `DISK_REPAIR_TIME` ASM initialization parameter.
2. Perform the required maintenance operation within the planned maintenance window.
3. Bring the affected ASM failure groups back online.

Cell Maintenance Overview (continued)

If the maintenance operation takes longer than planned you can either extend the maintenance window or if it has expired you will have to re-create the ASM disks and failure group.

Some failure scenarios can be automatically remedied by the Exadata server. For example, if the `cellsrv` process is killed, the restart server process (RS) will restart `cellsrv` and in most cases processing will continue uninterrupted.

In situations where failure cannot be automatically remedied, an unplanned maintenance operation is required. Unplanned maintenance may be required as a result of a disk failure, some other form of cell hardware failure or unrecoverable cell software failure.

If you are following the recommended pattern of maintaining ASM redundancy using failure groups on multiple Exadata cells, then the failure event will automatically cause your affected ASM disks to be taken offline. If possible, remedy the failure. After the failure is remedied then the affected ASM disks can either be brought back online or will need to be re-created depending on whether or not the failure is remedied before the amount of time specified in the `DISK_REPAIR_TIME` ASM initialization parameter. If the failure cannot be repaired then the failure groups associated with the failed cell will be dropped and should be re-created on another cell.

Note: The general procedures described in this overview rely on ASM redundancy (mirroring) across multiple Exadata cells to maintain at least one copy of data online during a planned or unplanned maintenance operation. If you do not implement ASM redundancy across multiple Exadata cells or you suffer simultaneous failures that affect all of your data copies then you will need to rebuild your database (in whole or in part) or perform a recovery operation.



Automated Cell Maintenance Operations

- Automatic addition of a replacement disk to the original disk group:
 - Cell disk and grid disks are automatically re-created
 - Each grid disk is added back to the original disk group
- Automatic cell restart:
 - Grid disks are brought online when a cell restarts
- Automatic firmware upgrades:
 - A golden firmware copy is kept on the cell and flashed to replacement components:
 - Immediately for a disk replacement
 - During reboot for other components such as the motherboard, InfiniBand HCA, disk controller, flash card, etc.

ORACLE

3 - 26

Copyright © 2009, Oracle. All rights reserved.

Automated Cell Maintenance Operations

Exadata V2 simplifies various cell maintenance operations by automating tasks that previously required administrator intervention:

- Automatic addition of a replacement disk to the original disk group: When a replacement disk is inserted after a physical disk failure, the cell disk and grid disks are automatically re-created, and each grid disk is automatically added back to the original disk group. The same occurs after the replacement of a flash card containing flash-based grid disks.
- Automatic cell restart: Grid disks are automatically changed to online when a cell recovers from a failure, or after a restart.
- Automatic firmware upgrades: A copy of the required firmware, called the golden version, is kept on the cell as part of the cell software distribution.
 - If a disk is replaced, the new disk is automatically flashed with the golden version of the firmware before being rebuilt as previously mentioned.
 - For other components, the cell must be shut down to replace the component. Once the Exadata cell is powered on, it will apply the golden version of the firmware on the new component and restart. For a replacement motherboard, the storage server will shutdown and the administrator will need to power on once again.
 - If a firmware change is made while the cell is running, a periodic check will raise an alert if a component does not match the golden firmware level. Once the cell is rebooted, it will update itself by reapplying the golden firmware version.



Replacing a Damaged Physical Disk

1 Determine the damaged disk.



```
CellCLI> LIST ALERTHISTORY -  
WHERE ALERTMESSAGE LIKE "Logical drive lost.*" DETAIL  
Logical drive lost. Lun:0_5. Status: normal. Physdisk: 20:5.  
Celldisk on it: CD_05_cell01. Griddisks on it: data_CD_05_cell01.  
The suggested action is: Refer to section Maintaining Physical Disks in  
the User Guide.
```

2 Replace physical disk.



```
LIST PHYSICALDISK → normal
```

3 Monitor ASM to confirm the re-addition of the disk.



```
SQL> SELECT NAME, STATE FROM V$ASM_DISK  
SQL> SELECT * FROM GV$OPERATION
```

ORACLE

Replacing a Damaged Physical Disk

Replacing a physical disk due to problem or failure is probably the most likely hardware maintenance operation that an Exadata cell might ever require. Assuming you are using ASM redundancy, the procedure to replace a problem disk is quite simple.

The first step requires that you identify the problem disk. This could occur in a number of ways:

- Hardware monitoring using ILOM may report a problem disk.
- If a disk fails, an Exadata alert is generated. The alert includes specific instructions for replacing the disk. If you have configured the system for alert notifications, then the alert will be sent to the designated email address or SNMP target. You can also use the `LIST ALERTHISTORY` command shown on the slide to identify the failed disk.
- The `LIST PHYSICALDISK` command may identify a disk reporting a status of warning or critical. Even if the cell is still functioning the problem may be a precursor to a disk failure.
- The `CALIBRATE` command may identify a disk delivering abnormally low throughput or IOPS. Even if the cell is still functioning, a single bad physical disk can degrade the performance of other good disks so you may decide to replace the identified disk. Note that running `CALIBRATE` at the same time as the cell is active will impact performance.

Once you have identified the problem disk you can replace it. When you remove the disk, you will get an alert. When you replace a physical disk, the disk must be acknowledged by the RAID controller before it can be used. This does not take a long time, but you can use the `LIST PHYSICALDISK` command to monitor the status until it returns to normal.

Replacing a Damaged Physical Disk (continued)

The grid disks and cell disks that existed on the previous disk in the slot will be automatically recreated on the new disk. If these grid disks were part of an Oracle ASM disk group with `NORMAL` or `HIGH` redundancy, then they will be added back to the cluster and the data will be rebalanced based on disk group redundancy and the `asm_power_limit` parameter.

Re-creating the ASM disk and rebalancing the data may take some time to complete. You can monitor the progress of these operations within ASM. You can monitor the status of the disk as reported by `V$ASM_DISK.STATE` until it returns to `NORMAL`. You can also monitor the rebalance progress using `GV$OPERATION`.

Review the following considerations when replacing a failed disk:

- If the repair timer (specified in the `DISK_REPAIR_TIME` ASM initialization parameter) has not expired, then the ASM disk could be offline (not dropped) and the disk group is yet to be rebalanced. In this case the prompt replacement of the failed disk can avoid a needless rebalance operation.
- The disk could be dropped by Oracle Automatic Storage Management (Oracle ASM), and the rebalance operation may have been successfully run. Check the Oracle ASM alert logs to confirm this. After the failed disk is replaced a second rebalance will be required.
- The disk could be dropped, and the rebalance operation is currently running. Check the `GV$ASM_OPERATION` view to determine if the rebalance operation is still running. In this case the rebalance operation following the disk replacement will be queued.
- The disk could be dropped by ASM, and the rebalance operation failed. Check the `V$ASM_OPERATION.ERROR` to determine why the rebalance operation failed. Monitor the rebalance operation following the disk replacement to ensure it runs.
- Rebalance operations from multiple disk groups can be done on different Oracle ASM instances in the same cluster if the physical disk being replaced contains grid disks from multiple disk groups. Multiple rebalance operations cannot be run in parallel on just one Oracle ASM instance. The operations will be queued for the instance.



Replacing a Damaged Flash Card

- 1 Determine the damaged flash card.



```
CellCLI> LIST PHYSICALDISK DETAIL
name: [9:0:2:0]
diskType: FlashDisk
...
slotNumber: "PCI Slot: 1; FDOM: 2"
status: critical
```



- 2 Power down the cell.



- 3 Replace the flash card.



- 4 Power up the cell.



- 5 If the card contained a flash-based grid disk, monitor ASM to confirm the re-addition of the disk.



```
SQL> SELECT NAME, STATE FROM V$ASM_DISK
SQL> SELECT * FROM GV$OPERATION
```

ORACLE

Replacing a Damaged Flash Card

Each Sun Oracle Exadata Storage Server is equipped with 4 PCI flash memory cards. Each card has 4 flash modules (FDOMs) for a total of 16 on each cell.

Identifying a damaged flash module is similar to identifying a damaged physical disk. Hardware monitoring using iLOM or a drop in performance indicated by the CALIBRATE command may indicate a problem. If an FDOM is detected to have failed, then an alert is generated. The alert message includes if any flash-based grid disks were on the flash module.

As shown in the slide, a damaged flash module can also be reported using the LIST PHYSICALDISK DETAIL command. The slotNumber attribute shows the PCI slot and the FDOM number. In this example, the status attribute indicates a critical fault.

If there were no grid disks on the flash module, then the flash module was probably being used as Exadata Smart Flash Cache. In this mode, the bad flash module results in a decreased amount of flash memory on the cell. The performance of the cell is affected proportional to the size of flash memory lost, but the database and applications are not at risk of failure.

While technically the PCI slots in a Sun Oracle Exadata Storage Server are hot-replaceable, it is recommended to power down the cell while servicing a damaged flash card. After replacing the card and powering up the cell, no additional steps are required to recreate any flash-based grid disks. Optionally, you can monitor ASM to confirm the re-addition of a flash-based grid disk.



Moving All Disks from One Cell to Another



1. Make the grid disks inactive:
`CellCLI> ALTER GRIDDISK ALL INACTIVE`
2. Back up the operating system configuration files that will change when the new cell is booted.
3. Move the disks from the original cell to the new cell.
 - Ensure the system disks occupy the first two slots.
4. Boot the new cell.
5. Restart Exadata cell services:
`CellCLI> ALTER CELL RESTART SERVICES ALL`
6. Import the cell disks:
`CellCLI> IMPORT CELLDISK ALL`

ORACLE

3 - 30

Copyright © 2009, Oracle. All rights reserved.

Moving All Disks from One Cell to Another

You may need to move all drives from one Exadata storage server to another Exadata storage server. This may be necessary when there is a chassis-level component failure, or when troubleshooting a hardware problem. To move the drives, perform the following procedure:

1. If possible, use the `ALTER GRIDDISK ALL INACTIVE` command to make the grid disks inactive.
2. If possible, back up the files in `/etc/hosts`, `/etc/modprobe.conf`, and `/etc/sysconfig/network`. This is a precautionary step if you wish to retain the settings associated with your original Exadata storage server in case you plan to move the disks back to the original Exadata storage server in the future.
3. Move the disks from the original Exadata cell to the new Exadata cell.
Caution: Ensure the first two disks, which are the system disks, are in the same first two slots. Failure to do so will cause the Exadata cell to not function properly.
4. Start the cell. The cell operating system will be automatically reconfigured to suit the new storage server hardware.
5. Restart the cell services using `ALTER CELL RESTART SERVICES ALL`.
6. Import the cell disks using `IMPORT CELLDISK ALL`.

If you are using ASM redundancy and the procedure is completed before the amount of time specified in the `DISK_REPAIR_TIME` ASM initialization parameter, then the ASM disks will be automatically brought back online and updated with any changes made during the cell outage.

Quiz

Grid disks are seen by ASM by using a discovery string that starts with:

1. c/*
2. o/*
3. g/*
4. e/*

ORACLE

3 - 31

Copyright © 2009, Oracle. All rights reserved.

Answers: 2

Quiz

The first grid disk you create on your cell uses the slowest tracks of the corresponding physical disk.

1. TRUE
2. FALSE

ORACLE

3 - 32

Copyright © 2009, Oracle. All rights reserved.

Answers: 2

Quiz

When you create a disk group for which you want Exadata smart storage capabilities, what three attributes must you specify?

1. `compatible.rdbms`
2. `compatible.asm`
3. `au_size`
4. `disk_repair_time`
5. `cell.smart_scan_capable`

ORACLE®

Answers: 1, 2, 5

Summary

In this lesson, you should have learned how to:

- Perform the initial Exadata Storage Server boot sequence
- Create and configure an Exadata cell
- Create and configure ASM disk groups using Exadata storage
- Use the Exadata administration tool CellCLI
- Handle probable maintenance scenarios

ORACLE

Demonstration Overview

The demonstrations associated with this lesson cover the following topics:

- A
- B
- C

ORACLE[®]