

Introducción a Oracle Exadata

Documento generado por

Hector Ulloa Ligarius

Para el sitio



Índice

1.	Introducción.....	2
2.	¿Qué es Oracle Exadata?	3
2.1.	COMPONENTES TÉCNICOS.....	4
2.1.1.	Database servers X2-2.....	4
2.1.2.	Exadata Storage servers X2-2	4
2.1.3.	Infiniband switches	5
2.2.	DESCRIPCIÓN DE COMPONENTES	6
2.2.1.	Nodos de bases de datos (Database nodes).....	6
2.2.2.	Nodos de Storage (Storage Cells).....	6
2.2.3.	Disks	7
2.2.4.	Flash Disks	7
2.2.5.	Infiniband Switch	8
2.2.6.	Ethernet Switch.....	8
3.	Exadata Storage Server	9
3.1.	OFFLOADING OF DATA SEARCH	9
3.2.	SMART SCAN	9
3.3.	iDB	9
3.4.	STORAGE INDEXES	10
3.5.	STORAGE CENTRALIZADO.....	10
3.6.	I/O RESOURCE MANAGEMENT	11
3.7.	OFFLOADING OF INCREMENTAL BACKUP	11
3.8.	SMART CACHE	11
4.	Estructura jerárquica de los discos	12

1. Introducción

El siguiente documento explica de una forma muy superficial las principales características de Oracle Exadata, ya sea su arquitectura y estructuras internas.

Además se mencionan las principales características de su componente estrella, el Exadata Storage Server, que en el fondo marca la diferencia con las bases de datos actuales.

Se mencionan los componentes técnicos y al final del documento se hace una breve reseña , como se estructuran jerárquicamente los distintos Storage Server para llegar a construir una base de datos.

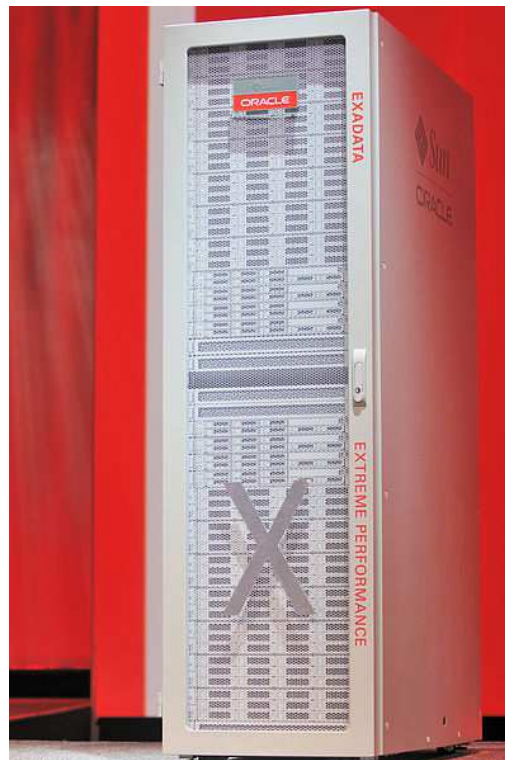
2. ¿Qué es Oracle Exadata?

Oracle Exadata es una máquina (rack) que da soporte de alta performance para aplicaciones OLTP y cargas OLAP.

En un principio fue un trabajo entre Oracle Corporation y Hewlett Packard, Oracle diseñaba todo lo que era base de datos y colocaba el sistema operativo (OEL), lo que correspondía a Storage era parte de HP junto con la arquitectura de esta máquina. Esto fue un primer release, a los meses Oracle compraba Sun, con lo cual anuncia una versión de su Oracle Exadata, donde deja fuera a HP y ocupa todas las tecnologías de Sun Microsystems.

Hoy en día Oracle Exadata se distribuye con elección de Sistema Operativo, ya sea, OEL o Solaris 11 Express.

Oracle Exadata luce de esta forma.



Oracle Exadata es un rack que junta una serie de componentes, los cuales conforman la gran infraestructura de base de datos, entre esos componentes se encuentran discos, servidores, networking, etc.

Siendo un poco más técnicos, los servidores de base de datos tienen la siguiente composición

2.1. Componentes técnicos

2.1.1. Database servers X2-2

(Sun Fire X4170 M2)

- 2 x Six-Core Intel Xeon X5675 Processors (3.06 GHz)
- 96 GB Memory (expandable to 144 GB with optional memory expansion kit)
- Disk Controller HBA with 512MB Battery Backed Write Cache
- 4 x 300 GB 10,000 RPM SAS Disks
- 2 x QDR (40Gbit/s) Ports
- 2 x 10 Gb Ethernet Ports based on the Intel 82599 10GbE Controller
- 4 x 1 Gb Ethernet Ports
- 1 x ILOM Ethernet Port
- 2 x Redundant Hot-Swappable Power Supplies
- 3 x 36 port QDR (40 Gbit/s) InfiniBand Switches (2 x in Quarter Rack configuration)

Los servidores de Storage

2.1.2. Exadata Storage servers X2-2

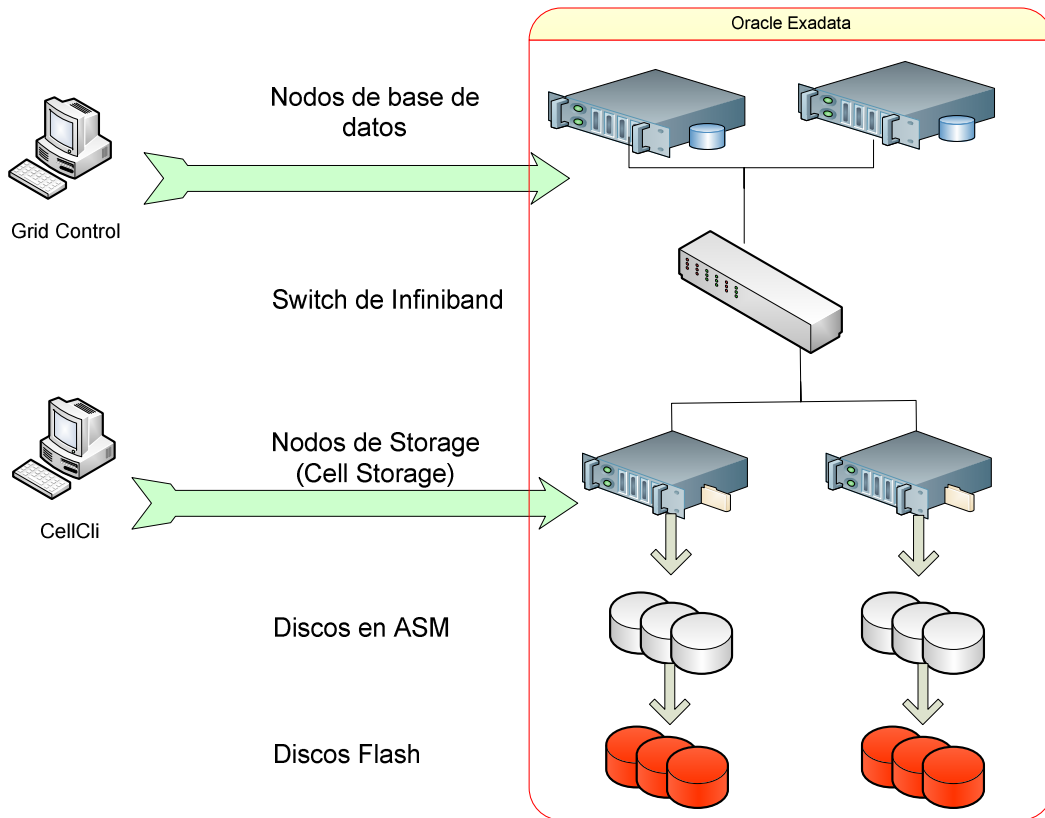
- 2 x Six-Core Intel Xeon L5640 (2.26 GHz) Processors
- Exadata Smart Flash Cache 384 GB
- System Memory 24 GB
- Disk Controller Disk Controller HBA with 512MB Battery Backed Write Cache
- InfiniBand Connectivity Dual-Port QDR (40Gbit/s) InfiniBand Host Channel Adapter
- Power Supplies Dual-redundant, hot-swappable power supply
- Remote Management Sun Embedded Integrated Lights Out Manager (ILOM)
- Disk Drives 12 x 600 GB 15,000 RPM High Performance SAS or
- 12 x 3 TB 7,200 RPM High Capacity SAS
- Integrated Lights Out Manager (ILOM) Ethernet port

Y los switches presentes en el Exadata

2.1.3. Infiniband switches

- Sun Datacenter Infiniband Switch 36
- 36 ports

Si quisiéramos mostrar un dibujo de los componentes, sería algo así



Grid Control es el componente para la administración de Oracle Exadata y Cellcli es un utilitario que está en el lado de los nodos de Storage

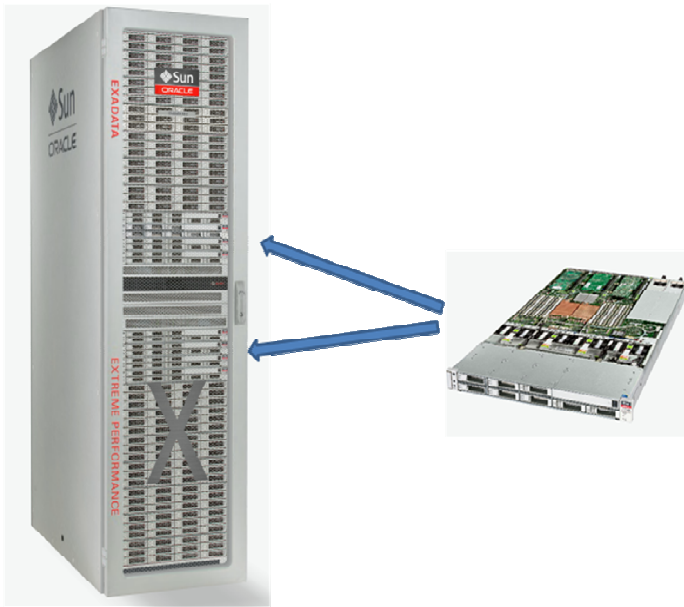
Para lo anterior existen 3 configuraciones en el mercado , que combinan nodos de base de datos, nodos de storage y switches, esas configuraciones son Full Rack , half rack o quarter rack, para los tres casos son los mismos componentes , sólo varia la cantidad.

2.2. Descripción de componentes

2.2.1. Nodos de bases de datos (Database nodes)

La base de datos y el cluster corren en los nodos conocidos como Databases nodes, en estas máquinas por defecto corre un RAC versión 11gr2 . Un full rack contiene 8 nodos de base de datos, un half rack tiene 4 nodos y un quarter rack posee 2 nodos.

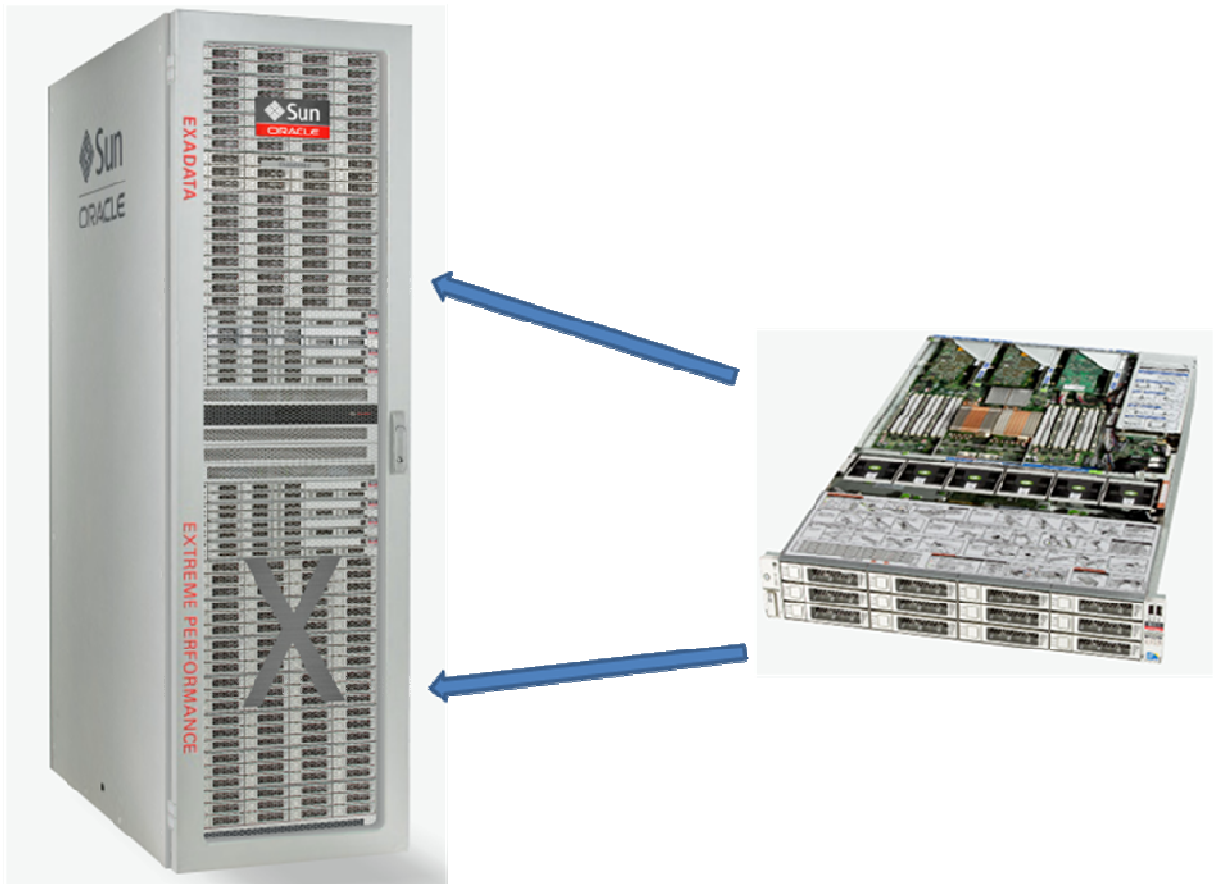
Para visualizarlo dentro del Exadata



2.2.2. Nodos de Storage (Storage Cells)

En estas máquinas es donde se encuentran atachados los discos que se utilizarán para almacenar la información de nuestras bases de datos, sobre estas máquinas corre todo el software que maneja los discos

Para visualizarlo dentro del exadata



2.2.3. Disks

Cada celda de Storage contiene 12 discos (dependiendo de la configuración), estos discos pueden ser de 600Gb o 2Tb.

2.2.4. Flash Disks

Cada celda de storage , contiene además 384Gb de flash disk, estos discos pueden ser presentados a los nodos (Database nodes) como storage o usados como cache secundaria por el cluster de base de datos, esto último es conocido como Smart Cache.

2.2.5. Infiniband Switch

Los databases nodes y las storage cell están conectadas a través de infiniband (para saber lo que es infiniband puede consultar el siguiente [link](#))

Existe 3 switches de infiniband para nuestro Oracle Exadata, lo cual claramente da una seguridad pues elimina puntos de falla único al momento de realizar las comunicaciones.

2.2.6. Ethernet Switch

Los usuarios y administradores de Oracle Exadata se comunican a través de Infiniband o a través de Ethernet.

3. Exadata Storage Server

He aquí el gran secreto de Oracle y por lo cual ha alcanzado niveles de performance increíbles, este Storage es el encargado de procesar la información solicitada desde los Database nodes y entregar la información de la manera más rápida y consistente. Es un conjunto de software que trabaja de una forma tal que trata de disminuir los tiempos de respuesta de las aplicaciones.

Los storage tradicionales tienen serios problemas a nivel de desarrollo de IT, o sea, cada vez que hay una base de datos de por medio, siempre habrá problemas de performance, sobre todo en arquitecturas clásicas de arreglos de Storage.

El problema principal, son los anchos de banda al momento de que todas las bases de datos conectadas a un storage, comienzan a realizar consultas, obviamente el recurso de ancho de banda queda corto y siempre irá en perjuicio de los tiempos de respuesta.

La gran cualidad de Exadata es que está orientado a obtener los mejores resultados posibles a nivel de tiempos de respuesta e I/O, para lo anterior los Exadata Storage Server utilizan las siguientes cualidades:

3.1. Offloading of Data Search

Esta característica tiene relación con la capacidad de procesamiento de las celdas de Storage, esto implica que ellas son capaces de pre-procesar la información para minimizar la cantidad de información que va a ser traspasada a través de la red que conecta los database nodes y los cell storage, las características que vienen a continuación, mejoran de una gran forma la información que se procesa, lo que obviamente impacta en el tráfico y en los tiempos de respuesta de las aplicaciones.

3.2. Smart Scan

En una base de datos común y corriente, cada vez que se hace una consulta por ejemplo de una columna en una fila, el bloque completo es extraído desde los datafiles y llevado a la SGA (buffer cache), indistinto de la cantidad de columnas que tenga la fila e indistinto de la cantidad de bloques que tenga el bloque Oracle.

En Oracle Exadata, lo anterior no cambia en casi nada, pero sí hay algunas cosas muy interesantes y extremadamente útiles que sí marcan la diferencia, por ejemplo Direct Path Access, full table scans y full index scans. En lo anterior se puede colocar una fila o específicamente una columna directamente desde disco y enviada a las bases de datos. Lo anterior es conocido como Smart Scan, lo anterior implica una reducción increíble de I/O.

3.3. iDB

iDB es la abreviación de Intelligent Database, el Smart Scan es la capacidad de enviar solo la información solicitada, pero para ello se debe saber que es lo que se envía, el iDB es un protocolo de comunación existente entre los databases nodes y los cell storage, para los casos en que se puede enviar solamente un pequeño grupo de información (o sea, que se puede aplicar un Smart Scan) el iDB envía los nombres de las tablas, las columnas, los predicados entre otros datos, con esta información las cell storage determinan de mejor forma la información a enviar más que solamente la dirección de los bloques a enviar, con lo anterior se envía solamente la fila o la columna en vez de enviar los bloques Oracle .

3.4. Storage Indexes

Cada Cell Exadata mantiene un Storage Index que contiene un resumen de toda la distribución de data en los discos. Esta información es mantenida de forma automática y es totalmente transparente para la base de datos.

Por cada región indexada, el storage index mantiene el mínimo y máximo valor de las columnas de una tabla (región de disco casi siempre de 1MB) , como son regiones distintas para cada Cell Storage , redundando en que el sistema es altamente escalable y nunca se producen esperas por contensiones de latch , puesto que a mayor cantidad de información , mayor la cantidad de Storage Indexes, los cuales no se deben confundir con los índices normales, que son estructuras totalmente distintas.

¿De qué sirve mantener el máximo y mínimo de las columnas indexadas? Pues bien, esto ayuda eliminando el I/O innecesario, este efecto se conoce como I/O filtering. Cada I/O que se produce en la celda de storage es almacenado en la vista V\$SYS_STAT y muestra el número de bytes de I/O que son “ahorrados” usando los Storage Index.

¿Qué consultas son mejoradas por los Storage Index? Pues todas las consultas que ejecuten cualquiera de las siguientes instrucciones :

- Igualdad (=)
- No iguales (< , != o >)
- Menor igual que (<=)
- Mayor o igual que (>=)
- Is null
- Is not null

3.5. Storage centralizado

Se puede usar Oracle Exadata Storage para centralizar todos los requerimientos de storage de una compañía, no importando la cantidad de bases de datos que lo utilicen.

Las celdas de exadata con ASM distribuyen toda la carga de I/O a través de todos los discos disponibles en el storage. Cada base de datos puede usar todos los discos disponibles con lo cual se alcanzan niveles de performance muy completos.

3.6. I/O Resource Management

I/O Resource Management de aquí en adelante IORM y Database Resource Management, permiten que múltiples bases de datos compartan el mismo storage, mientras se asegura que todos los recursos de I/O sean ocupados de buena forma a través de todas las bases de datos.

Lo anterior da como resultado que una base de datos no puede, entiéndase bien, no puede monopolizar los recursos de I/O cuando se accede a la información de las Cell Storage.

IORM es implementado y manejado a través de las políticas definidas en el Database resource management, Database resource management en una instancia de base de datos se comunica con el software de IORM en las storage cell, para manejar todas las políticas declaradas por el DBA (negocio) . Los database resource plan son manejados por la base de datos, mientras los interdatabase plans son manejados por las storage cell

3.7. Offloading of Incremental Backup

También se optimizan los respaldos a través de RMAN, dejando fuera los bloques que no sean necesarios respaldar.

Esto implica que no solamente deja fuera los bloques vacíos, si no que también deja fuera los bloques que no son necesarios en una restauración, esto se hace de forma automática y no requiere intervención del DBA.

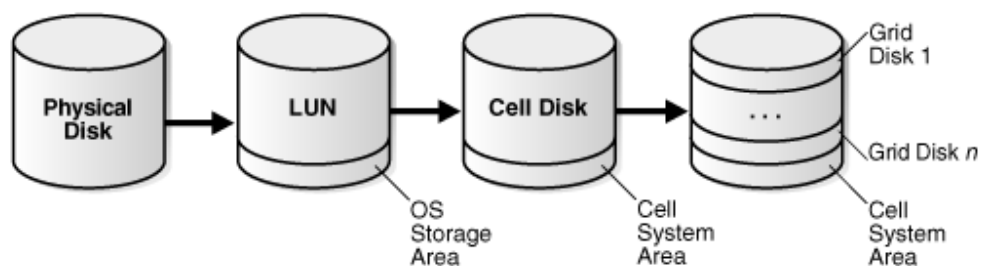
3.8. Smart Cache

El Database buffer cache es el lugar donde los bloques de datos son leídos en una primera instancia, si no se encuentran allí, pues sencillamente se leen desde disco y allí se produce el inefable I/O. Este es el caso común de la totalidad de las bases de datos Oracle pre-Exadata.

Pues bien Oracle Exadata crea un cache intermedio , entre el storage y el buffer cache de la SGA, esta cache se llama Smart Cache, esta porción de memoria almacena los datos más frecuentemente usados, claramente esto puede redundar en que muchas veces se reduce el I/O dado que la información si no esta en el buffer cache, se encuentra disponible en esta segunda área de cache.

4. Estructura jerárquica de los discos

Para entender un poco como se estructuran los discos hasta llegar a trabajar con ASM, se presenta el siguiente cuadro



El anterior cuadro refleja lo siguiente

- Oracle Exadata se compone de múltiples servidores de storage, si hablamos de un Full Rack son 14 servidores de Storage, para un half rack son 7 servidores de storage y para un quarte rack son 3 servidores de Storage
- Cada Oracle Exadata Server, tiene 12 discos (SAS o SATA de 3"5), cada uno de los tipos de discos tiene una capacidad distinta y performance distintos.
- A partir de los anteriores discos se generan LUNs (particiones)
- Las LUNs son identificadas y presentadas como Cell Disk
- Cada Cell Disk es presentada como Grid Disk (las cell disks se pueden dividir en múltiples Grid Disks)
- Los Grid Disk, son usados como ASM Disks
- Y los ASM Disks, pues usados para construir ASM Diskgroups
- Y con los ASM Diskgroups, yo puedo generar mi base de datos

Si los Diskgroups de ASM tienen redundancia normal o alta, los grupos de falla de los diskgroups, siempre se van a ubicar en distintas celdas del storage, esto implica que si una celda falla, la información estará disponible en la otra celda

También hay una variación, de cuando se requiere montar un filesystem

- Oracle Exadata se compone de múltiples servidores de storage, si hablamos de un Full Rack son 14 servidores de Storage, para un half rack son 7 servidores de storage y para un quarter rack son 3 servidores de Storage
- Cada Oracle Exadata Server, tiene 12 discos (SAS o SATA de 3"5), cada uno de los tipos de discos tiene una capacidad distinta y performance distintos.
- A partir de los anteriores discos se generan LUNs (particiones)
- Cada partición es presentada como un VOLUME
- Cada volumen es con lo que se construye un punto de montaje