Exadata for Oracle DBAs

Arup Nanda
Longtime Oracle DBA
Why this Session?

- I’m an Oracle DBA
  - Familiar with RAC, 11gR2 and ASM
- About to become a Database Machine Administrator (DMA)

**How much do I have to learn?**

- How much of my own prior knowledge I can apply?
- What’s different in Exadata?
- What makes it special, fast, efficient?
- Do I have to go through a lot of training?
What is Exadata

- Is an *appliance* containing
  - Storage
  - Flash Disks
  - Database Servers
  - Infiniband Switches
  - Ethernet Switches
  - KVM (some models)
- But is *not* an appliance. Why?
  - It contains additional software to make it a better database machine
- That’s why Oracle calls it a *Database Machine* (DBM)
- And DMA – Database Machine Administrator
Anatomy of an Oracle Database

Instance
Combination of
• Memory Areas
• Background Processes

SELECT NAME
FROM CUSTOMERS
WHERE STATUS = 'ANGRY'

UPDATE
CUSTOMERS
SET BONUS = 1M
WHERE STATUS = 'ANGRY'

Storage

datafile1
datafile2
Query Processing

SELECT NAME FROM CUSTOMERS WHERE STATUS = 'ANGRY'

Database Block

Storage
RAC Database

Oracle Clusterware (CRS)

Server 1
- Buffer cache
- DBWR
- PMON

Server 2
- Buffer cache
- DBWR
- PMON

Storage
- datafile1
- datafile2

Arup Nanda
Components for Performance

Storage has been and will continue to be the bane of all databases. Simply put, less I/O ➔ better performance.
What about SAN Caches?

- Success of SAN caches is built upon predictive analytics
- They work well, if a small percentage of disk is accessed most often
  - The emphasis is on disk; not data
- Most database systems
  - are way bigger than caches
  - need to get the data to the memory to process
    - I/O at the disk level is still high
- Caches are excellent for filesystems
  - or very small databases
What about In-Memory DBs

- Memory is still more expensive
- How much memory is enough?
- You have a 100 MB database and 100 MB buffer cache
- The whole database will fit in the memory, right?
- NO!
- Oracle database fills up to 7x DB size buffer cache

The Solution

- A typical query may:
  - Select 10% of the entire storage
  - Use only 1% of the data it gets
- To gain performance, the DB needs to shed weight
- It has to get less from the storage
  - Filtering at the storage level
  - The storage must be cognizant of the data

```sql
SELECT NAME
FROM CUSTOMERS
WHERE STATUS = 'ANGRY'
```
The communication between CPU and Disk carries the information on the query – columns and predicates. This occurs as a result of a special protocol called iDB.
Magic #2 Storage Cell Server

- Cells are Sun Blades
- Run Oracle Enterprise Linux
- Software called Exadata Storage Server (ESS) which understands iDB
Magic #3 Storage Indexes

Storage Indexes store in memory of the Cell Server the areas on the disk and the MIN/MAX value of the column and whether NULL exists. They eliminate disk I/O.

```
SELECT ... 
FROM TABLE 
WHERE COL1 = 1
```
Magic #4 Flash Cache

These are flash cards presented as disks; not memory to the Storage Cells. They are similar to SAN cache; but Oracle controls what goes on there and how long it stays.
## Components

<table>
<thead>
<tr>
<th>CPU</th>
<th>Database Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td>(Sun Blade, OEL) Oracle 11gR2 RAC</td>
</tr>
<tr>
<td>Network</td>
<td>InfiniBand Switch</td>
</tr>
<tr>
<td>I/O Controller</td>
<td>Storage Cell</td>
</tr>
<tr>
<td>Disk</td>
<td>Exadata Storage Server Disks, Flash</td>
</tr>
</tbody>
</table>

**Arup Nanda**
Put Together: One Full Rack

Clients connect to the database nodes.
Disk Layout

- Disks (hard and flash) are connected to the cells.
- The disks are partitioned at the cell.
- Some partitions are presented as filesystems.
- The rest are used for ASM diskgroups.

All these disks/partitions are presented to the compute nodes.
Disk Presentation

- **Cell**
  - Griddisk
  - Celldisk
  - LUN
  - Partition 1
  - Partition 2
  - Physical Disk

- **Node**
  - ASM Disk
  - filesystem

- **ASM Diskgroup**
  - Disk
  - filesystem
Command Components

**Linux Commands** – `vmstat`, `mpstat`, `fdisk`, etc.

**ASM Commands** – SQL*Plus, ASMCMD, ASMCA

**Database Commands** – `startup`, `alter database`, etc.

**Clusterware Commands** – CRSCTL, SRVCTL, etc.

**CellCLI** – command line tool to manage the Cell

5-part Linux Commands article series

One Cluster?

One Cluster

QA1  QA2  QA3  Prod1  Prod2  Prod3  Dev1  Int1
QA1  QA2  Prod4  Prod1  Prod2  Prod3  Dev1  Int1
QA1  QA2  QA3  Prod1  Prod2  Prod3  Dev1  Int1

Dev2
Many Clusters?
Disk Failures

Datafile

Cell 1
block1

Cell 2
block1
Other Questions

Q: Do clients have to connect using Infiniband?
A: No; Ethernet is also available

Q: How do you back it up?
A: Normal RMAN Backup, just like an Oracle Database

Q: How do you create DR?
A: Data Guard is the only solution

Q: Can I install any other software?
A: Nothing on Cells. On nodes – yes

Q: How do I monitor it?
A: Enterprise Manager, CellCLI, SQL Commands
Summary

- Exadata is an Oracle Database running 11.2
- The storage cells have added intelligence about data placement
- The compute nodes run Oracle DB and Grid Infra
- Nodes communicate with Cells using iDB which can send more information on the query
- Smart Scan, when possible, reduces I/O at cells even for full table scans
- Cell is controlled by CellCLI commands
- DMA skills = 60% RAC DBA + 15% Linux + 20% CellCLI + 5% miscellaneous
Thank You!

Arup Nanda
Blog: Arup.Blogspot.com
Twitter: arupnanda
Email: arup@proligence.com