Why this Session?

• If you are
  – an Oracle DBA
    • Familiar with RAC, 11gR2 and ASM
  – about to be a Database Machine Administrator (DMA)

• How much do you have to learn?
• How much of you own prior knowledge I can apply?
• What’s different in Exadata?
• What makes it special, fast, efficient?
• Do you 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?
  – additional software to make it a better database machine
  – Components can be managed independently

• 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
RAC Database

Oracle Clusterware (CRS)

Storage
datafile1
datafile2

Server1
Buffer cache
DBWR
PMUN

Server2
Buffer cache
DBWR
PMUN

SELECT NAME
FROM CUSTOMERS
WHERE STATUS = 'ANGRY'

Query Processing

Server
Buffer cache
DBWR
PMUN

JILL

CUSTID NAME STATUS
1 JOHNN HAPPY
2 JILL ANGRY
3 JOE HAPPY
500 JIM HAPPY

Database Block

Storage
datafile1
datafile2

Arup Nanda
Components for Performance

- CPU
- Memory
- Network
- I/O Controller
- Disk

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

```
SELECT NAME
FROM CUSTOMERS
WHERE STATUS = 'ANGRY'
```
The Magic #1

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
```

Output

<table>
<thead>
<tr>
<th>STAT_NAME</th>
<th>STAT_VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI Savings</td>
<td>5120.45</td>
</tr>
<tr>
<td>Smart Scan</td>
<td>1034.00</td>
</tr>
</tbody>
</table>
Why Not?

- Pre-requisite for Smart Scan
  - Direct Path
  - Full Table or Full Index Scan
  - > 0 Predicates
  - Simple Comparison Operators
- Other Reasons
  - Cell is not offload capable
    - The diskgroup attribute `cell.smart_scan_capable` set to FALSE;
  - Not on clustered tables, IOTs, etc.

Disabling Smart Scans

```sql
cell_offload_processing = true;
_kcfis_storageidx_disable d = true;
```

---

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.
Magic #5 Process Offloading

• Bloom Filters
• Functions Offloading
  – Get the functions that can be offloaded
    • V$SQLFN_METADATA
• Decompression
  – (Compression handled by Compute Nodes)
• Virtual Columns

Components

CPU

Memory

Network

I/O Controller

Disk

Database Node
(Sun Blade. OEL)
Oracle 11gR2 RAC

InfiniBand Switch

Storage Cell
Exadata Storage Server
Disks, Flash
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
**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  

4-part Exadata Command Reference article series  
Administration Skills

<table>
<thead>
<tr>
<th>Skill</th>
<th>Needed</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Administrator</td>
<td>15%</td>
</tr>
<tr>
<td>Storage Administrator</td>
<td>0%</td>
</tr>
<tr>
<td>Network Administrator</td>
<td>5%</td>
</tr>
<tr>
<td>Database Administrator</td>
<td>60%</td>
</tr>
<tr>
<td>Cell Administration</td>
<td>20%</td>
</tr>
</tbody>
</table>

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

Cell 1

Cell 2

Datafile

block1

block1

Exadata Demystified
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
Resources

• My Articles

• OTN Page on Exadata

• Tutorials

• OTN Exadata Forum

• Exadata SIG

Thank You!

My Blog: arup.blogspot.com
My Tweeter: arupnanda