Exadata:
from Beginner to Advanced
in 3 Hours

Arup Nanda
Longtime Oracle DBA
(and now DMA)
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 your own prior knowledge can I 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

- It is like an *appliance* containing
  - Storage, Flash Disks, Database Servers, Infiniband Switches, Ethernet Switches, KVM (some models)
- But it 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

SELECT NAME
FROM CUSTOMERS
WHERE STATUS = 'ANGRY'

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

Exadata: from Beginner to Advanced in 3 Hours
RAC Database

Oracle Clusterware (CRS)

Exadata: from Beginner to Advanced in 3 Hours

Arup Nanda
SELECT NAME FROM CUSTOMERS WHERE STATUS = 'ANGRY'
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

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

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
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

Exadata: from Beginner to Advanced in 3 Hours
Checking Storage Index Use

```sql
select name, value/1024/1024 as stat_value
from v$mystat s, v$statname n
where s.statistic# = n.statistic#
and n.name in ('cell physical IO bytes saved by storage index', 'cell physical IO interconnect bytes returned by smart scan')
```

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>
Checking Offloading of an SQL

```
select
    sql_id,
    child_number child#,
    plan_hash_value plan_hash,
    executions execs,
    (elapsed_time/1000000)/decode(nvl(executions,0),0,1,executions)/
    decode(px_servers_executions,0,1,px_servers_executions/decode(nvl(executions,0),0,1,executions)) avg_elapsed_time_in_secs,
    px_servers_executions/decode(nvl(executions,0),0,1,executions) avg_par_deg,
    decode(io_cell_offload_eligible_bytes,0,'No','Yes') Offloaded,
    decode(io_cell_offload_eligible_bytes,0,0,100*(io_cell_offload_eligible_bytes-io_interconnect_bytes)/decode(io_cell_offload_eligible_bytes,0,1,io_cell_offload_eligible_bytes)) "%age IO Saved",
    buffer_gets/decode(nvl(executions,0),0,1,executions) avg_lio
from v$sql
where sql_text like '<SQL Statement Comes Here>%'```

Exadata: from Beginner to Advanced in 3 Hours
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

```plaintext
cell_offload_processing = false;
_kcfis_storageidx_disabled = true;
```

Exadata: from Beginner to Advanced in 3 Hours
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

- **Database Node**
  (Sun Blade, OEL)
  Oracle 11gR2 RAC

- **InfiniBand Switch**

- **Storage Cell**
  Exadata Storage Server
  Disks, Flash

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

Exadata: from Beginner to Advanced in 3 Hours
Put Together: One Full Rack

Clients connect to the database nodes.

Exadata: from Beginner to Advanced in 3 Hours
## How it Looks

<table>
<thead>
<tr>
<th></th>
<th>X2-2 Qtr</th>
<th>X2-2 Half</th>
<th>X2-2 Full</th>
<th>X2-8 Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Compute Nodes</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Total Compute Node Processor Cores</td>
<td>24</td>
<td>48</td>
<td>96</td>
<td>160</td>
</tr>
<tr>
<td>Total Compute Node Memory</td>
<td>196 GB</td>
<td>384 GB</td>
<td>768 GB</td>
<td>4 TB</td>
</tr>
<tr>
<td>Number of Storage Servers</td>
<td>3</td>
<td>7</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Number of SAS Disks in Storage</td>
<td>36</td>
<td>84</td>
<td>168</td>
<td>168</td>
</tr>
<tr>
<td>Storage Capacity - HP</td>
<td>21.6 TB</td>
<td>50.4 TB</td>
<td>100.8 TB</td>
<td>100.8 TB</td>
</tr>
<tr>
<td>Storage Capacity - HC</td>
<td>108 TB</td>
<td>252 TB</td>
<td>504 TB</td>
<td>504 TB</td>
</tr>
<tr>
<td>Number of InfiniBand Switches</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Source: upcoming book Exadata Recipes by Clarke from Apress

Arup Nanda
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.

5-part Linux Commands article series

http://bit.ly/k4mKQS

4-part Exadata Command Reference article series

http://bit.ly/lljFl0

Exadata: from Beginner to Advanced in 3 Hours
## 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>

Exadata: from Beginner to Advanced in 3 Hours
One Cluster?

Exadata: from Beginner to Advanced in 3 Hours

Arup Nanda
Many Clusters?
Disk Failures

Cell 1

Cell 2

Datafile

block1

block1

block1

block1

Exadata: from Beginner to Advanced in 3 Hours
Disk Failures

Cell 1
- block1

Cell 2
- block1

Cell 3
- block1

Datafile
- block1

High Redundancy
Playing Nice

• Database Resource Manager
• I/O Resource Manager
• Cell Fencing

Compute Nodes

Storage Cells

Exadata: from Beginner to Advanced in 3 Hours
<table>
<thead>
<tr>
<th>Category</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database</td>
<td>DBA</td>
</tr>
<tr>
<td>Machine</td>
<td>System Admin</td>
</tr>
<tr>
<td>Network</td>
<td>Network Admin</td>
</tr>
<tr>
<td>Storage</td>
<td>??</td>
</tr>
</tbody>
</table>

Exadata: from Beginner to Advanced in 3 Hours

Arup Nanda
Combined Skills

New Role

Database Machine Administrator (DMA)

- Database
- Machine
- Network
- Storage

Exadata: from Beginner to Advanced in 3 Hours

Arup Nanda
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
Backup and DR

- No SAN connectivity
- Only NAS
  - Infiniband
  - Tape, Disk Pool
- DR
  - No Storage Level Replication
  - Only Data Guard
  - Supplemental Logging
  - Force Logging
- Golden Gate

Exadata: from Beginner to Advanced in 3 Hours
Overall Activities

• Physical Aspects
  – Delivery, power, network components, etc.

• Layout Planning

• Installation and Configuration

• Data Migration

• Administration
  – Who manages it
  – Backup and Disaster Recovery

• Application Development
Power Delivery Units

• Over or Under the unit

• Power Requirement
  – Single-Phase Low Voltage Americas / Japan / Taiwan
  – Single-Phase High Voltage (EMEA & APAC (excluding Japan / Taiwan)
  – Three-Phase Low Voltage (Americas / Japan / Taiwan)
  – Three-Phase High Voltage (EMEA & APAC (excluding Japan / Taiwan)
Network Ports

- **NET0**
  - Admin Interface

- **NET1, NET2**
  - Network Access to Nodes

- **NET3**
  - Backup Network

- **IB**
  - Infiniband Network
  - IP Addr: Qtr Rack: 5; Half Rack: 11; Full rack: 22
Installation Activities

1. Configuration Worksheet
2. Pre-delivery Survey
3. Generate config files
4. Run checkip.sh
5. Power on and validate components
6. Configure KVM
7. Configure IB
8. Configure Cisco Switch
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9.</td>
<td>Configure IP to PDUs</td>
</tr>
<tr>
<td>10.</td>
<td>Validate Storage Cells</td>
</tr>
<tr>
<td>11.</td>
<td>Validate Compute Nodes</td>
</tr>
<tr>
<td>12.</td>
<td>Config files from USB</td>
</tr>
<tr>
<td>13.</td>
<td>Firstboot and applyconfig.sh</td>
</tr>
<tr>
<td>14.</td>
<td>Stage Oracle Software on Node 1</td>
</tr>
<tr>
<td>15.</td>
<td>Run OneCommand</td>
</tr>
</tbody>
</table>
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
  – 5-part Linux Commands article series http://bit.ly/k4mKQS
  – 4-part Exadata Reference article series http://bit.ly/IljFl0

• OTN Page on Exadata

• Tutorials

• OTN Exadata Forum

• Exadata SIG
  – http://www.linkedin.com/groups?home=&gid=918317
Thank You!

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