Exadata 101 - What You Need to Know!

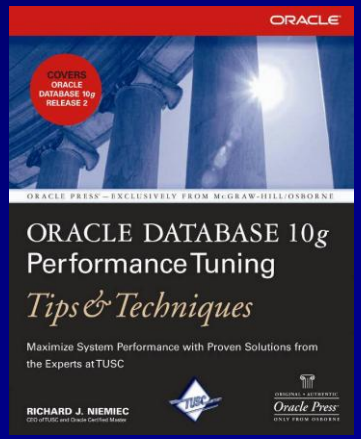
New York Oracle Users Group - 2011

Rich Niemiec, Rolta TUSC

(Thanks: Jim Viscusi, Milton Wan, Damon Grube, Mike Messina, Sri Avantsa, & Shyam Varan Nath + Oracle Learning Library for Examples)

Rich's Overview (rich@tusc.com)

- Advisor to Rolta International Board
- Former President of TUSC
  - Inc. 500 Company (Fastest Growing 500 Private Companies)
  - 10 Offices in the United States (U.S.); Based in Chicago
  - Oracle Advantage Partner in Tech & Applications
- Former President Rolta TUSC & President Rolta EICT International
- Author (3 Oracle Best Sellers – #1 Oracle Tuning Book for a Decade):
  - Oracle Performing Tips & Techniques (Covers Oracle7 & 8i)
  - Oracle9i Performance Tips & Techniques
  - Oracle Database 10g Performance Tips & Techniques
- Former President of the International Oracle Users Group
- Current President of the Midwest Oracle Users Group
- Chicago Entrepreneur Hall of Fame - 1998
- E&Y Entrepreneur of the Year & National Hall of Fame - 2001
- IOUG Top Speaker in 1991, 1994, 1997, 2001, 2006, 2007
- MOUG Top Speaker Twelve Times
- National Trio Achiever award - 2006
- Oracle Certified Master & Oracle Ace Director



Overview

- Terminology & the Basics about Exadata
- Flash Cache
- Storage Index
- Smart Scans
- Hybrid Columnar Compression (HCC)
- Enterprise Manager & Grid Control
- Enterprise Manager Exadata Simulation
- I/O Resource Manager
- Security
- Utilities
- Best Practices
- Summary

Goals ...

- Goals
  - Overview of Exadata - Easy
- Non-Goals
  - Making you the Exper

Audience Knowledge

Exadata V1?

X2-8?

10x faster than any Oracle DW

Exadata V2-2?

5x faster than V1

Big Difference… Much Improved!

Exadata Version 1?

Exadata Version

Audience Knowledge

Full RACK?

Half or Quarter

Terminology & The Basics

Some Terms

- SATA Disk (337T)–

Big & Slow – Like a 33 1/3 <7200 RPM>

- SATA=Serial Advanced Technology Attachment

- SAS Disk (100T) –

Small & Fast – Like a 45 <15K RPM>

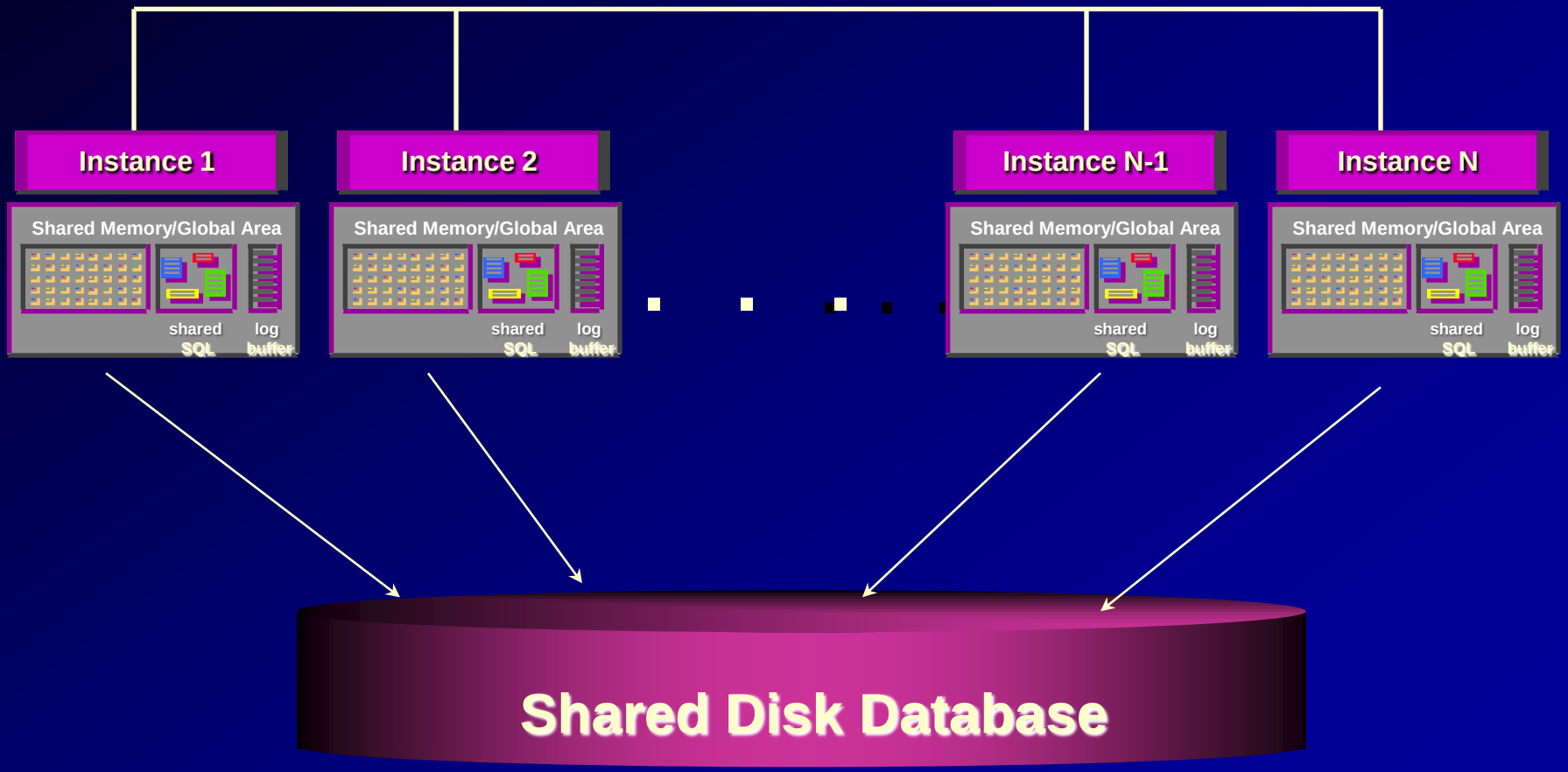
WHAT is it?

- A prebuilt 8-Node RAC cluster with Super SA
- All the CPU power you need (64 cores)
- Mega DRAM Server Memory (576G)
- Super-Mega Flash Memory (5.3T)
- Super fast interconnect (40Gb/s)
- 100T of SAS disk (28T useable)
- Database could be MUCH larger with compression!

- **If you need it & can afford it – You want it!!**

Introduction to RAC - Shared Data Model

Exadata puts it back into One Machine



How BIG is it?

- 64 Cores (16 quad core CPUs) on compute server +112 cores on storage servers (+28 quad core) = **176 cores** total-full rack
- 576G server & 400G of useable server **DRAM (100G/sec)**
- 5.3T of **flash cache (50G/sec)**
- 100T SAS disk (**28T useable**) –15K RPM (21G/s; 50K IO/s)

OR

- 336T disk space (**100T useable**) – SATA 7.2K RPM

- SATA = Serial Advanced Technology Attachment

How FAST can it be?

- ALL Disks Combined:
  - SAS – 21G/s (**50,000 IOPS** = 300 IOPS x 12 disk x 14)
  - SATA – 12 G/s (**20,000 IOPS**)
- ALL Flash Cache Combined (3.6G/s per cell):
  - 50G/s (**1,000,000 IOPS**); <20x more random I/O; 2x seq)
- Max Data Bandwidth with Disk + Cache + Compress:
  - **500G/s (10x compression)**
- Data Load Rate:

How FAST is it?

Compared to the competition:

- 5 – 100x for Data Warehousing
- 20x faster for OLTP

- Also - Miscellaneous:
  - Hot Swappable **Redundant Power**
  - Each Database Server - Dual Port InfiniBand 40Gb/s card
  - Database Servers have Disk Controller HBA (Host Bus Adapter) has 512M battery backed up cache
  - **Each DB Server has 4 x 1GbE interfaces & ILOM** (Integrated Lights Out Management – Remote power on)

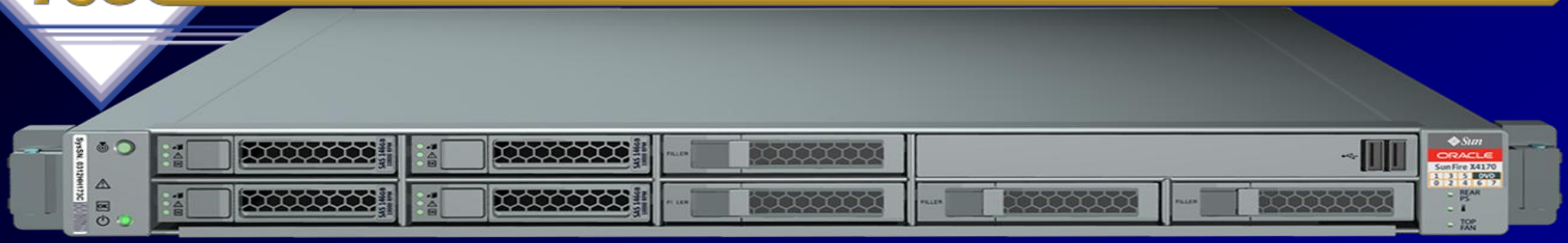
What's Making it FAST?

- Fast Hardware!
- Many CPUs
- Flash Cache
- Lot's of DRAM (Parallel Query in DRAM in 11.2)
- Compression (save 10x-70x)
- Partition Pruning (save 10x-100x)
- Storage Indexes (save 5x-10x)
- Smart Scan (save 4x-10x)
- Turn a 1T search into a 500M search or even

How they got these NUMBERS?

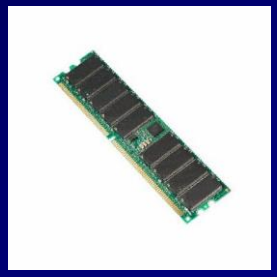
- 8 compute servers (x4170's)
  - 8 servers x 2 CPU sockets x 4 cores = 64 cores (E5540 2.53 GHz)
- 8 servers x 72G DRAM = 576G DRAM (400G useable)
- 14 Storage Servers total 336G DRAM = 912G Total DRAM
- 3 InfiniBand Switches x 36 ports = 108 ports
- 14 Storage Servers (100-336T) with Flash Cache (5T+)
  - 96G x 4 banks = 394G flash cache per storage server
  - 14 storage servers x 394G = 5.376T Flash Cache
  - 12 disks per storage server x 14 servers = 168 disks
  - 168 disks x 600G SAS = 101T SAS
  - 168 disks x 2T SATA = 336T SATA
  - Additional total storage of 4.672T on Database Servers (146G 16 drives)

Compute Servers – Like 8 Node RAC!



- 8 compute servers (x4170's)
  - 8 servers x 2 CPU sockets x 4 cores = **64 cores**
- 8 compute servers x 72G (18x4G) DRAM = **576G DRAM**
- 4 x 146G drives x 8 = **4.67T** (in addition to storage servers)

x8



DRAM (4G each)

Storage Servers – Full Rack

- 14 Storage Servers (x4275's) with Flash Cache
  - 96G x 4 cards = 394G per storage server of flash cache
  - 24Gx14= **336G of DRAM** (in addition to database servers)
  - 14 storage servers x 394G = **5.376T Flash Cache**
  - 12 disks per storage server x 14 servers = **168 disks**
  - 168 disks x 600G SAS = **101T SAS**
  - 168 disks x 2T SATA = **336T SAS**

12 Disks Hot Swappable

Flash Cache (96G each)

InfiniBand - 40G/s Each way

- 3 InfiniBand Switches x 36 ports = 108 ports
- Leaf and spine switches wired at factory depending on needs and how many Racks you'll have – careful!

36 Ports

Put it all together – Oracle's picture of the Sun Oracle Database Machine

8 Compute Servers

- 8 x 2 sockets x 4 cores = 64 cores
- 576 GB DRAM

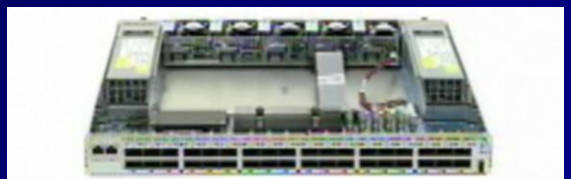
14 Storage Servers

- 14x12=168 Disks
- 100T SAS or
- 336T SATA

- 5TB+ flash storage!

InfiniBand Network

- 40 Gb/sec each direction
- Fault Tolerant



One more time... How they got these NUMBERS?

- 8 compute servers (x4170's)
  - 8 servers x 2 CPU sockets x 4 cores = 64 cores
- 8 compute servers x 72G DRAM = 576G DRAM
- 3 InfiniBand Switches x 36 ports = 108 ports
- 14 Storage Servers with 112 CPU cores & Flash Cache
  - 96G x 4 banks = 394G DRAM per storage server
  - 14 storage servers x 394G = 5.376T Flash Cache
  - 12 disks per storage server x 14 servers = 168 disks
  - 168 disks x 600G SAS = 101T SAS

How will the X2-8 change these … How they got these NUMBERS?

- 2 compute servers (7560 CPU at 2.26 GHz & 5T SAS)
  - 2 servers (x4800's) x 8 CPU sockets x 8 cores = 128 cores
- 2 compute servers x 1T DRAM = 2T DRAM
- 3 InfiniBand Switches x 36 ports = 108 ports
- 14 Storage Servers with 112 CPU cores & Flash Cache
  - 96G x 4 banks = 394G DRAM per storage server
  - 14 storage servers x 394G = 5.376T Flash Cache
  - 12 disks per storage server x 14 servers = 168 disks

Where did all my disk space go?

- Lost Space:
  - 100T SAS = 28T usable
  - 336T SATA = 100T usable

- Apply some compression & get it back:
  - 28T usable x 10 = 280T SAS
  - 100T usable x 10 = 1P SATA

Full Rack or start with ½ or ¼

	Full	Half	Quarter
Compute Servers/cores	8/64	4/32	2/16
Storage Servers/disks*	14/168	7/84	3/36
Storage SAS /IOPs 21.6T	100T	50T	
Storage SATA	336T	168T	72T
Flash IOPs (max)	1,000,000	500,000	225,000
InfiniBand Switches	3	2**	2
Data Load Rates 1T/hr	5T/hr	2.5T/hr	

Benefits Multiply*

10 TB of user data
Requires 10 TB of IO

1 TB
with compression

100 GB
with partition pruning

20 GB
with Storage Indexes

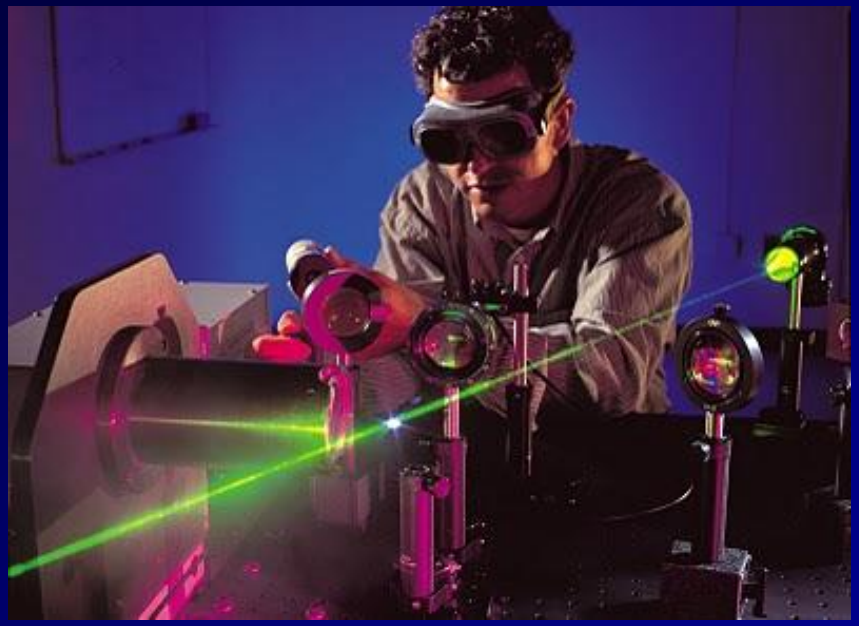
5 GB
with Smart Scans

Sub second
On Database
Machine

Data is 10x Smaller, Scans are 2000x faster

**Oracle Slide – Thanks!*

Smart Scans



Smart Scans – 10x savings common

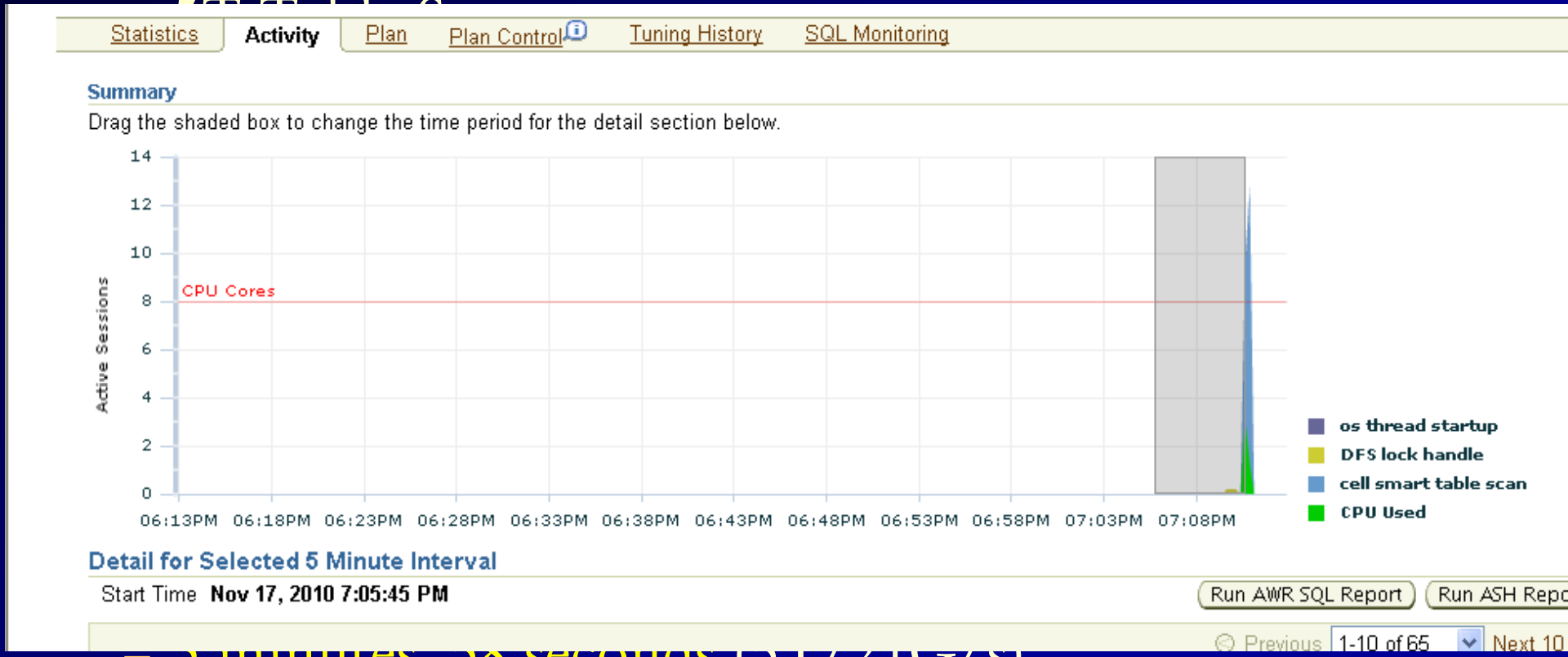
- HARDWARE Scans with **NO Code Change**:
  - Filters based on WHERE clause (predicates)
  - Filters on row / column / join condition
  - Incremental Backup Filtering
- Works with:
  - Uncommitted data
  - Locked rows
  - Chained rows
  - Compressed Data
  - Encrypted Data (11.2)
- You can SEE the benefit with Grid Control (OEM)

Smart Scans

- Bloom Filters used for Join Filtering
  - A way to quickly search for matches (simplistic meaning)
  - Saves space & is transparent to the user
  - Makes things faster – hardware level filtering
  - Tests if elements to search for are in a set
  - Many types out there including *Bloomier* filters
  - False positives are possible (rechecks to be perfect – addl. disk)
  - Google BigTable uses Bloom filters to reduce disk lookups

Oracle performance test…

- Without Smart Scan (Push whole table via network)



The SMART Flash Cache

ALL Flash Cache Combined (3.6G/s per cell): 50G/s (1,000,000 IOPS)
20x more random I/O; 2x more sequential I/O (vs. disk)

Flash Cache – 20x-50x faster than disk

- Caches HOT Data – Does as LAST step!
- PCIe based Flash cards (PCI = Peripheral Component Interconnect express)
- Knows which objects NOT to cache (FTS)
- Can specify WHAT you want to cache
  - STORAGE (CELL_FLASH_CACHE KEEP)
  - Table/Partition level with CREATE or ALTER
- Write through caches is used to accelerate reads – Data written to disk also written to cache for future reads.

Flash Cache

- Caches
  - Hot Data/Index Blocks
  - Control File reads/writes
  - File header reads/writes
- Does NOT cache
  - Mirror copies / Backups / Data Pump
  - Tablespace Formatting
  - Table Scans (rare)

24G x 4 doms = 96G (dom = disk on module – "*solid state*")

96G x 4 flash cards = 394G per storage server of flash

Flash Cache LRU

- CELL_FLASH_CACHE storage clause
  - DEFAULT (normal – large I/O's not cached)
  - KEEP (use flash cache more aggressive / May not occupy > 80%)
  - NONE (flash cache not used)
- CACHE (NOCACHE) Hint
  - I/O cached/not-cached in the flash cache
  - *SELECT /*+ CACHE */ …*
- EVICT Hint – Data removed from the flash cache
- ASM rebalance data is evicted from cache when done
- Large I/O (Full Table Scans) on objects with CELL_FLASH_CACHE set to DEFAULT are not

Using the KEEP cache

```
ALTER TABLE CUSTOMER
STORAGE (CELL_FLASH_CACHE KEEP);

Table Altered.


SELECT      TABLE_NAME, TABLESPACE_NAME,
            CELL_FLASH_CACHE
FROM        USER_TABLES
WHERE       TABLE_NAME = 'CUSTOMER';

TABLE_NAME          TABLESPACE_NAME          CELL_FL
------------        --------------------     -------
CUSTOMER            R_TEST                   KEEP
```

How it works...

- DB Request comes to CELLSRV (Cell storage server)
- CELLSRV (first time) gets data from disk
  - Data cached based on settings, hints ... etc.
  - Data to WRITE may also be cached after written if it is deemed that it may be needed again.
- CELLSRV (next time) checks:
  - In Memory Hash Table that lists what is cached
  - If cached – goes to flash cache
  - In not cached ...may cache based on settings...etc.

- CELLCLI> list flashcache detail (allows you to monitor)
- CELLCLI> list flachcachecontent where ObjectNumber=62340 detail

(Select DATAOBJ# – from obj$ where name = 'CUSTOMER';)

Is it working for me...

```
SELECT      NAME, VALUE
FROM        V$SYSSTAT
WHERE       NAME IN (
            'physical read total IO requests',
            'physical read requests optimized');
```

```
Name                                                  Value
----------------------------------------------------  --------
physical read total IO requests                       36240
physical read requests optimized                      23954
```

(this second line (*8192) is flash cache used)

It IS working ... 4G query

```
SELECT          NAME, VALUE, VALUE*8192 VALUE2
FROM            V$SYSSTAT
WHERE           NAME IN (
                'physical read total IO requests',
                'physical read requests optimized');
```

```
NAME                                VALUE       VALUE2
----------------------------------- ----------- ---------------
physical read total IO requests     10,862,844  88,988,418,048
physical read requests optimized     2,805,003  22,978,584,576
run2.....
physical read total IO requests     11,320,185  92,734,955,520
physical read requests optimized     3,203,224  26,240,811,008
run4 .....
physical read total IO requests     11,993,845  98,253,578,240
physical read requests optimized     3,793,000  31,072,256,000
```

It IS working ... V$SQL

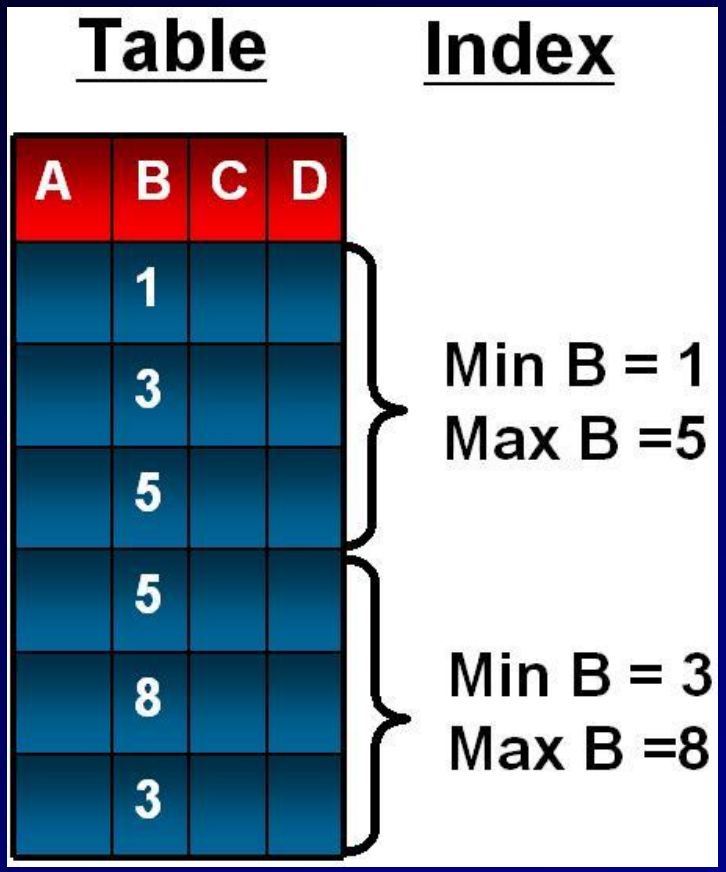
```
Select      sql_text, optimized_phy_read_requests,
            physical_read_requests,
            io_cell_offload_eligible_bytes
from        v$sql
where       sql_text like '%FIND YOUR SQL%'
```

```
SQL_TEXT     OPTIMIZED_PHY_READ_REQUESTS PHYSICAL_READ_REQUESTS
------------ --------------------------- ----------------------
IO_CELL_OFFLOAD_ELIGIBLE_BYTES
------------------------------
SELECT....   567790                      688309
4.2501E+10
Run 2.....
SELECT...    762747                      906729
4.9069E+10
run 4 ....
SELECT...    1352166                     1566537
6.8772E+10
```

FYI... NOT DB Flash Cache

- Note: Exadata PCIe card Smart Flash Cache (Exadata hardware PCIe Card Cache stored) is NOT the same as 11gR2 Database Flash Cache (file stored) that can be used with Oracle Enterprise Linux (OEL) and Solaris. In Database Flash Cache, a file can be used as data is aged out of SGA.
- To learn more about the 11gR2 Database Flash Cache, see the initialization parameters db_flash_cache_size & db_flash_cache_size.

Storage Indexes (11.2)



** *Thanks Oracle for this image*

Storage Index - 10x is common (11.2)

- Storage Indexes maintain summary information about the data– *(like Meta Data in a way)*
- A CELL LEVEL (storage) Memory Structure
- Groups things into Min/Max for various columns
- Eliminates I/Os where there is no match
- Transparent to the user
- Done at the hardware level
- Typically one index for every 1M of disk
- NOT like a B-Tree Index…more like partition elimination to skip data NOT meeting conditions
- 100% done by Oracle – **NO COMMANDS NEEDED!!**

Is it working for me...

```
SELECT      NAME, VALUE
FROM        V$SYSSTAT
WHERE       NAME LIKE ('%storage%');

NAME                                                VALUE
--------------------------------------------------  -------
cell physical IO bytes saved by storage index       25604736
```

(actual savings from Exadata built storage index)

Check BOTH servers...

```
SELECT      NAME, VALUE
FROM        GV$SYSSTAT
WHERE       NAME LIKE ('%storage%');

NAME                                                  VALUE
----------------------------------------------------- -----------
cell physical IO bytes saved by storage index 19693854720
cell physical IO bytes saved by storage index               0
```

(actual savings from Exadata built storage index)

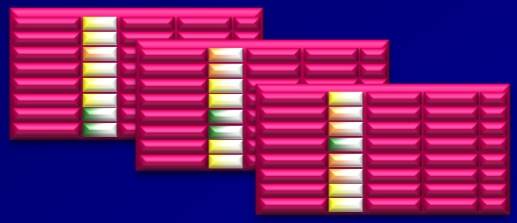
Hybrid Columnar Compression (11.2)

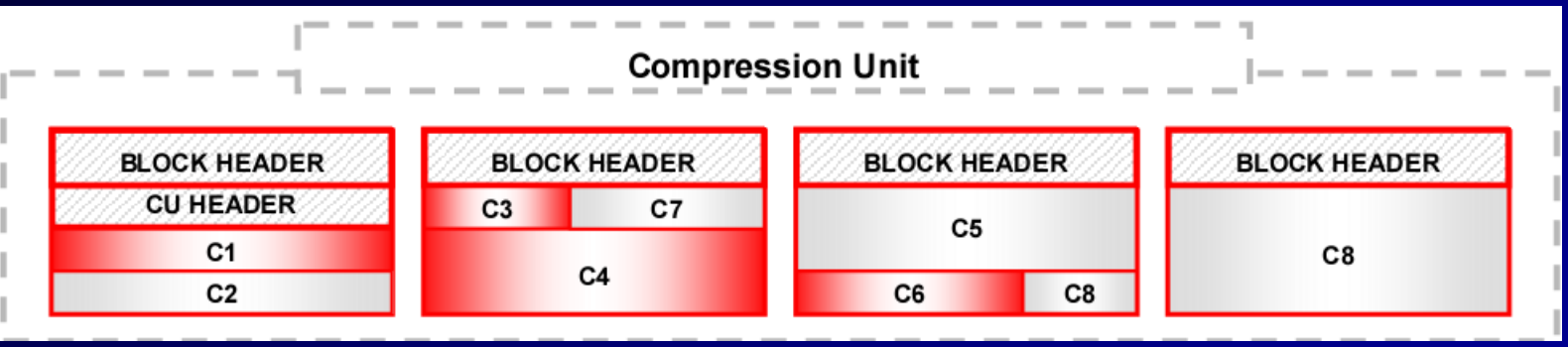
Compression (EHCC) – 4-10x & 30x is common

- What is it (a HYBRID of column & row storage)?
  - Data organized by column and compressed vs. row
  - Tables organized in Compression Units (CU)-1000 rows?
  - CU's span many blocks (32K)
  - **Good for data bulk loaded (not for OLTP – single block)**
- What's it for?
  - Query Data / DWHS (**NOT frequently Updated**)
- How much does it compress (old OLTP was 2-3x)?
  - 10x in a typical data warehouse compression; *(we got 4-11)*
  - 15x to 70x in archive compression (*cold* data); *(we got 32)*

Hybrid Columnar Compression

1. Column Data Compressed





(Warehouse)

2. Stored in Compression Units
(Better compression when column data stored together)

(Archive)

*** Thanks Oracle for these images*

Hybrid Columnar Compression

- Faster Operations: Query runs without decompression
  - Compressed/Processed in FLASH CACHE; lower I/O!
  - Compressed when sent over InfiniBand!
  - Cloned compressed!
  - Backed Up compressed!
  - Scans MUCH less (compressed) data
- Worth Noting:
  - Use standard table compression for OLTP
  - Single block lookup FASTER than other columnar

Hybrid Columnar Compression

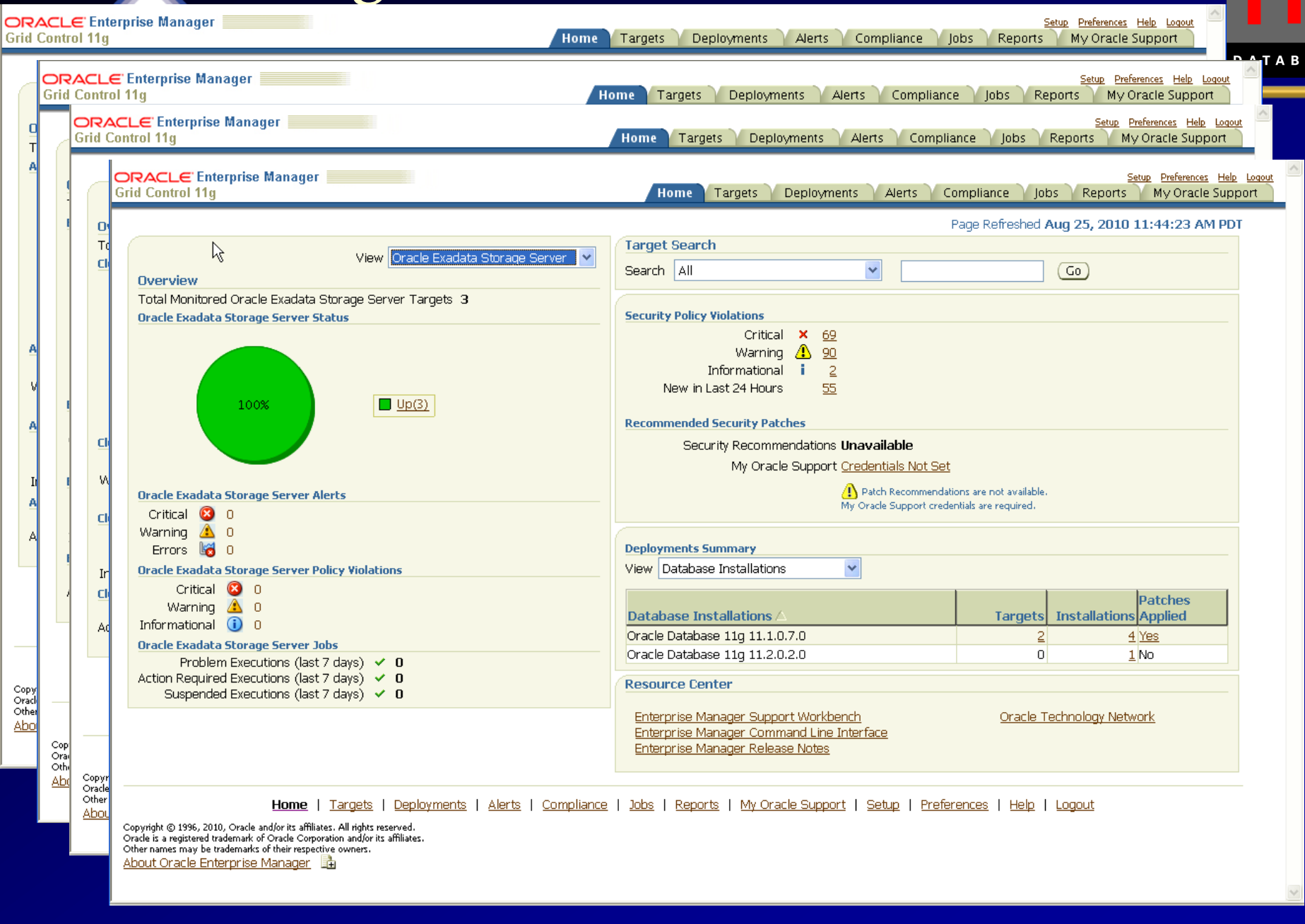
- Fully supported:
  - B-Tree Indexes
  - Bitmap Indexes
  - Text Index
  - Materialized Views
  - Partitioning
  - Parallel Query
  - Data Guard Physical Standby
  - *Logical Standby and Streams (FUTURE release)*
  - ***Smart Scans of HCC tables!***

Other Oracle Compression

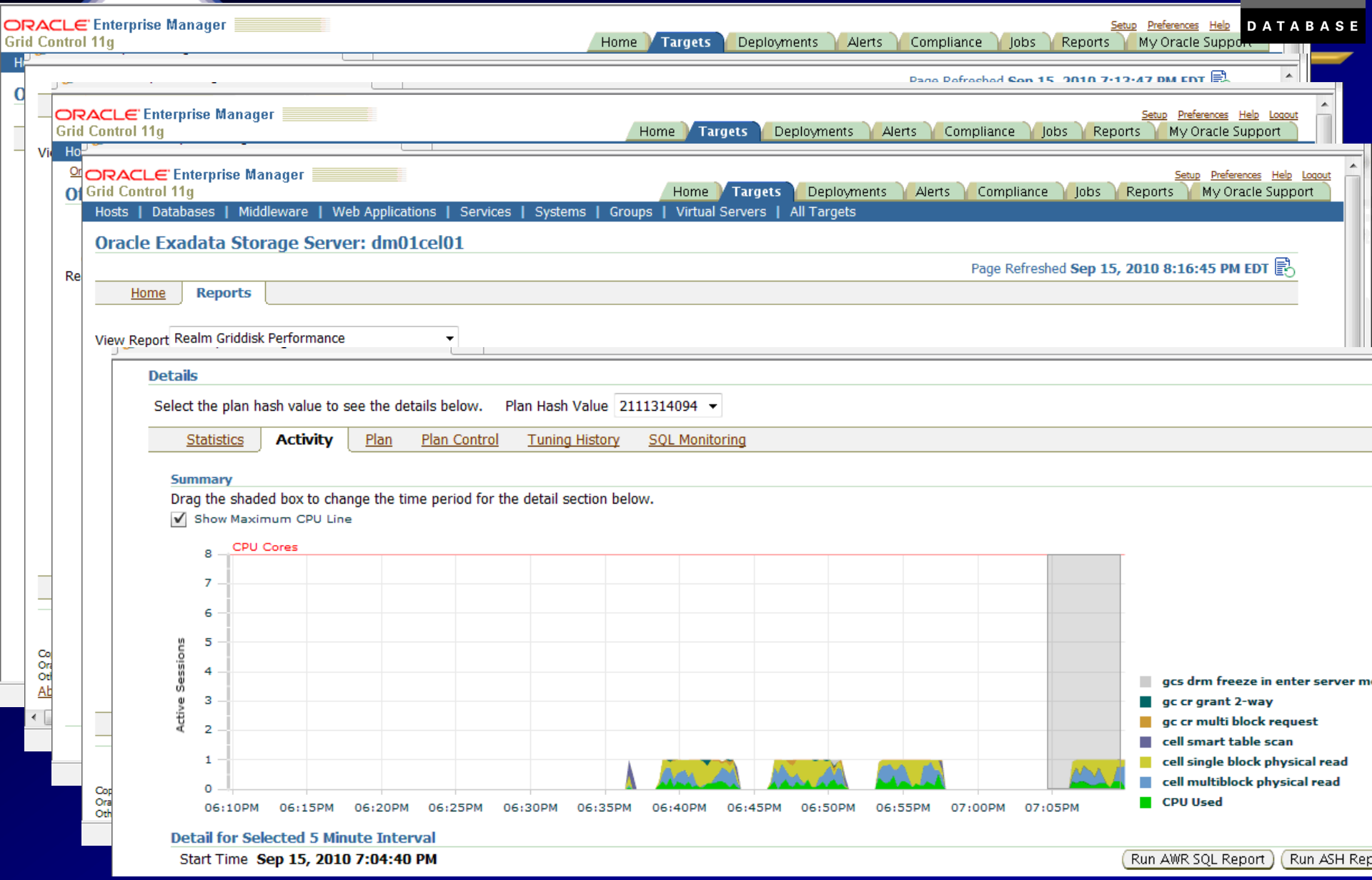
- Data Pump Compression
  - Compression = {ALL | DATA_ONLY | NONE}
- RMAN Backup Compression
  - Compression Level LOW/HIGH (New in 11.2)
- Secure File Compression
  - LOW/MEDIUM/HIGH (2-3x compression)
  - Deduplication & Encryption
- Normal OLTP Table Compression (since 9.2)
  - 11g now supports INSERT/UPDATE
  - FASTER Algorithm
- Data Guard Redo Transport Compression

Enterprise Manager & Grid Control for Exadata

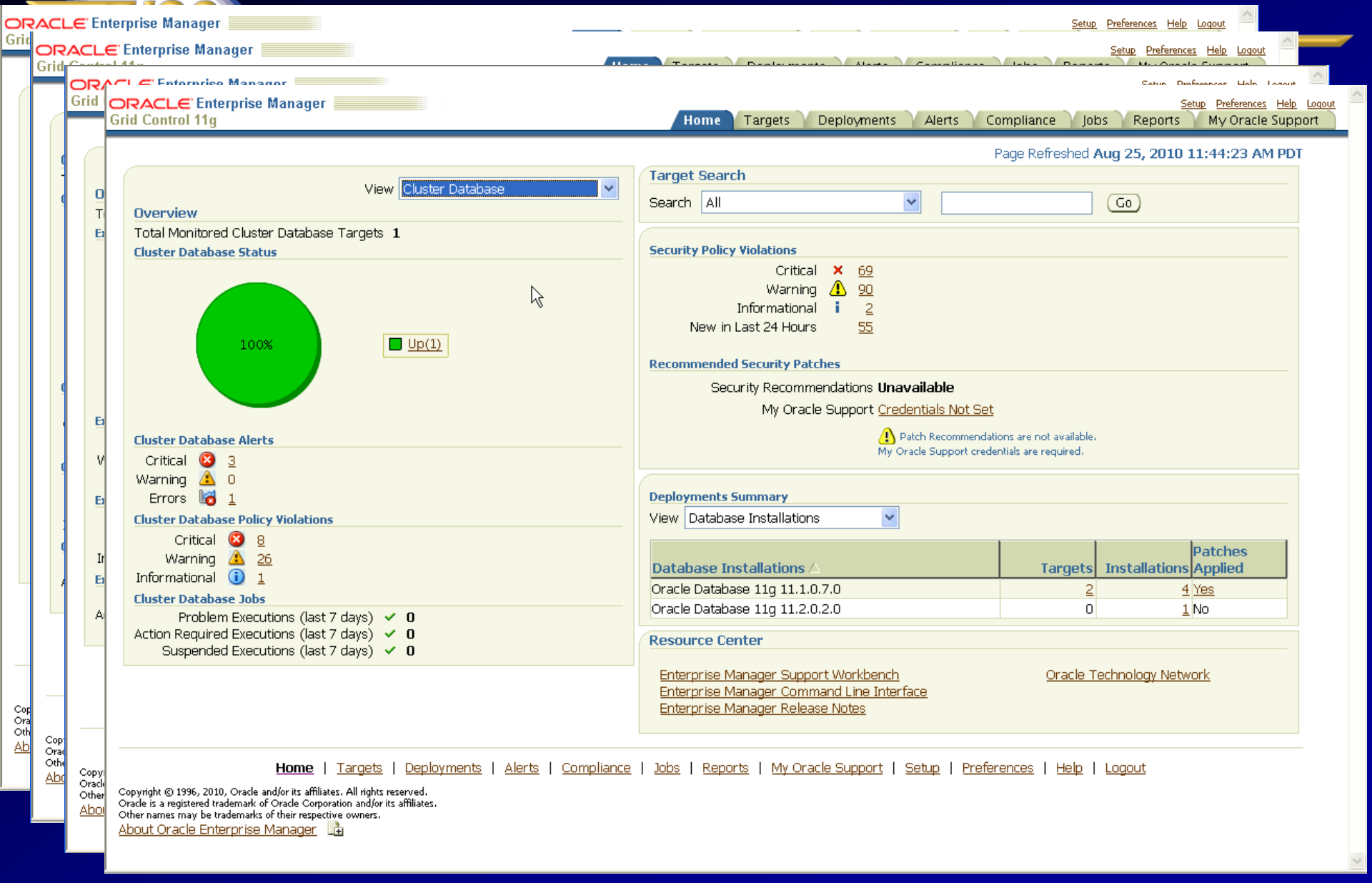
11gR2 Exadata



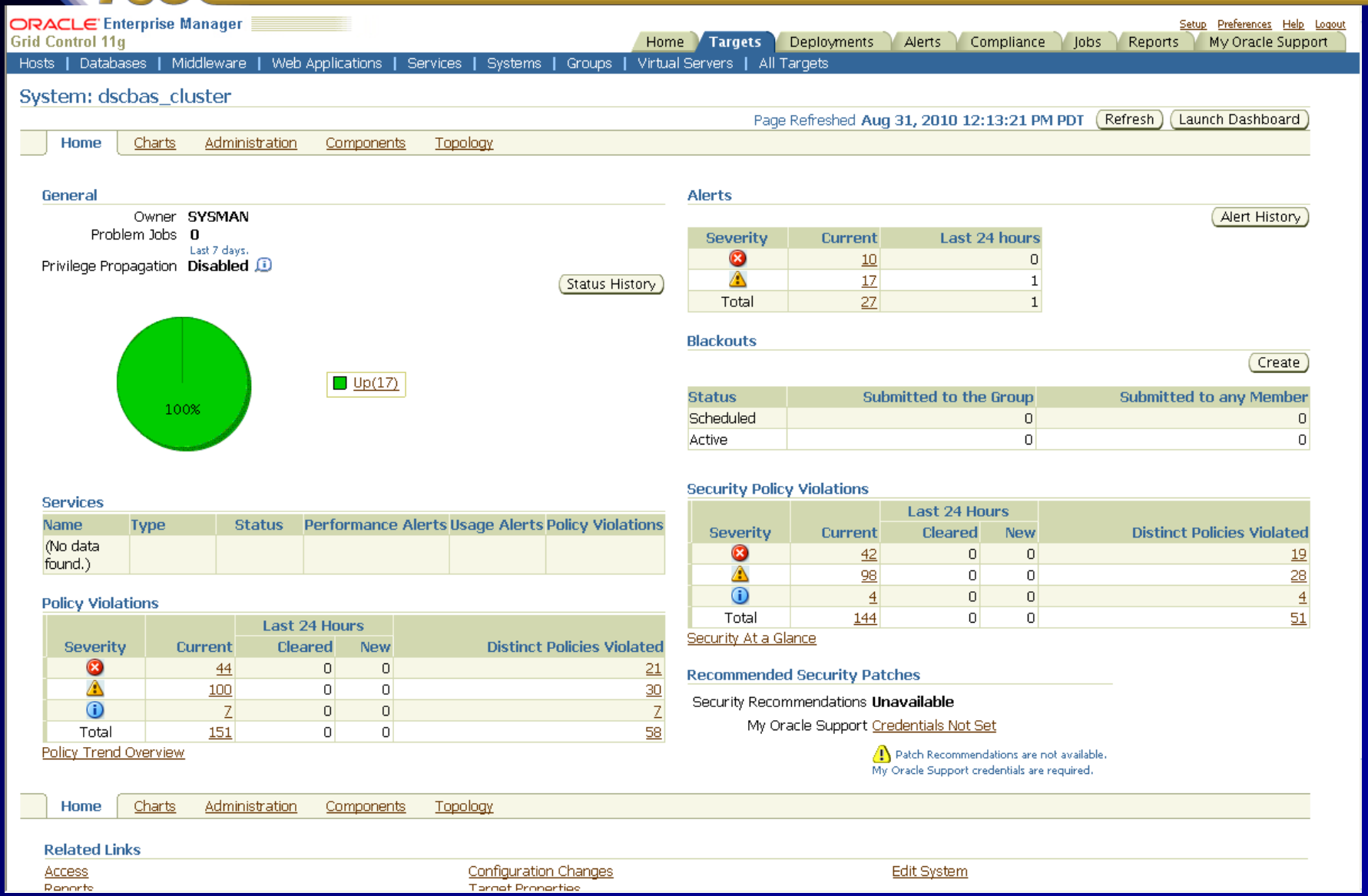
11gR2 Exadata



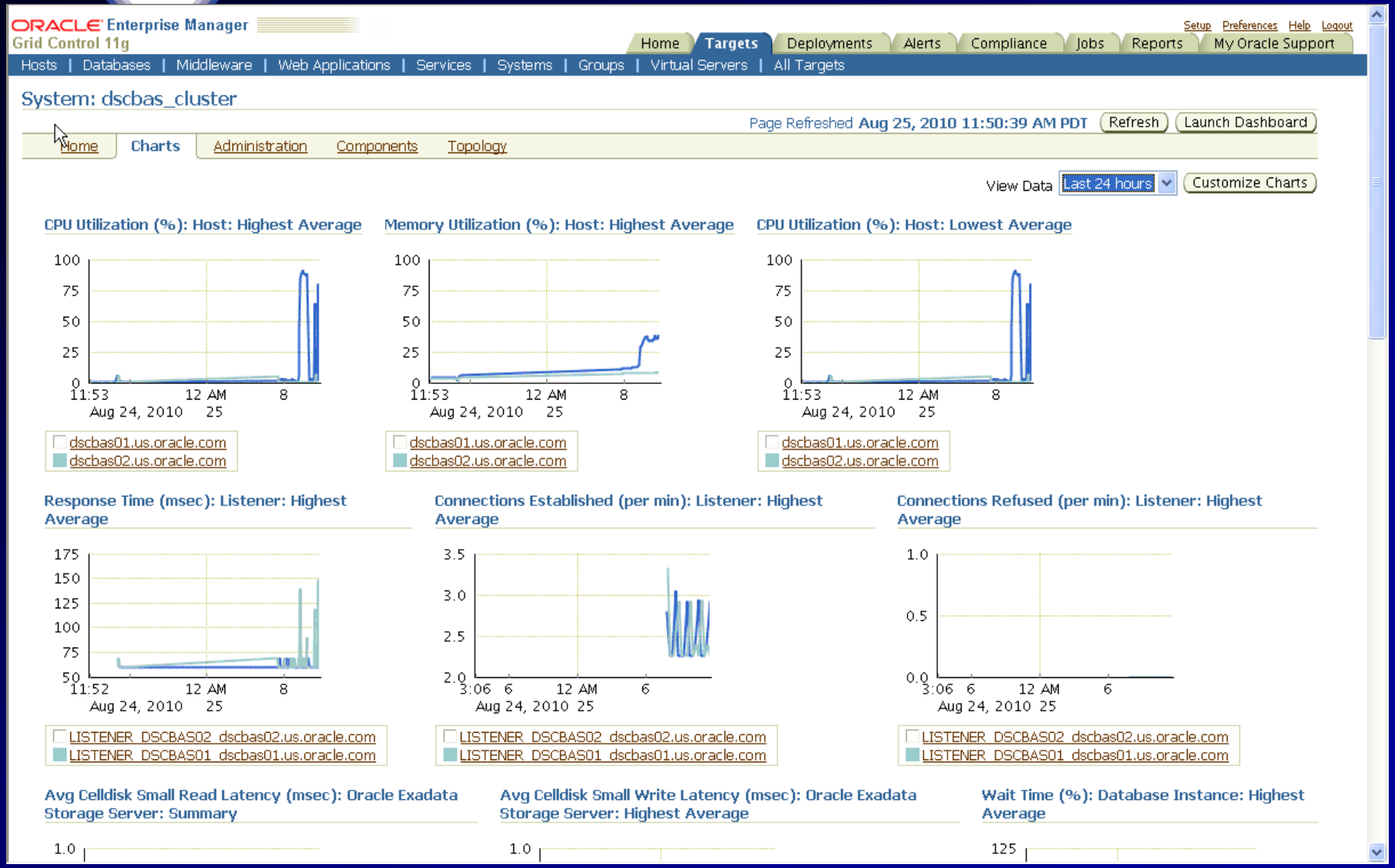
Monitor Targets
(Next slides – Some are coming soon)



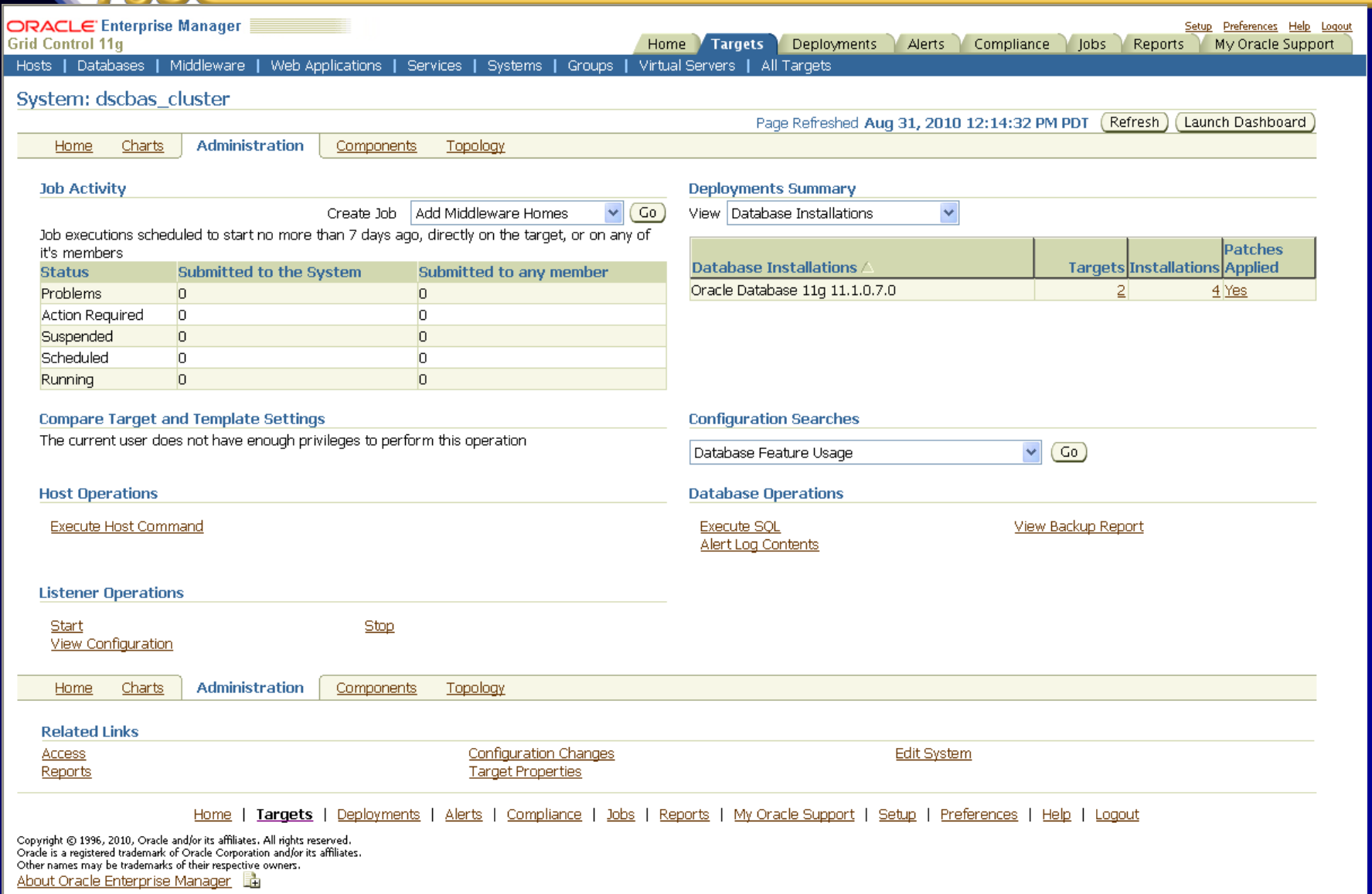
Cluster Home – Shows Alerts… etc.



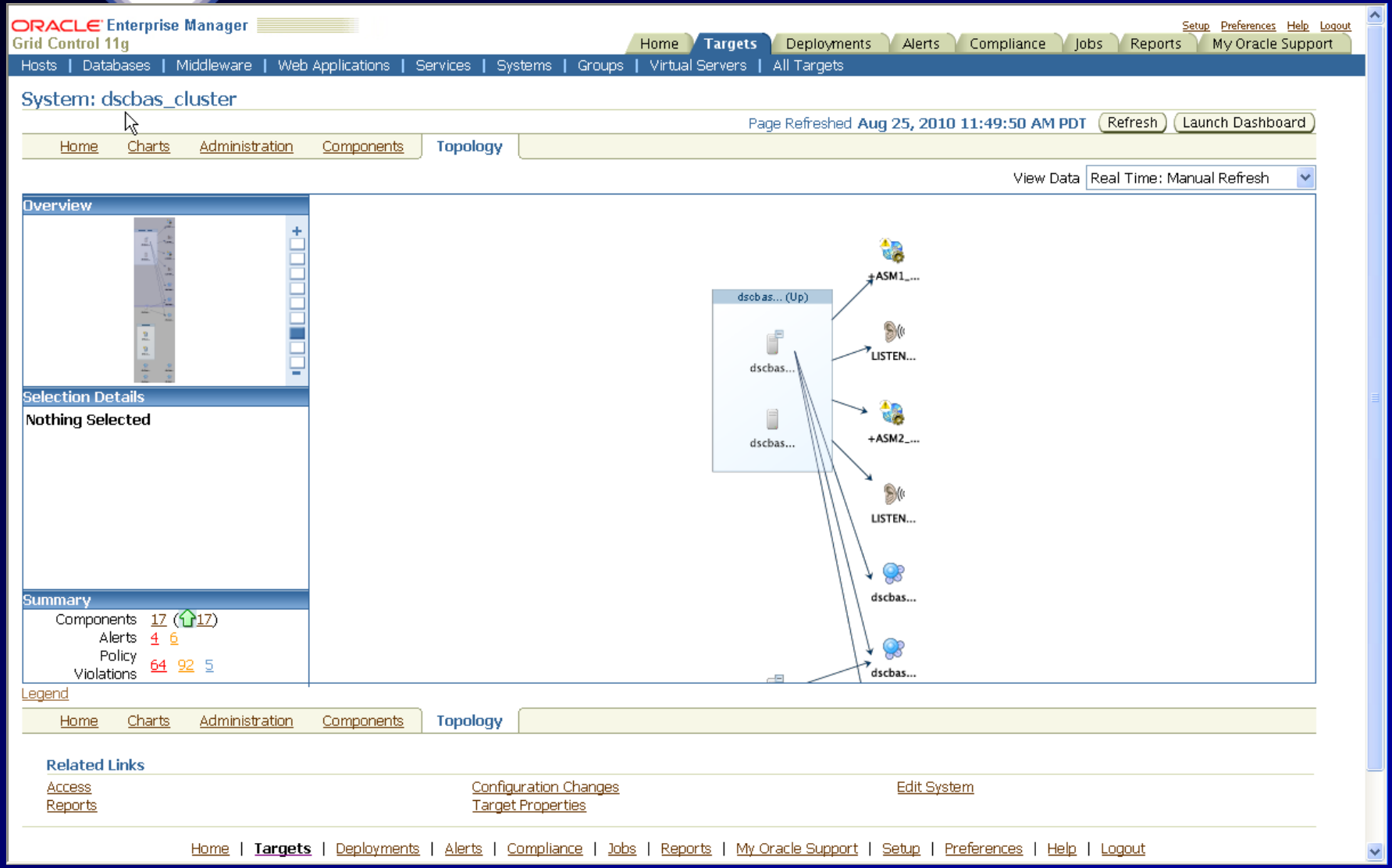
Cluster Charts – Shows Performance (showing 2 DB Servers on same graph)



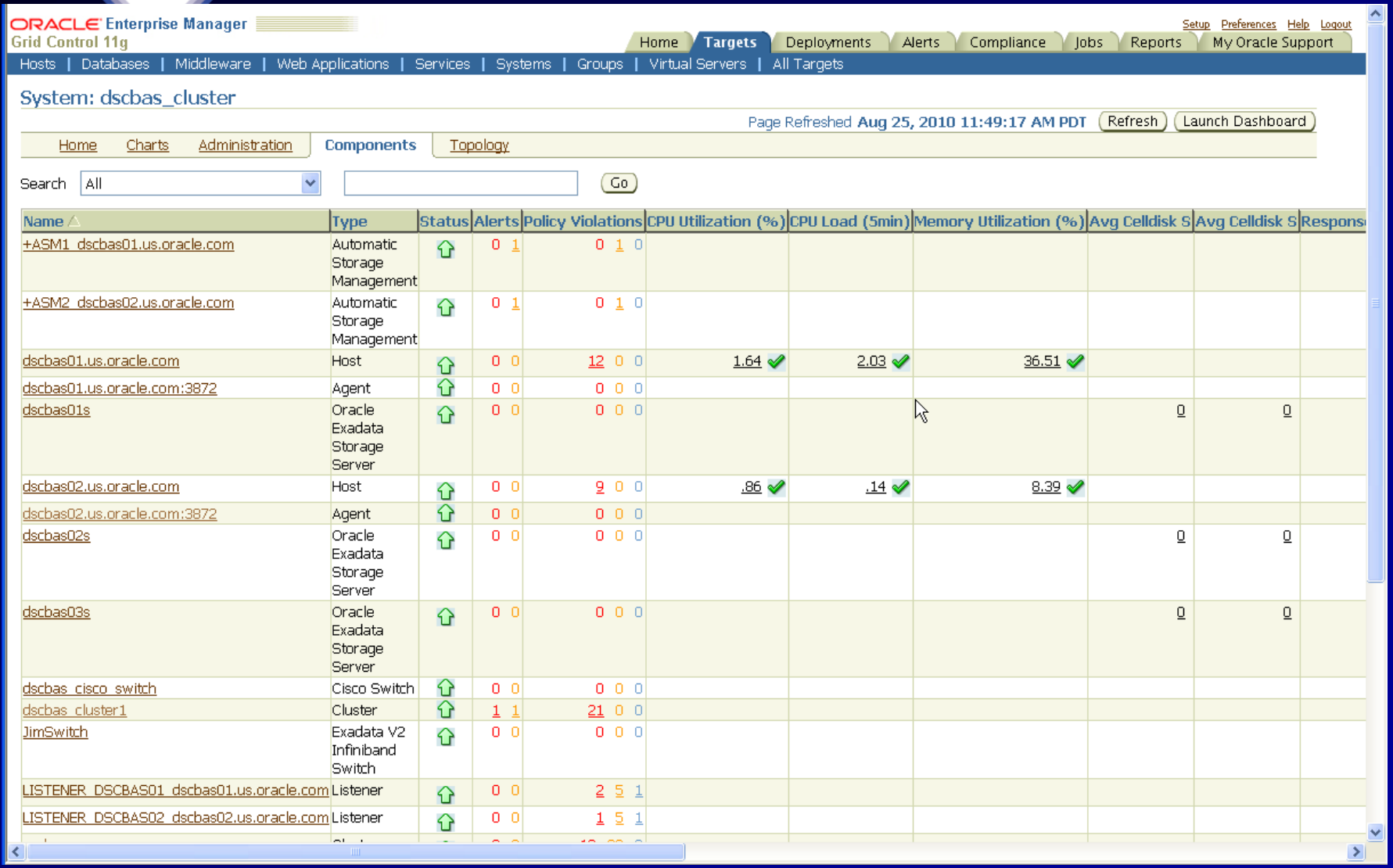
Cluster Admin – Shows Issues



Cluster Topology

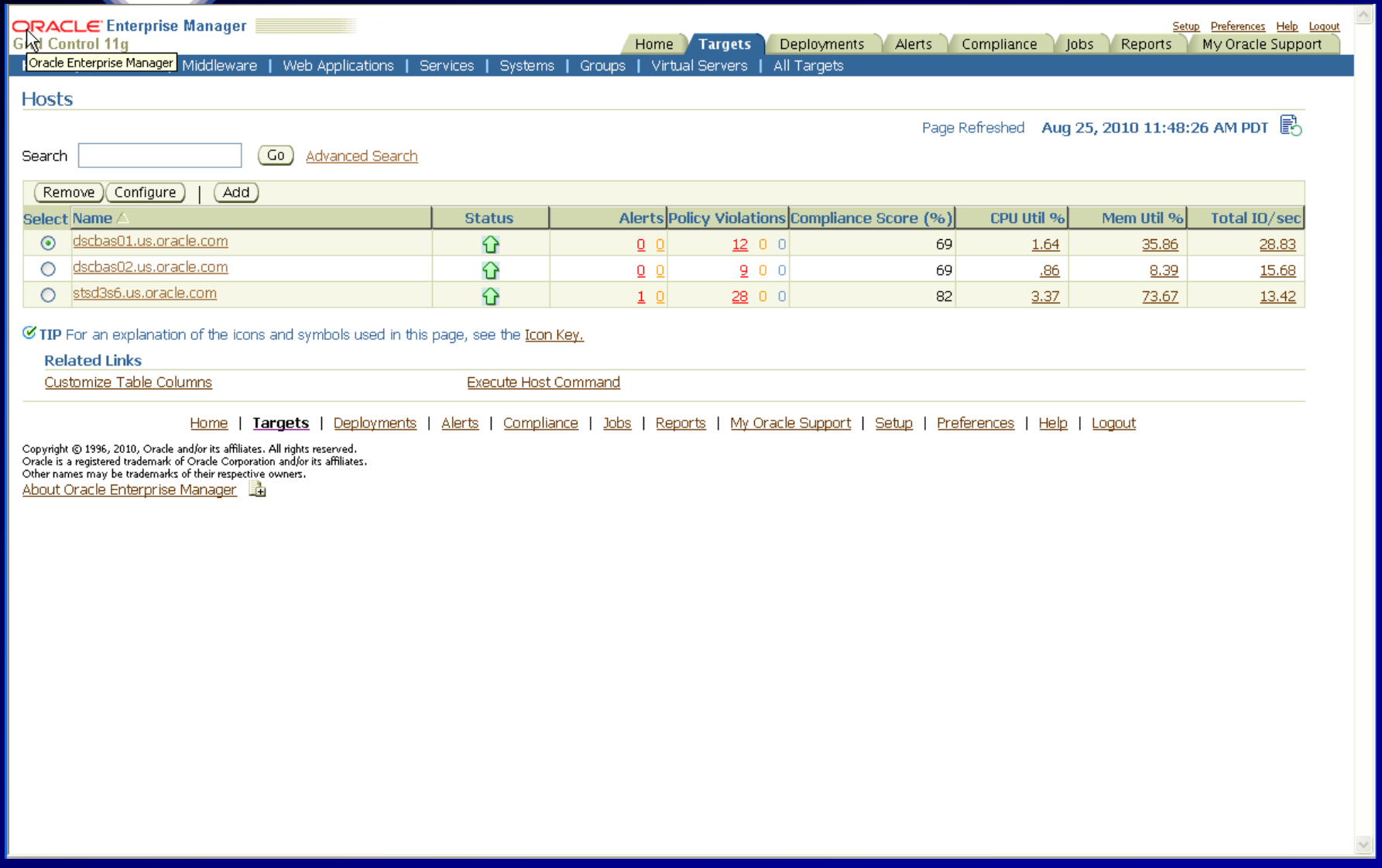


Targets – CLUSTER Components



Name	Type	Status	Alerts	Policy Violations	CPU Utilization (%)	CPU Load (5min)	Memory Utilization (%)	Avg Celldisk S	Avg Celldisk S	Respons
+ASM1_dscbas01.us.oracle.com	Automatic Storage Management		0 1	0 1 0						
+ASM2_dscbas02.us.oracle.com	Automatic Storage Management		0 1	0 1 0						
dscbas01.us.oracle.com	Host		0 0	12 0 0	1.64	2.03	36.51			
dscbas01.us.oracle.com:3872	Agent		0 0	0 0 0						
dscbas01s	Oracle Exadata Storage Server		0 0	0 0 0				0	0	
dscbas02.us.oracle.com	Host		0 0	9 0 0	.86	.14	8.39			
dscbas02.us.oracle.com:3872	Agent		0 0	0 0 0						
dscbas02s	Oracle Exadata Storage Server		0 0	0 0 0				0	0	
dscbas03s	Oracle Exadata Storage Server		0 0	0 0 0				0	0	
dscbas_cisco_switch	Cisco Switch		0 0	0 0 0						
dscbas_cluster1	Cluster		1 1	21 0 0						
JimSwitch	Exadata V2 Infiniband Switch		0 0	0 0 0						
LISTENER_DSCBAS01_dscbas01.us.oracle.com	Listener		0 0	2 5 1						
LISTENER_DSCBAS02_dscbas02.us.oracle.com	Listener		0 0	1 5 1						

Hosts


ORACLE Enterprise Manager
Grid Control 11g
Setup Preferences Help Logout
Home Targets Deployments Alerts Compliance Jobs Reports My Oracle Support
Oracle Enterprise Manager
Middleware | Web Applications | Services | Systems | Groups | Virtual Servers | All Targets
Hosts
Page Refreshed Aug 25, 2010 11:48:26 AM PDT
Search Go Advanced Search
Remove Configure | Add
Select Name Status Alerts Policy Violations Compliance Score (%) CPU Util % Mem Util % Total IO/sec
dscbas01.us.oracle.com 0 0 12 0 0 69 1.64 35.86 28.83
dscbas02.us.oracle.com 0 0 9 0 0 69 .86 8.39 15.68
stsd3s6.us.oracle.com 1 0 28 0 0 82 3.37 73.67 13.42
TIP For an explanation of the icons and symbols used in this page, see the Icon Key.
Related Links
Customize Table Columns
Execute Host Command
Home | Targets | Deployments | Alerts | Compliance | Jobs | Reports | My Oracle Support | Setup | Preferences | Help | Logout
Copyright © 1996, 2010, Oracle and/or its affiliates. All rights reserved.
Oracle is a registered trademark of Oracle Corporation and/or its affiliates.
Other names may be trademarks of their respective owners.
About Oracle Enterprise Manager

ASM

ORACLE Enterprise Manager
Grid Control 11g

Setup | Preferences | Help | Logout

Home | Targets | Deployments | Alerts | Compliance | Jobs | Reports | My Oracle Support

Hosts | Databases | Middleware | Web Applications | Services | Systems | Groups | Virtual Servers | All Targets

Host: dscbas01.us.oracle.com >

Automatic Storage Management: +ASM1_dscbas01.us.oracle.com

Home | Performance | Disk Groups | Configuration | Users

Data Retrieved Aug 25, 2010 11:54:20 AM PDT Refresh

General

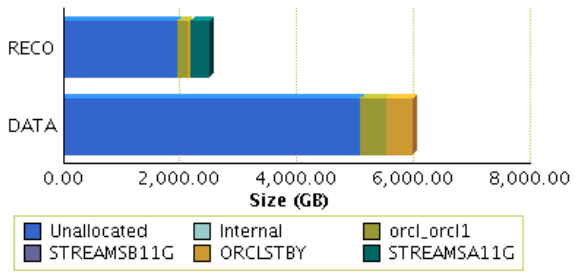
Startup/Shutdown

Current Status	Up
Up Since	Aug 25, 2010 7:12:21 AM PDT
Availability (%)	61.84 (Last 24 hours)
Instance Name	+ASM1
Version	11.1.0.7.0
Host	dscbas01.us.oracle.com
Oracle Home	/u01/app/oracle/product/11.1.0/asm_1

Diagnostic Summary

Alert Log: No ORA- errors
Active Incidents: 0

Disk Group Usage (GB)



Serviced Databases

Name	Disk Groups	Failure Groups	Allocated Space (GB)	Availability	Alerts
orcl_orcl1	DATA, RECO	6 (0 down)	632.06	↑	1 0
STREAMSB11G	RECO	3 (0 down)	0.15		Not Monitored
ORCLSTBY	DATA, RECO	6 (0 down)	507.56		Not Monitored
STREAMSA11G	RECO	3 (0 down)	328.84		Not Monitored

Alerts

Severity	Category	Name	Message	Alert Triggered
⚠	Failure Group Imbalance Status	Disk Size Imbalance (%)	Disk Group RECO has failure groups with disks of different sizes which may lead to suboptimal space usage. Changing the configuration may alleviate this problem.	May 25, 2010 12:16:18 PM

Host Alerts

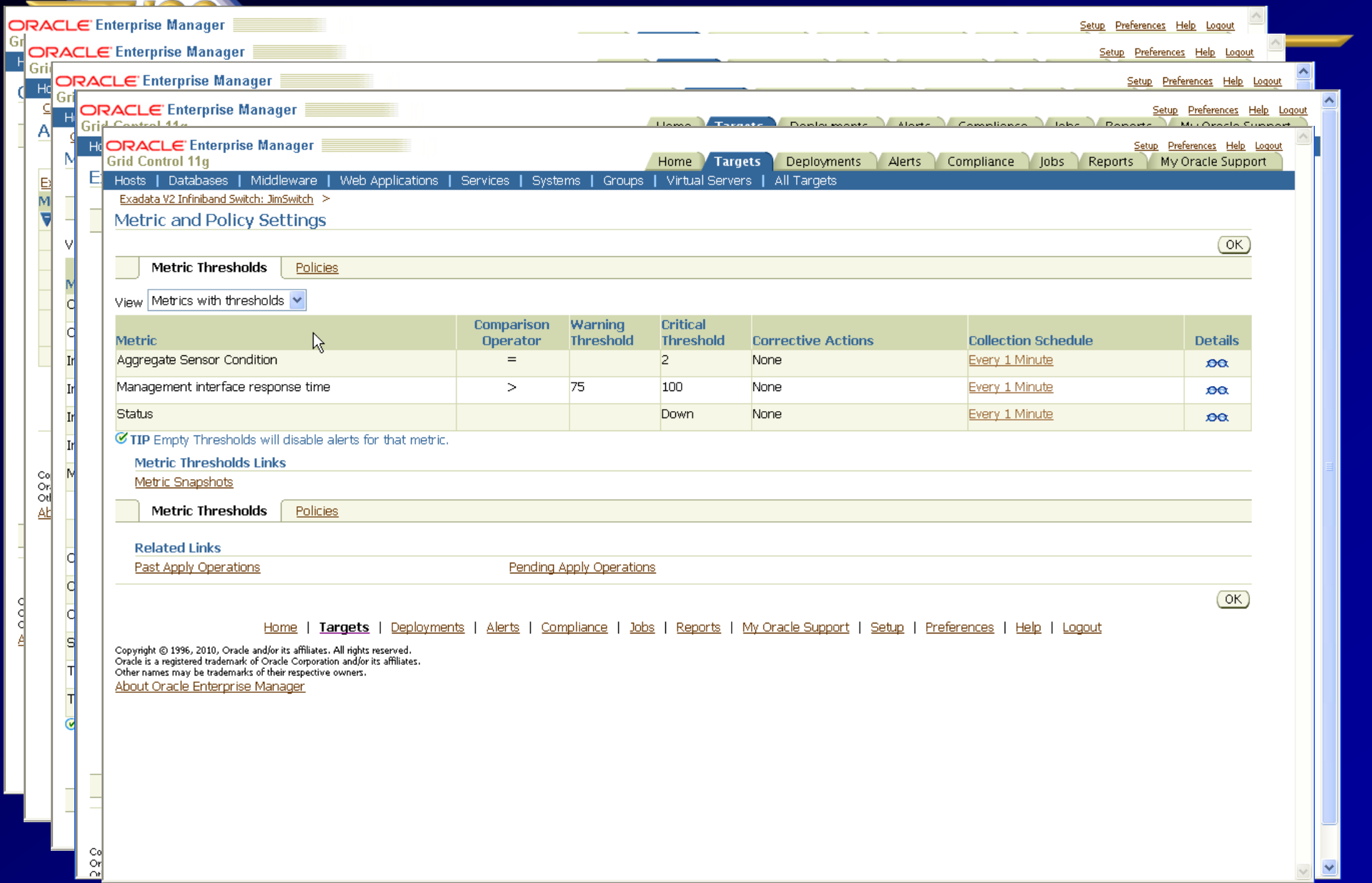
Severity	Category	Name	Message	Alert Triggered
(No alerts)				

Policy Violations

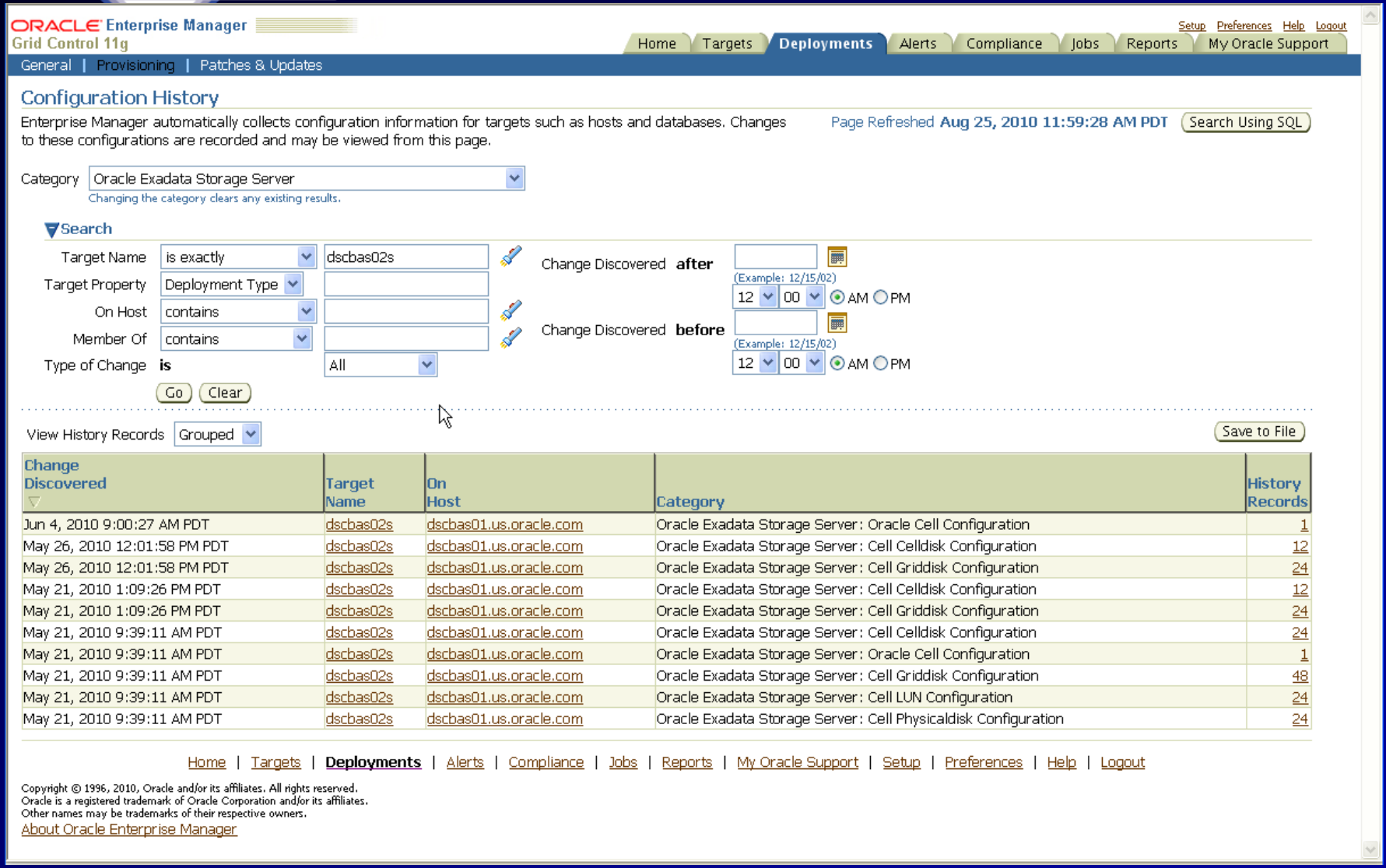
Storage Servers 1-3 (Exadata Plug-In)

Database Instance Monitoring

Cisco Switch & InfiniBand Switch



Configuration History

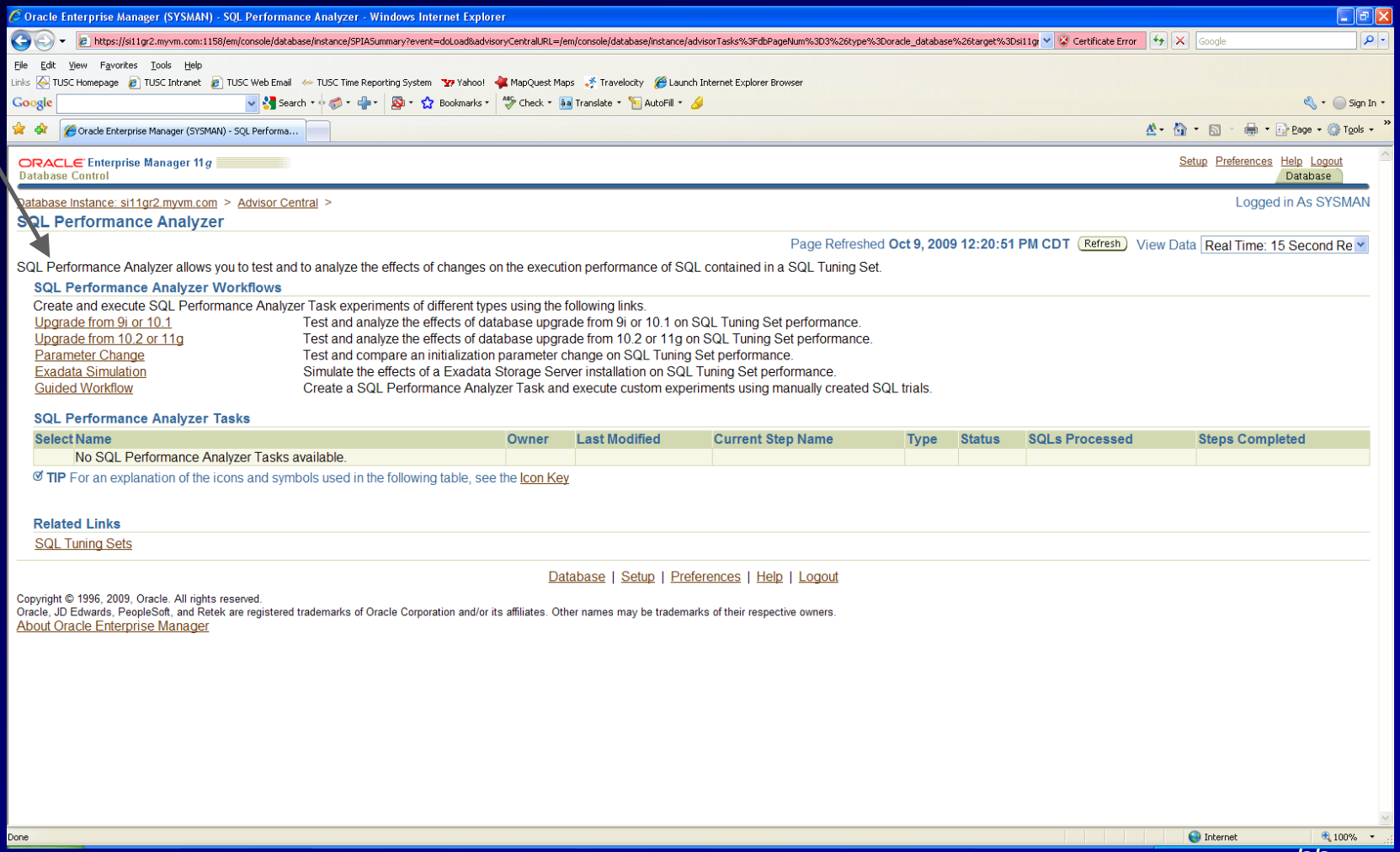


Enterprise Manager
Exadata Simulation

“11g R1/R2 Best Features” (more on this)

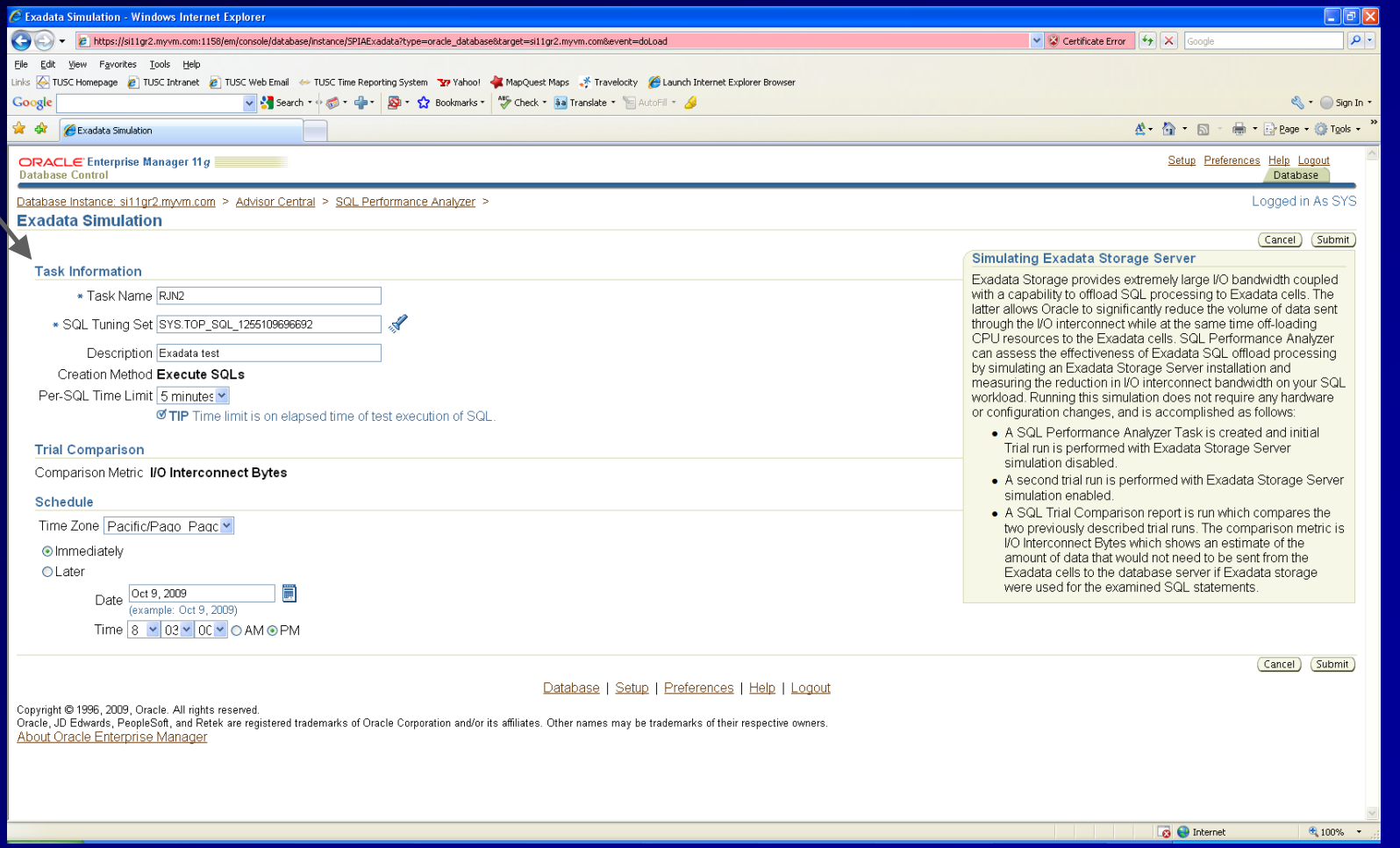
SQL Performance Analyzer 11gR2 - Options

Upgrade Options



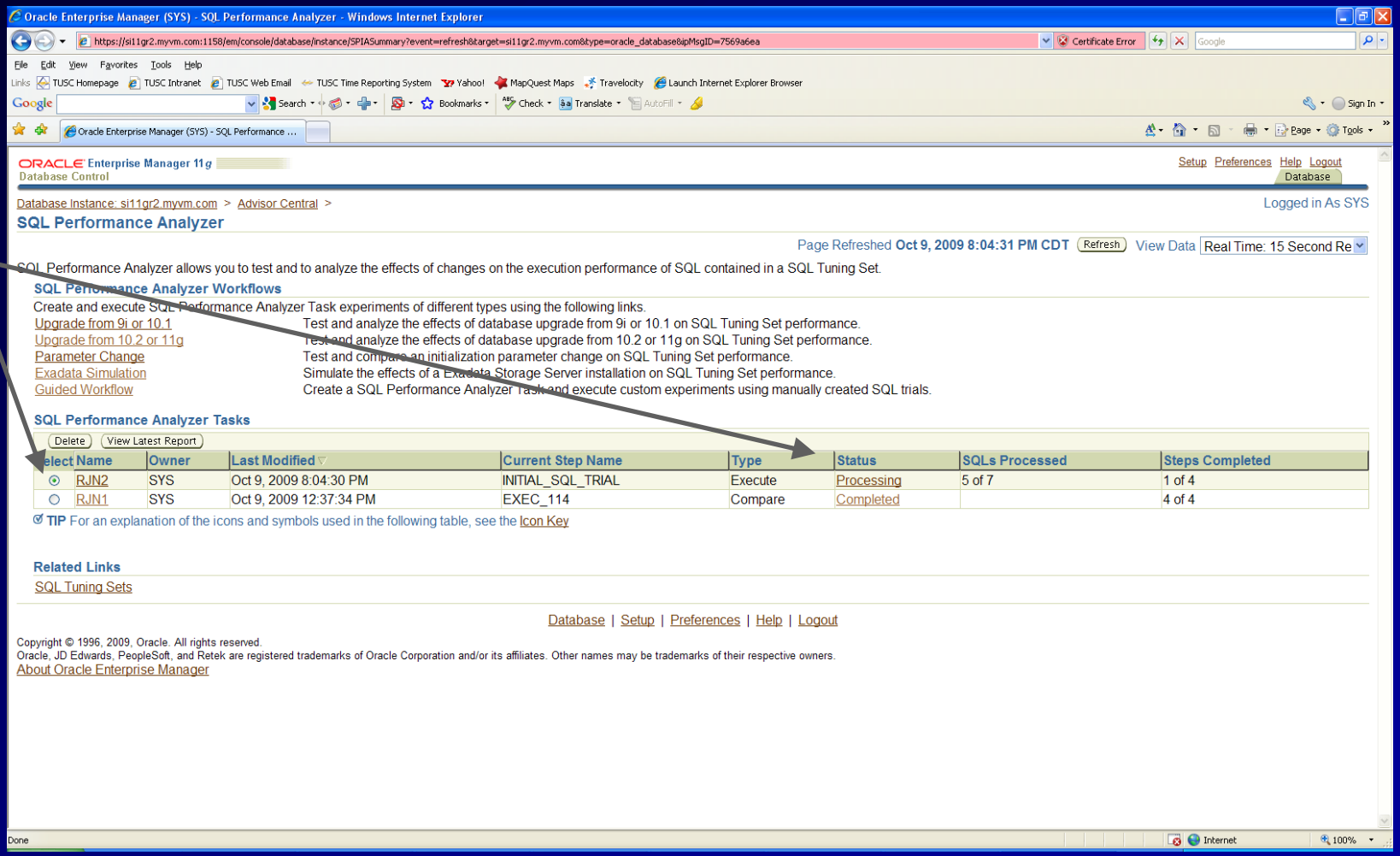
SQL Performance Analyzer 11gR2 – Exadata Simulation

Test a Tuning Set that I've used in the past



SQL Performance Analyzer 11gR2 – Exadata Simulation

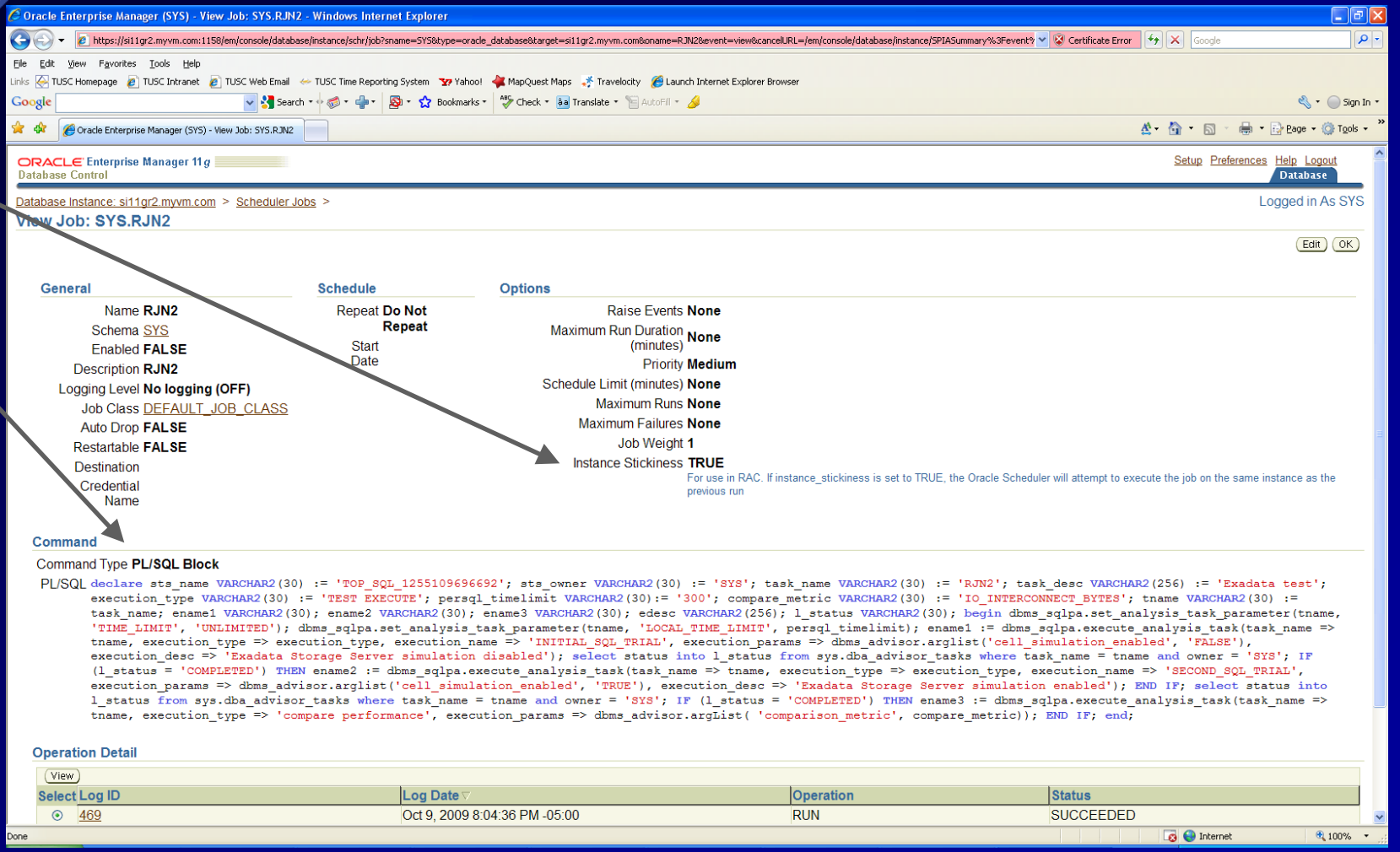
Job is running



SQL Performance Analyzer 11gR2 – Exadata Simulation

New!

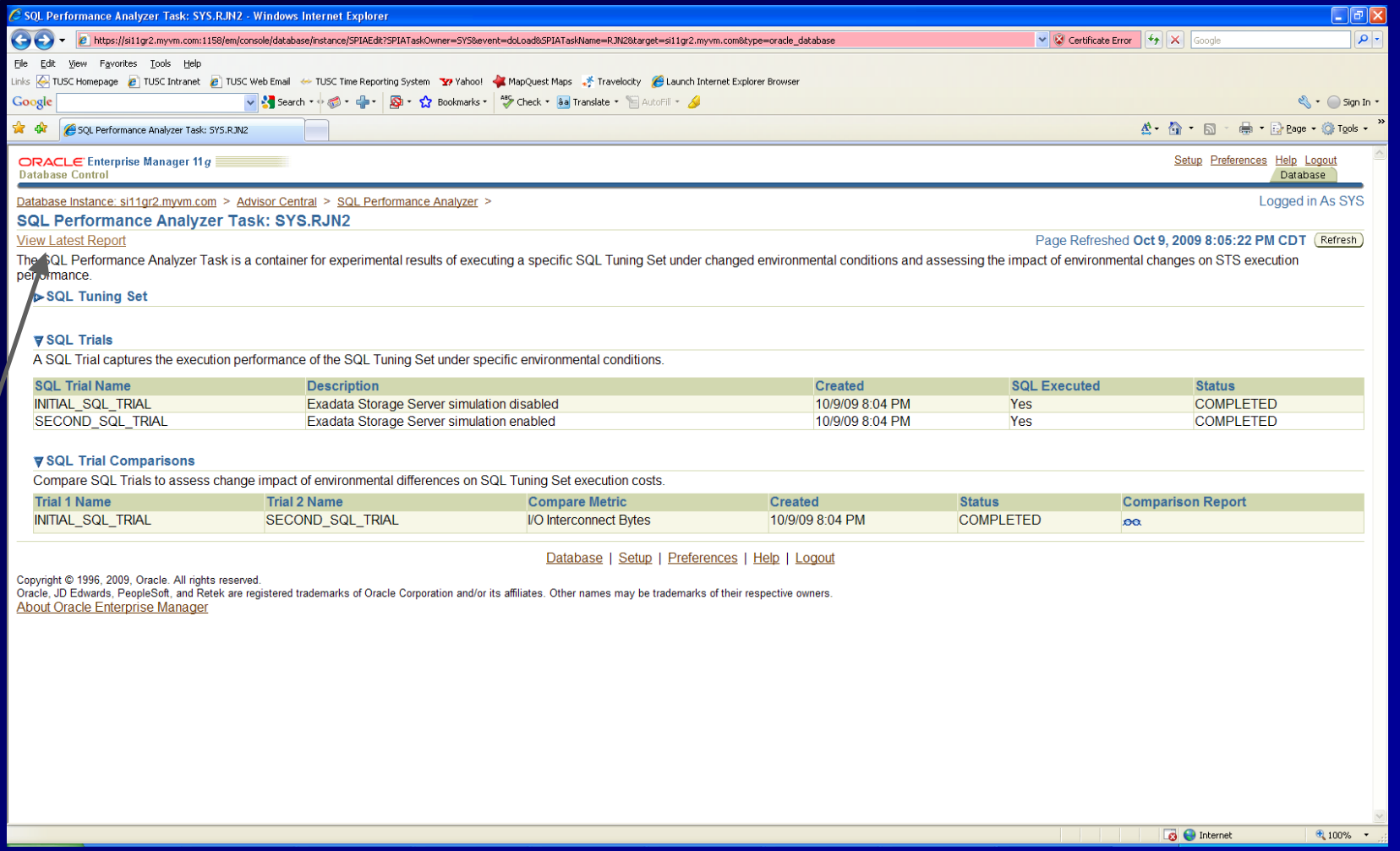
View PL/SQL



SQL Performance Analyzer 11gR2 – Exadata Simulation

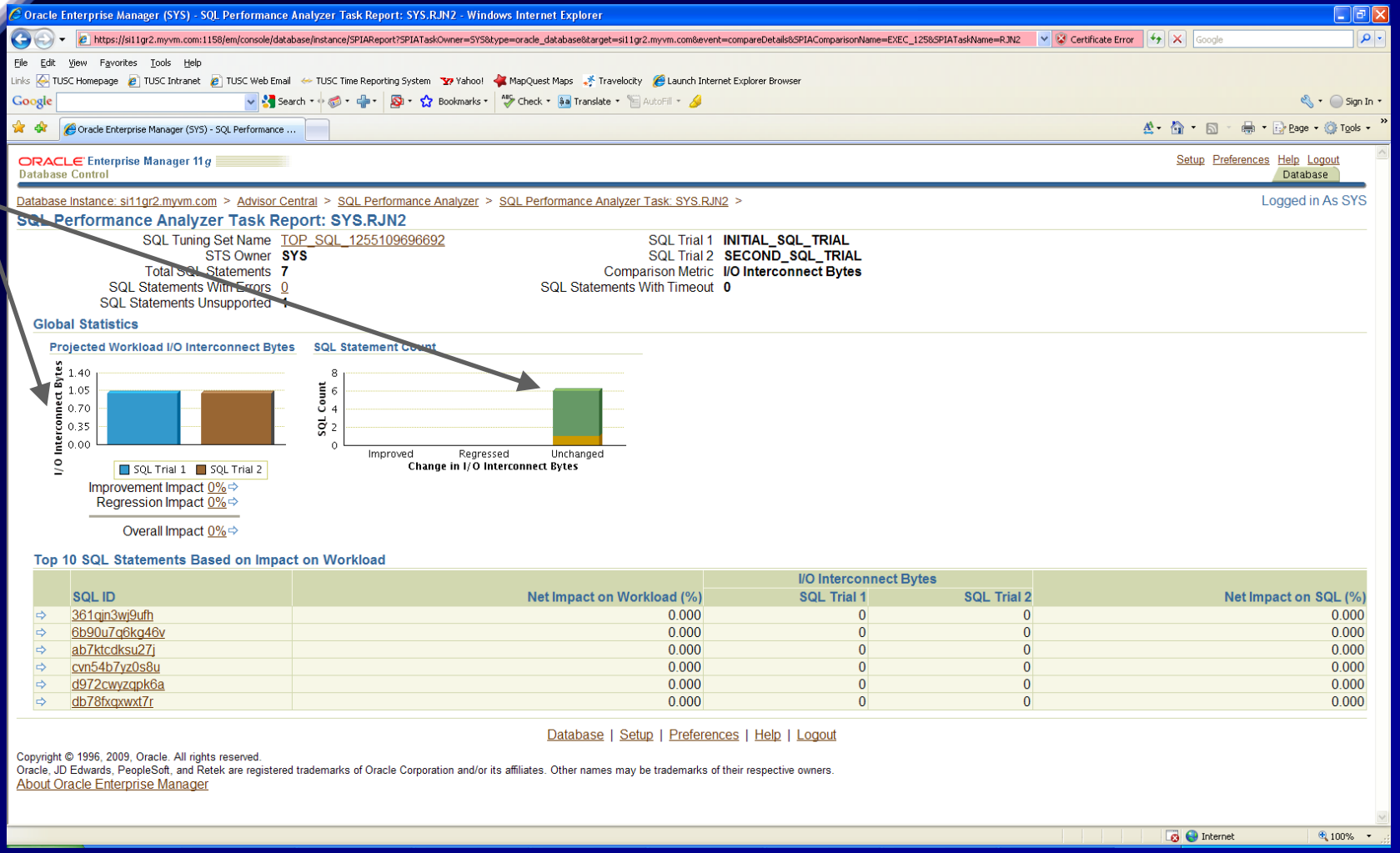
Click on Job after complete

View Report



SQL Performance Analyzer 11gR2 – Exadata Simulation

Simple Job so no benefit



Resource Management (IORM) (FYI Only)

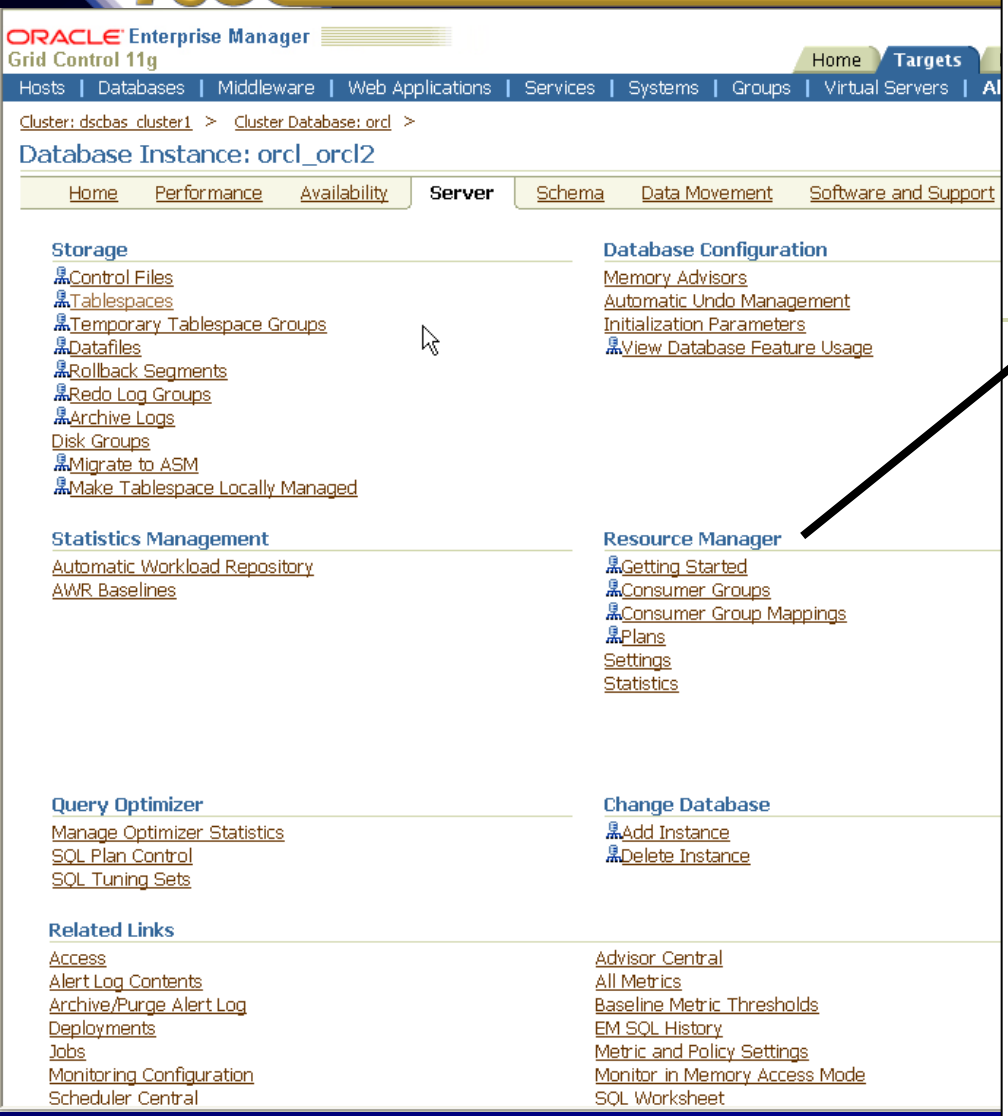
IORM - I/O Resource Management

- Set I/O resources for different instance
  - Instance A = 50%
  - Instance B = 30%
  - Instance C = 20%
- Further set I/O based on users and tasks
  - Instance A Interactive = 50%
  - Instance A Reporting = 25%
  - Instance A Batch = 15%
  - Instance A ETL = 15%
- Best Solution for MIXED workloads & many instances

DBRM – Database Resource Manager

- Enhanced for Exadata
- Allows management of inter and intra DB I/O
- Inter-DB – Managed via IORM & Exadata storage software
- Intra-DB - Managed via Consumer Group
- CPU
- Undo
- DOP (Degree of Parallelism)
- Active Sessions

Grid Control - Resource Manager



Resource Manager

- Getting Started
- Consumer Groups
- Consumer Group Mappings
- Plans

Settings

Statistics

Change Database

Security – FYI Only

Oracle Database Security*

Built over MANY years...

Oracle Audit Vault

Oracle Database Vault

DB Security Evaluation #19

Transparent Data Encryption

EM Configuration Scanning

Fine Grained Auditing (9i)

Secure application roles

Client Identifier / Identity propagation

Oracle Label Security (2000)

Proxy authentication

Enterprise User Security

Global roles

Virtual Private Database (8i)

Database Encryption API

Strong authentication (PKI, Kerberos, RADIUS)

Native Network Encryption (Oracle7)

Database Auditing

1977 Government customer

2007+

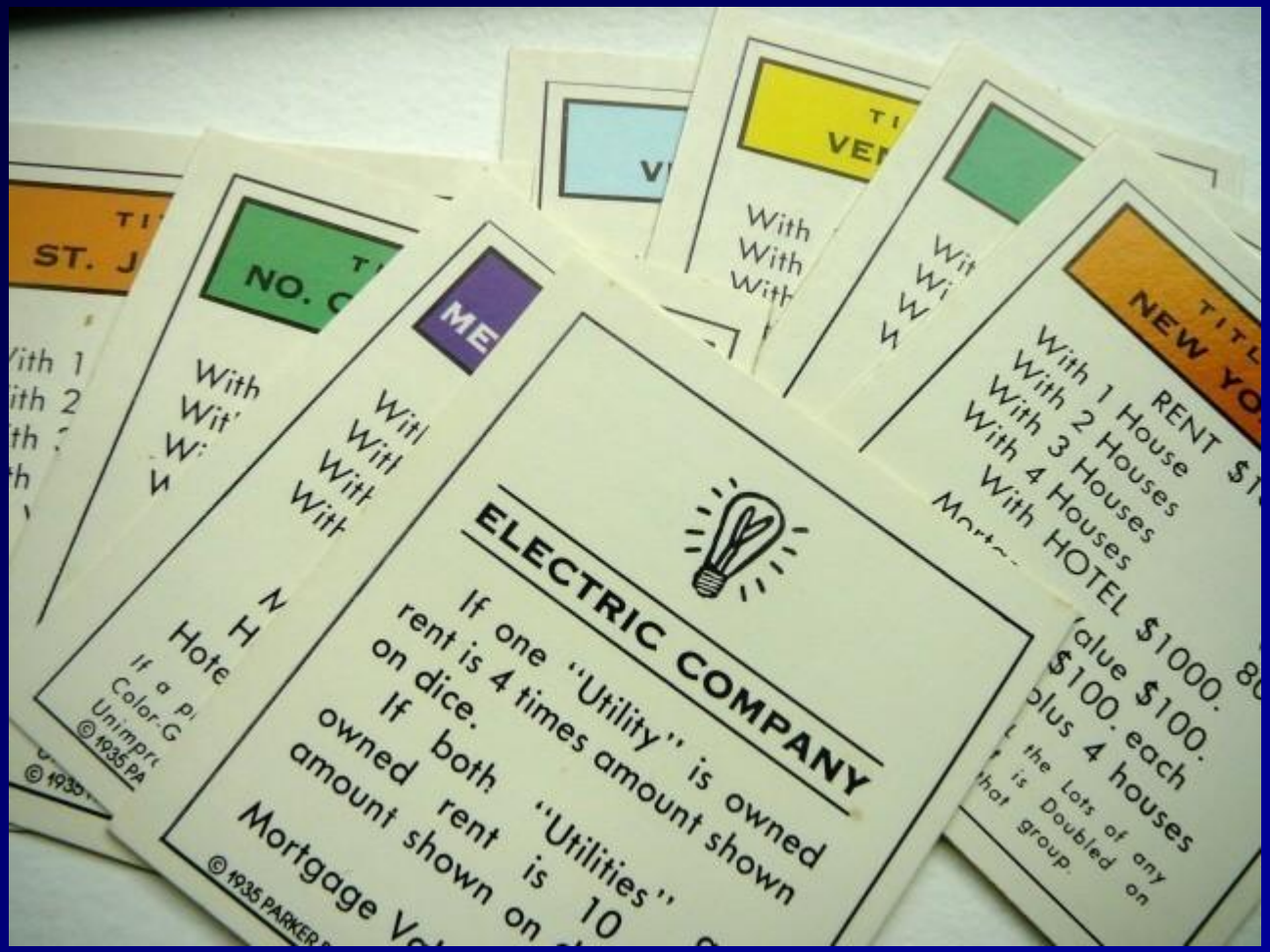
*Oracle Slide – Thanks!

Security

- Audit Vault
- Total Recall / Flashback
- Database Vault
- Label Security
- Advanced Security
- Secure encrypted backup (also available: incremental backup with Change Tracking File – much faster)
- Data Masking
- Data Guard
- Failure Groups – (automatic-for storage cell failure)

Utilities - FYI Only...

(Screen Shots - Oracle Learning Library)



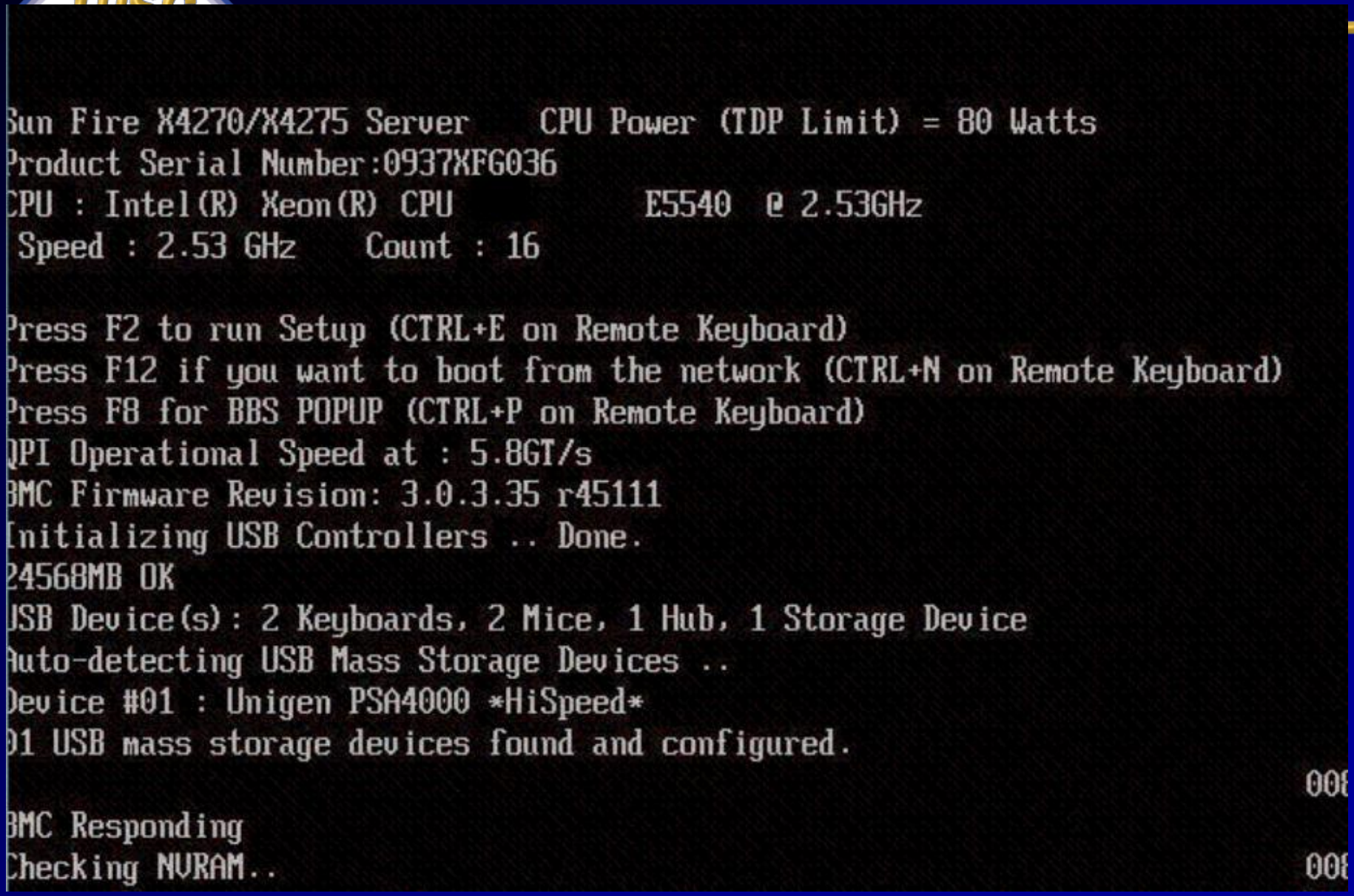
Utilities You'll need to Use - FYI

- **CELLCLI** – Cell Command Line Interface (CLI)
- **DCLI** – **Run the same command on multiple cells** at the same time

  From Oracle:

  Oracle Exadata Storage Server includes the DCLI utility on each cell. You can use the DCLI utility to execute commands or scripts in parallel across a defined set of cells. The DCLI tool simplifies any operations that must be run across a subset or all cells. Configuration of SSH user equivalency across all cells is an important prerequisite for optimizing the use of DCLI commands. DCLI provides the -k option to automate the distribution of SSH private keys into the AUTHORIZED_KEYS file.
- **ADRCI** – Automatic Diagnostic Repository Command line Interface; Quickly get diagnostics reports to send to Oracle

Turn Exadata on...



Checks continue...

```
Starting console mouse services:                           [  OK  ]
Starting crond:                                            [  OK  ]
Starting xfs:                                              [  OK  ]
Starting anacron:                                          [  OK  ]
Starting atd:                                              [  OK  ]
Starting exachkcfg:                                        [FAILED]
Starting HAL daemon:                                       [  OK  ]
Starting celld:
Starting the RS, CELLSRV, and MS services...
Getting the state of RS services...
 running
Starting CELLSRV services...
The STARTUP of  CELLSRV services was not successful. Error: Start Failed
Starting MS services...
The STARTUP of MS services was successful.
                                                           [  OK  ]

Logging started to /var/log/cellos/validations.log
Run validation biosbootorder - PASSED
Run validation misceachboot - PASSED
Run validation createcell - RDS/IB: connected to 192.168.0.111 version 3.
RDS/IB: connected to 192.168.0.111 version 3.1
```

CELLCLI – Commands

- CellCLI> list cell detail
- CellCLI> list lun
- CellCLI> list physicaldisk
- CellCLI> list flashcache detail
- CellCLI> list celldisk
- CellCLI> calibrate force
- CellCLI> create celldisk all
- CellCLI> create griddisk all harddisk prefix='data' ,size=100g
- CellCLI> create griddisk all harddisk prefix='data' ,size=100g
- CellCLI> list griddisk
- CellCLI> list griddisk attributes name,size

Run CELLCLI – Cell Detail

```
[root@sgsas1 ~]# cellcli
CellCLI: Release 11.2.1.2.0 - Production on Fri Nov 06 09:31:15 PST 2009

Copyright (c) 2007, 2009, Oracle.  All rights reserved.
Cell Efficiency Ratio: 1

CellCLI> list cell detail
         name:                   sgsas1
         bmcType:                IPMI
         cpuCount:               16
         fanCount:               12/12
         fanStatus:              normal
         id:                     f59022f8-969d-4c20-bed0-ed54ba0d2079
         interconnectCount:      3
         interconnect1:          bond0
         iormBoost:              0.0
         ipaddress1:             192.168.214.193/22
         makeModel:              SUN MICROSYSTEMS SUN FIRE X4275 SERVER SAS
         metricHistoryDays:      7
         offloadEfficiency:      1.0
         powerCount:             2/2
         powerStatus:            normal
         status:                 online
         temperatureReading:     29.0
         temperatureStatus:      normal
         upTime:                 7 days, 9:32
         cellsrvStatus:          running
         msStatus:               running
         rsStatus:               running

CellCLI>
```

One Storage Server – 12 Disks + Flash

12 Physical Disks

16 Flash Modules
4 – flash cards
4 – flash modules ea. card

```
CellCLI> list lun
         0_0     0_0     normal
         0_1     0_1     normal
         0_2     0_2     normal
         0_3     0_3     normal
         0_4     0_4     normal
         0_5     0_5     normal
         0_6     0_6     normal
         0_7     0_7     normal
         0_8     0_8     normal
         0_9     0_9     normal
         0_10    0_10    normal
         0_11    0_11    normal
         1_0     1_0     normal
         1_1     1_1     normal
         1_2     1_2     normal
         1_3     1_3     normal
         2_0     2_0     normal
         2_1     2_1     normal
         2_2     2_2     normal
         2_3     2_3     normal
         4_0     4_0     normal
         4_1     4_1     normal
         4_2     4_2     normal
         4_3     4_3     normal
         5_0     5_0     normal
         5_1     5_1     normal
         5_2     5_2     normal
         5_3     5_3     normal

CellCLI>
```

CALIBRATE – check performance

Check Performance

```
CellCLI> calibrate force
Calibration will take a few minutes...
Aggregate random read throughput across all hard disk luns: 1597 MBPS
Aggregate random read throughput across all flash disk luns: 4212.28 MBPS
Aggregate random read IOs per second (IOPS) across all hard disk luns: 4930
Aggregate random read IOs per second (IOPS) across all flash disk luns: 149690
Controller read throughput: 1683.11 MBPS
Calibrating hard disks (read only) ...
Lun 0_0  on drive [20:0      ] random read throughput: 152.93 MBPS, and 421 IOPS
Lun 0_1  on drive [20:1      ] random read throughput: 156.59 MBPS, and 410 IOPS
Lun 0_10 on drive [20:10     ] random read throughput: 157.07 MBPS, and 426 IOPS
Lun 0_11 on drive [20:11     ] random read throughput: 150.99 MBPS, and 408 IOPS
Lun 0_2  on drive [20:2      ] random read throughput: 151.67 MBPS, and 425 IOPS
Lun 0_3  on drive [20:3      ] random read throughput: 155.05 MBPS, and 424 IOPS
Lun 0_4  on drive [20:4      ] random read throughput: 154.51 MBPS, and 425 IOPS
Lun 0_5  on drive [20:5      ] random read throughput: 152.84 MBPS, and 420 IOPS
Lun 0_6  on drive [20:6      ] random read throughput: 149.28 MBPS, and 414 IOPS
Lun 0_7  on drive [20:7      ] random read throughput: 155.20 MBPS, and 427 IOPS
Lun 0_8  on drive [20:8      ] random read throughput: 154.13 MBPS, and 425 IOPS
Lun 0_9  on drive [20:9      ] random read throughput: 155.88 MBPS, and 423 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: 269.03 MBPS, and 19650 IOPS
Lun 1_1  on drive [[10:0:1:0]] random read throughput: 268.72 MBPS, and 19635 IOPS
Lun 1_2  on drive [[10:0:2:0]] random read throughput: 268.54 MBPS, and 19635 IOPS
Lun 1_3  on drive [[10:0:3:0]] random read throughput: 268.96 MBPS, and 19633 IOPS
Lun 2_0  on drive [[12:0:0:0]] random read throughput: 269.86 MBPS, and 20441 IOPS
Lun 2_1  on drive [[12:0:1:0]] random read throughput: 270.08 MBPS, and 20397 IOPS
Lun 2_2  on drive [[12:0:2:0]] random read throughput: 269.19 MBPS, and 20437 IOPS
Lun 2_3  on drive [[12:0:3:0]] random read throughput: 269.49 MBPS, and 20418 IOPS
Lun 4_0  on drive [[9:0:0:0] ] random read throughput: 268.54 MBPS, and 19674 IOPS
Lun 4_1  on drive [[9:0:1:0] ] random read throughput: 268.31 MBPS, and 19703 IOPS
Lun 4_2  on drive [[9:0:2:0] ] random read throughput: 268.10 MBPS, and 19698 IOPS
Lun 4_3  on drive [[9:0:3:0] ] random read throughput: 268.74 MBPS, and 19683 IOPS
Lun 5_0  on drive [[11:0:0:0]] random read throughput: 268.82 MBPS, and 19690 IOPS
Lun 5_1  on drive [[11:0:1:0]] random read throughput: 268.27 MBPS, and 19697 IOPS
Lun 5_2  on drive [[11:0:2:0]] random read throughput: 268.57 MBPS, and 19704 IOPS
Lun 5_3  on drive [[11:0:3:0]] random read throughput: 268.36 MBPS, and 19689 IOPS
CALIBRATE results are within an acceptable range.

CALIBRATE stress test is now running...
Calibration has finished.
```

Create Celldisks

We have Celldisks & Flash Disks now

```
CellCLI> create celldisk all
CellDisk CD_00_sgsas1 successfully created
CellDisk CD_01_sgsas1 successfully created
CellDisk CD_02_sgsas1 successfully created
CellDisk CD_03_sgsas1 successfully created
CellDisk CD_04_sgsas1 successfully created
CellDisk CD_05_sgsas1 successfully created
CellDisk CD_06_sgsas1 successfully created
CellDisk CD_07_sgsas1 successfully created
CellDisk CD_08_sgsas1 successfully created
CellDisk CD_09_sgsas1 successfully created
CellDisk CD_10_sgsas1 successfully created
CellDisk CD_11_sgsas1 successfully created

CellCLI> list celldisk
         CD_00_sgsas1    normal
         CD_01_sgsas1    normal
         CD_02_sgsas1    normal
         CD_03_sgsas1    normal
         CD_04_sgsas1    normal
         CD_05_sgsas1    normal
         CD_06_sgsas1    normal
         CD_07_sgsas1    normal
         CD_08_sgsas1    normal
         CD_09_sgsas1    normal
         CD_10_sgsas1    normal
         CD_11_sgsas1    normal
         FD_00_sgsas1    normal
         FD_01_sgsas1    normal
         FD_02_sgsas1    normal
         FD_03_sgsas1    normal
         FD_04_sgsas1    normal
         FD_05_sgsas1    normal
         FD_06_sgsas1    normal
         FD_07_sgsas1    normal
         FD_08_sgsas1    normal
         FD_09_sgsas1    normal
         FD_10_sgsas1    normal
         FD_11_sgsas1    normal
         FD_12_sgsas1    normal
         FD_13_sgsas1    normal
         FD_14_sgsas1    normal
         FD_15_sgsas1    normal

CellCLI>
```

Create Griddisks

Create griddisk (first 100G – fast part) DATA

Create griddisk (the rest of the disk) RECO

```
CellCLI> create griddisk all harddisk prefix='data',size=100g
GridDisk data_CD_00_sgsas1 successfully created
GridDisk data_CD_01_sgsas1 successfully created
GridDisk data_CD_02_sgsas1 successfully created
GridDisk data_CD_03_sgsas1 successfully created
GridDisk data_CD_04_sgsas1 successfully created
GridDisk data_CD_05_sgsas1 successfully created
GridDisk data_CD_06_sgsas1 successfully created
GridDisk data_CD_07_sgsas1 successfully created
GridDisk data_CD_08_sgsas1 successfully created
GridDisk data_CD_09_sgsas1 successfully created
GridDisk data_CD_10_sgsas1 successfully created
GridDisk data_CD_11_sgsas1 successfully created

CellCLI> create griddisk all harddisk prefix='fra'
GridDisk fra_CD_00_sgsas1 successfully created
GridDisk fra_CD_01_sgsas1 successfully created
GridDisk fra_CD_02_sgsas1 successfully created
GridDisk fra_CD_03_sgsas1 successfully created
GridDisk fra_CD_04_sgsas1 successfully created
GridDisk fra_CD_05_sgsas1 successfully created
GridDisk fra_CD_06_sgsas1 successfully created
GridDisk fra_CD_07_sgsas1 successfully created
GridDisk fra_CD_08_sgsas1 successfully created
GridDisk fra_CD_09_sgsas1 successfully created
GridDisk fra_CD_10_sgsas1 successfully created
GridDisk fra_CD_11_sgsas1 successfully created

CellCLI>
```

Check griddisk size

First Part (first 100G)

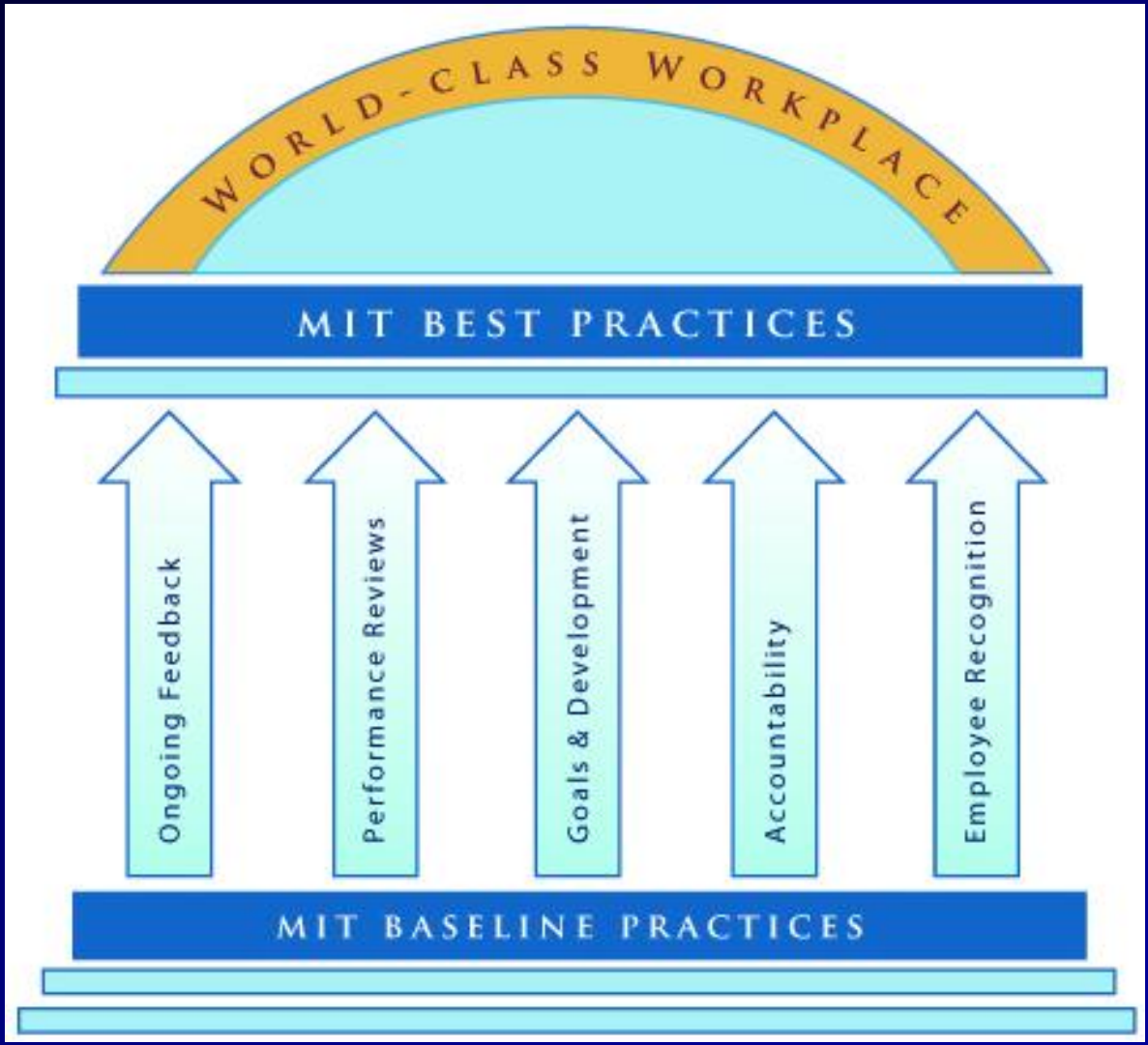
Second 458G (the rest of the disk)

```
CellCLI> list griddisk attributes name,size
         data_CD_00_sgsas1       100G
         data_CD_01_sgsas1       100G
         data_CD_02_sgsas1       100G
         data_CD_03_sgsas1       100G
         data_CD_04_sgsas1       100G
         data_CD_05_sgsas1       100G
         data_CD_06_sgsas1       100G
         data_CD_07_sgsas1       100G
         data_CD_08_sgsas1       100G
         data_CD_09_sgsas1       100G
         data_CD_10_sgsas1       100G
         data_CD_11_sgsas1       100G
         fra_CD_00_sgsas1        429.234375G
         fra_CD_01_sgsas1        429.234375G
         fra_CD_02_sgsas1        458.359375G
         fra_CD_03_sgsas1        458.359375G
         fra_CD_04_sgsas1        458.359375G
         fra_CD_05_sgsas1        458.359375G
         fra_CD_06_sgsas1        458.359375G
         fra_CD_07_sgsas1        458.359375G
         fra_CD_08_sgsas1        458.359375G
         fra_CD_09_sgsas1        458.359375G
         fra_CD_10_sgsas1        458.359375G
         fra_CD_11_sgsas1        458.359375G
```

Quick Disk Basics Overview

- Start with a Physical Disk or LUN (Logical Unit Number)
- Create a Cell Disk on a single LUN
- Create two Grid Disk slices on the Cell Disk – One Hot (first – outer ring) & One Cold
- Create 2 ASM Disk Groups (Hot/Cold – Data/Reco) across many Grid Disks to distribute the I/O across Grid Disks
- Add mirroring, DG, Flashback; Failure Groups (auto) ensure mirrored ASM extents are placed on different exadata cells.
- First 2 cells need 29G SYSTEM Area x 12 disks (coldest

Best Practices



MUST haves & DON'T do!

- Must have Bundle Patch 5 (See **note: 888828.1** for latest)
- Must have ASM
- Must have the correct data center **COOLING**!
  - 3 tiles with holes for full rack (400 CFM/tile) – don't melt it!
- Must have the **correct power needs**
- Must use Oracle Linux 5.3 (x86_64) & Oracle DB 11.2
- Must use RMAN for backups
- Consider StorageTek SL500 Tape backup
- Use an ASM allocation unit (AU) size of 4M

- Don't add any foreign hardware or – No Support!

Best Practices

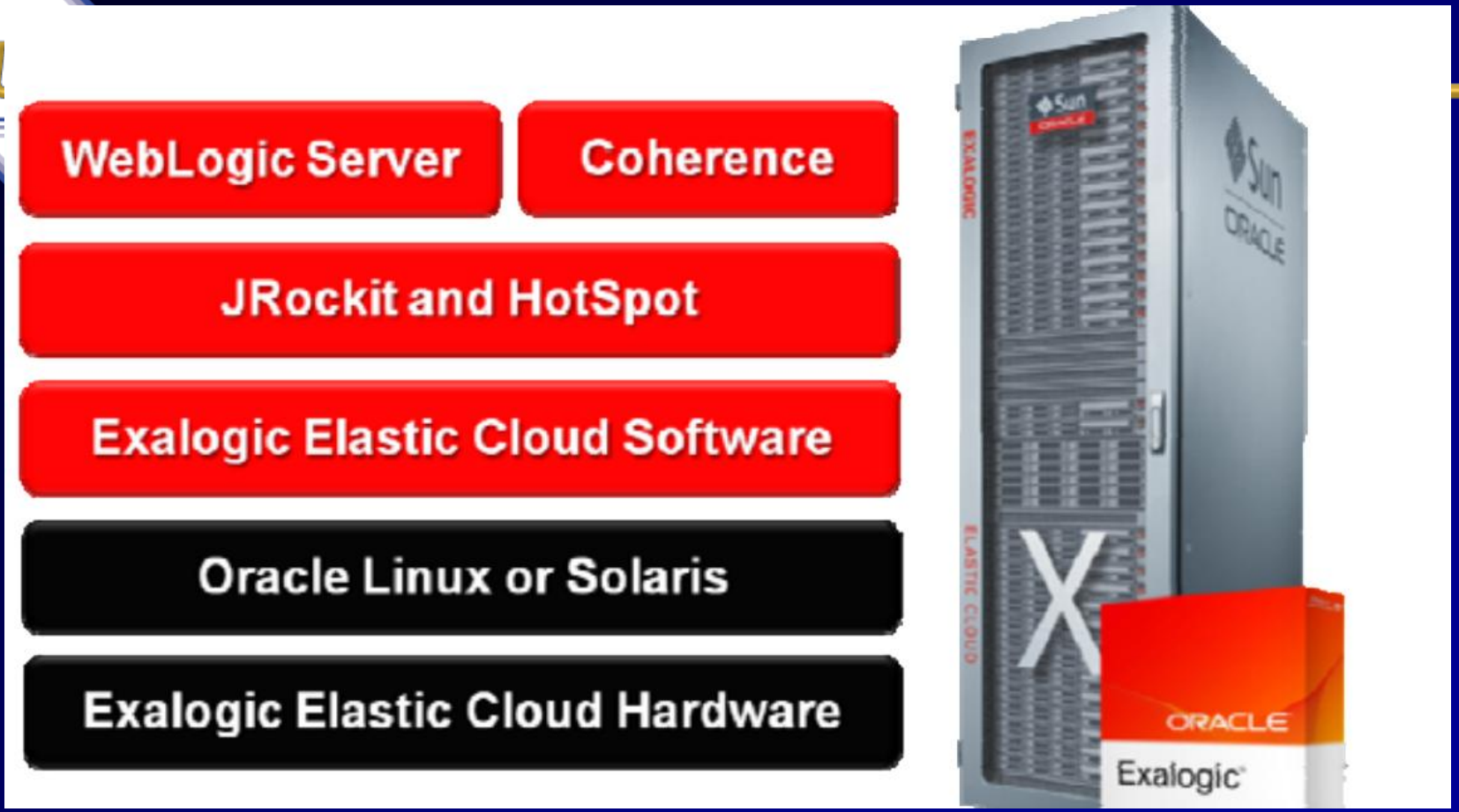
- Create ALL celldisk and griddisks
- Use DCLI to run on ALL Storage Servers at once
- Use IORM
- Decide Fast Recovery Area (FRA) & MAA Needs
- Database 11.2.0.1+ (11.2.1.3.1) and ASM 11.2.0.1+
- COMPATIBLE 11.2.0.1+
- Logfile size at 32G (Whoa!)
- LMT (Locally Managed Tablespaces) with at 4M uniform extents
- Move Data with Data Pump (or use INSERT /*+ APPEND */)

It's the Real Deal!!

- Fast Hardware!
- Many CPUs!
- Fast Flash Cache!
- Lot's of DRAM on Database Servers and Storage
- Compression (save 10x-70x)
- Partition Pruning (save 10-100x)
- Storage Indexes (save 5-10x)
- Smart Scan (save 4-10x)
- Turn a 1T search into a 500M search or even 50M

Exadata = Paradigm Shift!

What's Next – Exalogic Elastic Cloud!



- Some points here – Leveraging those acquisitions!
  - Coherence is a great product / NEW Linux – **Unbreakable Enterprise Kernel!**
  - 360 CPUs, 2.8T DRAM, 980G FlashFire SSD, 40T SAS – Will help Fusion Apps Smoke!
  - **1M HTTP/sec** – could fit Facebook on 2 of these even thought there are 500M people on Facebook

What's Next – Exadata X2-8

- 2 compute servers (7560 CPU at 2.26 GHz & 5T SAS)
  - 2 servers x 8 CPU sockets x 8 cores = **128 cores**
- 2 compute servers x 1T DRAM = **2T DRAM**
- **Same storage numbers…**

(FUTURE?? 8 servers = 512 CPUs & 8T of DRAM)

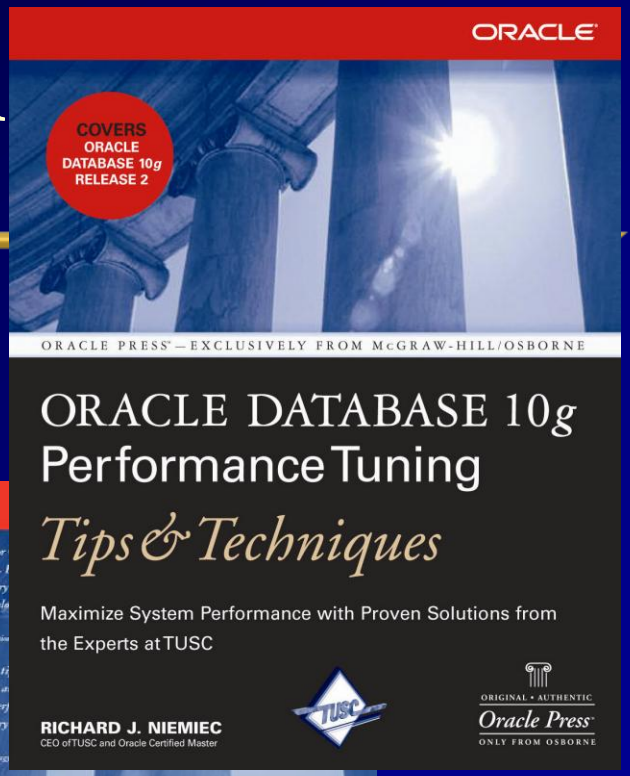
Summary – We Covered…

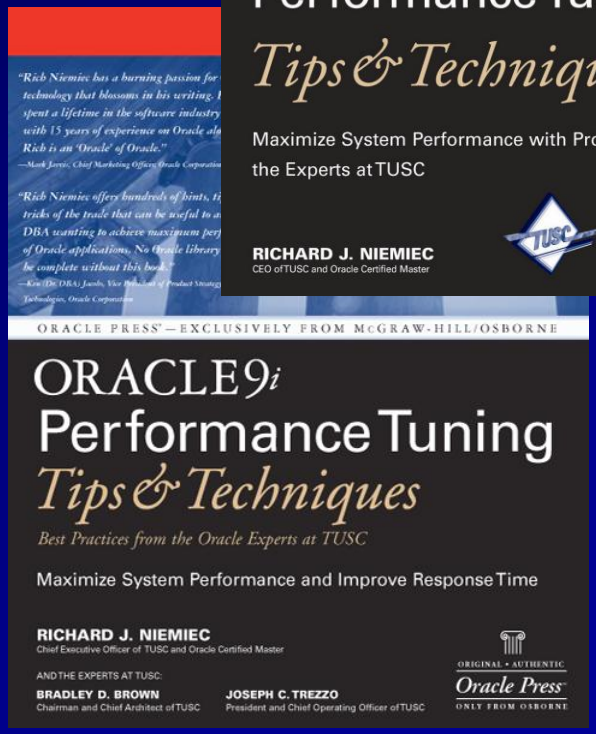
- Terminology & the Basics about Exadata
- Flash Cache
- Storage Index
- Smart Scans
- Hybrid Columnar Compression (HCC)
- Enterprise Manager & Grid Control
- Enterprise Manager Exadata Simulation
- I/O Resource Manager
- Security
- Utilities
- Best Practices

"We make a Living by what we get; We make a Life by what we give."

For More Information

- **www.tusc.com**
- *Oracle9i Performance Tuning Tips & Techniques; Richard J. Niemiec; Oracle Press (May 2003)*
- *Oracle 10g Tuning (June 11, 2007)*

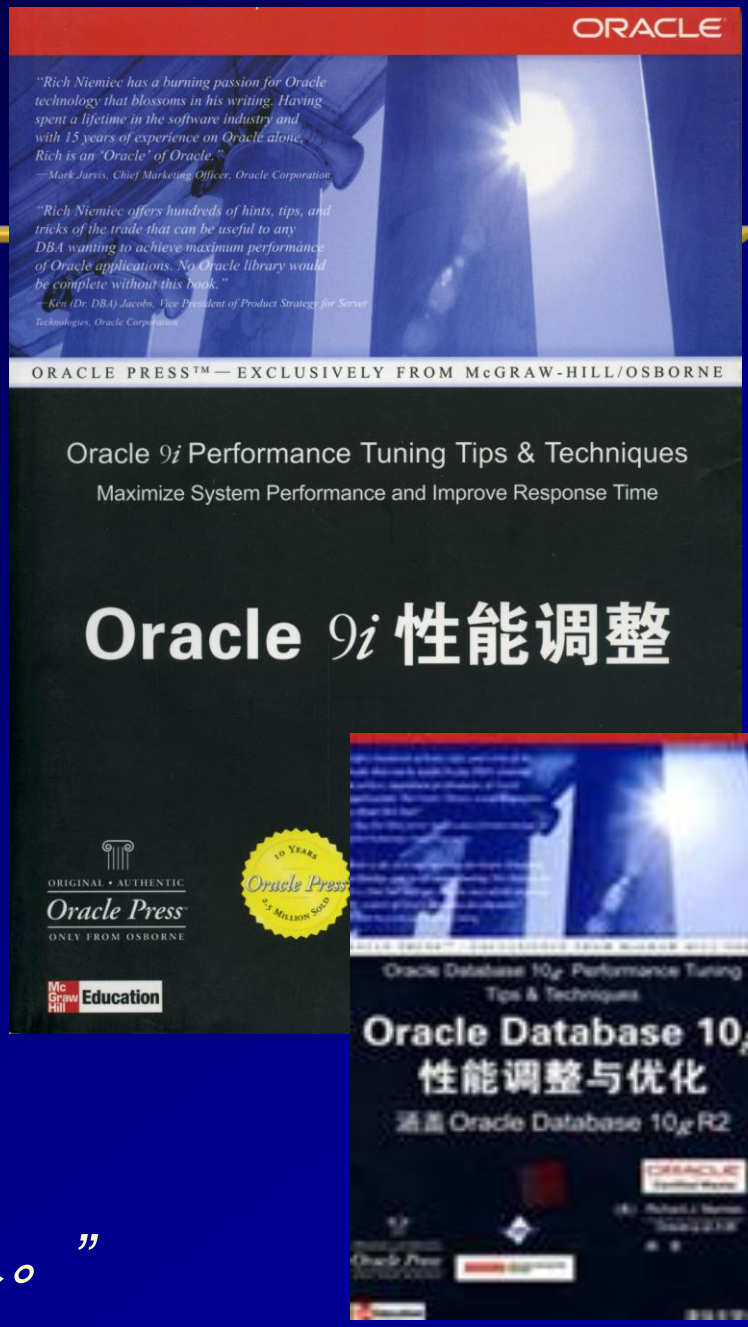




"If you are going through hell, keep going" - Churchill

更多信息

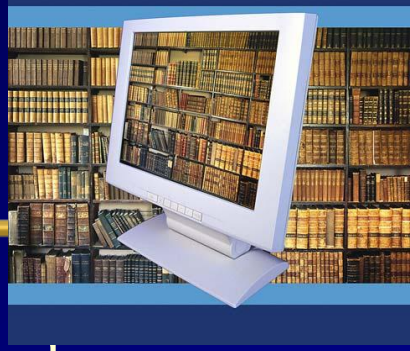
- www.tusc.com
- *Oracle9i Performance Tuning Tips & Techniques; Richard J. Niemiec; Oracle Press (May 2003)*
- *Oracle 10g Tuning (June 11, 2007)*



"成功只访问那些没空追求它的人。"

- Henry David Thoreau

References



- Exadata V2 – Sun Oracle Database Machine, Oracle
- Oracle Exadata Implementation Workshop, Oracle Corporation, McLean, Virginia - Multiple Exadata sessions
- Oracle Learning Library – multiple sessions/topics
- Oracle 11g R1/R2 Best Features, Rich Niemiec
- Oracle Enterprise Manager Deployment and High Availability Best Practices, Jim Viscusi (Oracle Corporation), Jim Bulloch (Oracle Corporation), Steve Colebrook-Taylor (Barclays Global Investors)
- *Oracle10g Performance Tuning Tips & Techniques*, Rich Niemiec, Oracle Press McGraw-Hill
- Advanced Compression with Oracle Database 11g Release 2, Oracle Corporation, Steven Lu
- Tech Crunch

Rolta TUSC – *Your* Partner …. Accomplished in Oracle!

2010 Oracle Partner of the Year (7 Titans Total)

Prior Years Winner 2002, 2004*, 2007*, 2008

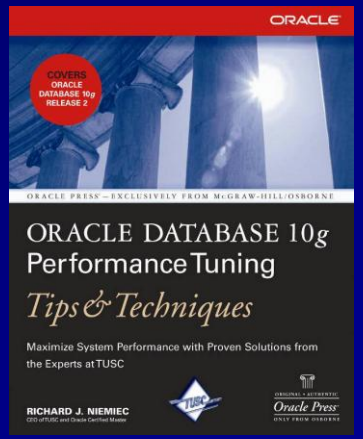
*Won 2 Awards

Rolta TUSC Services

- **Oracle**
  - E-Business Suite implementation, R12 upgrades, migration & support
  - Fusion Middleware and Open Systems development
  - Business Intelligence (OBIEE) development
  - Hyperion Financial Performance Management
  - DBA and Database tactical services
  - Strategic Global Sourcing
- **IT Infrastructure**
  - IT Roadmap - Security & Compliance - Infrastructure Management
  - Enterprise Integration / SOA - High Availability and Disaster Planning
- **Profitability & Cost Management**
  - Financial Consolidation - Budgeting & Forecasting
  - Profitability & Risk Analysis - Enterprise Performance Management
  - Operational, Financial & Management Reporting
- **Rolta Software Solutions**
  - *i*Perspective™ - rapid data & systems integration
  - Geospatial Fusion™ - spatial integration & visualization
  - OneView™ - business & operational intelligence

Rich's Overview (rich@tusc.com)

- Advisor to Rolta International Board
- Former President of TUSC
  - Inc. 500 Company (Fastest Growing 500 Private Companies)
  - 10 Offices in the United States (U.S.); Based in Chicago
  - Oracle Advantage Partner in Tech & Applications
- Former President Rolta TUSC & President Rolta EICT International
- Author (3 Oracle Best Sellers – #1 Oracle Tuning Book for a Decade):
  - Oracle Performing Tips & Techniques (Covers Oracle7 & 8i)
  - Oracle9i Performance Tips & Techniques
  - Oracle Database 10g Performance Tips & Techniques
- Former President of the International Oracle Users Group
- Current President of the Midwest Oracle Users Group
- Chicago Entrepreneur Hall of Fame - 1998
- E&Y Entrepreneur of the Year & National Hall of Fame - 2001
- IOUG Top Speaker in 1991, 1994, 1997, 2001, 2006, 2007
- MOUG Top Speaker Twelve Times
- National Trio Achiever award - 2006
- Oracle Certified Master & Oracle Ace Director



Copyright Information

- Neither Rolta TUSC nor the author guarantee this document to be error-free. Please provide comments/questions to rich@tusc.com. I am always looking to improve!
- Rich Niemiec/ Rolta TUSC © 2011. This document cannot be reproduced without expressed written consent from Rich Niemiec or an officer of Rolta TUSC, but may be reproduced or copied for presentation/conference use.

Contact Information

Rich Niemiec: rich@tusc.com

www.tusc.com